Statistical models and stochastic algorithms for the analysis of longitudinal Riemanian manifold valued data with multiple dynamic

Juliette Chevallier

▶ **To cite this version:**

Juliette Chevallier. Statistical models and stochastic algorithms for the analysis of longitudinal Riemanian manifold valued data with multiple dynamic. Statistics [math.ST]. Université Paris Saclay (COmUE), 2019. English. NNT : 2019SACLX059 . tel-02380752

HAL Id: tel-02380752

https://theses.hal.science/tel-02380752v1

Submitted on 26 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NNT : 2019SACLX059

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : École polytechnique

Laboratoire d'accueil : Centre de mathématiques appliquées de polytechnique (CMAP, UMR 7641 CNRS)

Spécialité de doctorat : Mathématiques appliquées

Juliette Chevallier

Modèles statistiques et algorithmes stochastiques pour l'analyse de données longitudinales à dynamiques multiples et à valeurs sur des variétés riemanniennes

Date de soutenance : 26 septembre 2019 *Lieu de soutenance :* École polytechnique (Palaiseau)

Après avis des rapporteurs :

Jean-Michel MARIN (Professeur des universités, Université de Montpellier)
Stefan SOMMER (Associate professor, University of Copenhagen)

Jury de soutenance :

Stéphanie ALLASSONNIÈRE	(Professeur, Université Paris-Descartes, CRC)	Directrice de thèse
Laure FOURNIER	(Professeur, Hôpital Européen Georges Pompidou)	Invitée
Jean-Michel MARIN	(Professeur, Université de Montpellier, IMAG)	Rapporteur
Eric MOULINES	(Professeur, École polytechnique, CMAP)	Président du jury
Nicolas VAYATIS	(Professeur, ENS Paris-Saclay, CMLA)	Examinateur
Christian ROBERT	(Professeur, Université Paris-Dauphine, Ceremade)	Examinateur

Juliette Chevallier
juliette.chevallier@polytechnique.edu
juliette.chevallier.perso.math.cnrs.fr

Thèse préparée au
Centre de mathématiques appliquées de l'École polytechnique
(CMAP, UMR 7641 CNRS)

Équipe SIMPAS : Signal IMage Probabilités Apprentissage Statistique,
Initiative Data Science de l'École polytechnique,
Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau

– Juliette Chevallier –

Modèles statistiques et algorithmes stochastiques pour l'analyse de données longitudinales à dynamiques multiples et à valeurs sur des variétés riemanniennes

Modèles statistiques et algorithmes stochastiques pour l'analyse de données longitudinales à dynamiques multiples et à valeurs sur des variétés riemanniennes

Mots clefs. Géométrie riemannienne, Données longitudinales, Optimisation stochastique, Modèles non-linéaires à effets mixtes, Algorithmes de type EM, Analyse spatio-temporelle, Estimation bayésienne.

Résumé. Par delà les études transversales, étudier l'évolution temporelle de phénomènes connait un intérêt croissant. En effet, pour comprendre un phénomène, il semble plus adapté de comparer l'évolution des marqueurs de celui-ci au cours du temps plutôt que ceux-ci à un stade donné. Le suivi de maladies neuro-dégénératives s'effectue par exemple par le suivi de scores cognitifs au cours du temps. C'est également le cas pour le suivi de chimiothérapie : plus que par l'aspect ou le volume des tumeurs, les oncologues jugent que le traitement engagé est efficace dès lors qu'il induit une diminution du volume tumoral. L'étude de données longitudinales n'est pas cantonnée aux applications médicales et s'avère fructueuse dans des cadres d'applications variés tels que la vision par ordinateur, la détection automatique d'émotions sur un visage, les sciences sociales, *etc.*

Les modèles à effets mixtes ont prouvé leur efficacité dans l'étude des données longitudinales, notamment dans le cadre d'applications médicales. Des travaux récent (Schiratti et al., 2015, 2017) ont permis l'étude de données complexes, telles que des données anatomiques. L'idée sous-jacente est de modéliser la progression temporelle d'un phénomène par des trajectoires continues dans un espace de mesures, que l'on suppose être une variété riemannienne. Sont alors estimées conjointement une trajectoire moyenne représentative de l'évolution globale de la population, à l'échelle macroscopique, et la variabilité inter-individuelle. Cependant, ces travaux supposent une progression unidirectionnelle et échouent à décrire des situations telles que la sclérose en plaques ou le suivi de chimiothérapie. En effet, pour ces pathologies, vont se succéder des phases de progression, de stabilisation et de remision de la maladie, induisant un changement de la dynamique d'évolution globale.

Le but de cette thèse est de développer des outils méthodologiques et algorithmiques pour l'analyse de données longitudinales, dans le cas de phénomènes dont la dynamique d'évolution est multiple et d'appliquer ces nouveaux outils pour le suivi de chimiothérapie. Nous proposons un modèle non-linéaire à effets mixtes dans lequel les trajectoires d'évolution individuelles sont vues comme des déformations spatio-temporelles d'une trajectoire géodésique par morceaux et représentative de l'évolution de la population. Nous présentons ce modèle sous des hypothèses très génériques afin d'englober une grande classe de modèles plus spécifiques.

L'estimation des paramètres du modèle géométrique est réalisée par un estimateur du *maximum a posteriori* dont nous démontrons l'existence et la consistance sous des hypothèses standards. Numériquement, du fait de la non-linéarité de notre modèle, l'estimation est réalisée par une approximation stochastique de l'algorithme EM, couplée à une méthode de Monte-Carlo par chaînes de Markov (MCMC-SAEM). La convergence du SAEM vers les maxima locaux de la vraisemblance observée ainsi que son efficacité numérique ont été démontrées. En dépit de cette performance, l'algorithme SAEM est très sensible à ses conditions initiales. Afin de palier ce problème, nous proposons une nouvelle classe d'algorithmes SAEM dont nous démontrons la convergence vers des minima locaux. Cette classe repose sur la simulation par une loi approchée de la vraie loi conditionnelle dans l'étape de simulation. Enfin, en se basant sur des techniques de recuit simulé, nous proposons une version tempérée de l'algorithme SAEM afin de favoriser sa convergence vers des minima globaux.

Statistical Models and Stochastic Algorithms for the Analysis of Longitudinal Riemannian Manifold Valued Data with Multiple Dynamic

Keys words. Riemannian geometry, Longitudinal data, Stochastic optimization, Nonlinear mixed effect models, EM-like algorithms, Spatio-temporal analysis, Bayesian estimation.

Abstract. Beyond transversal studies, temporal evolution of phenomena is a field of growing interest. For the purpose of understanding a phenomenon, it appears more suitable to compare the evolution of its markers over time than to do so at a given stage. The follow-up of neurodegenerative disorders is carried out *via* the monitoring of cognitive scores over time. The same applies for chemotherapy monitoring: rather than tumors aspect or size, oncologists asses that a given treatment is efficient from the moment it results in a decrease of tumor volume. The study of longitudinal data is not restricted to medical applications and proves successful in various fields of application such as computer vision, automatic detection of facial emotions, social sciences, *etc.*

Mixed effects models have proved their efficiency in the study of longitudinal datasets, especially for medical purposes. Recent works (Schiratti et al., 2015, 2017) allowed the study of complex data, such as anatomical data. The underlying idea is to model the temporal progression of a given phenomenon by continuous trajectories in a space of measurements, which is assumed to be a Riemannian manifold. Then, both a group-representative trajectory and inter-individual variability are estimated. However, these works assume an unidirectional dynamic and fail to encompass situations like multiple sclerosis or chemotherapy monitoring. Indeed, such diseases follow a chronic course, with phases of worsening, stabilization and improvement, inducing changes in the global dynamic.

The thesis is devoted to the development of methodological tools and algorithms suited for the analysis of longitudinal data arising from phenomena that undergo multiple dynamics and to apply them to chemotherapy monitoring. We propose a nonlinear mixed effects model which allows to estimate a representative piecewise-geodesic trajectory of the global progression and together with spatial and temporal inter-individual variability. Particular attention is paid to estimation of the correlation between the different phases of the evolution. This model provides a generic and coherent framework for studying longitudinal manifold-valued data.

Estimation is formulated as a well-defined *maximum a posteriori* problem which we prove to be consistent under mild assumptions. Numerically, due to the non-linearity of the proposed model, the estimation of the parameters is performed through a stochastic version of the EM algorithm, namely the Markov chain Monte-Carlo stochastic approximation EM (MCMC-SAEM). The convergence of the SAEM algorithm toward local maxima of the observed likelihood has been proved and its numerical efficiency has been demonstrated. However, despite appealing features, the limit position of this algorithm can strongly depend on its starting position. To cope with this issue, we propose a new version of the SAEM in which we do not sample from the exact distribution in the expectation phase of the procedure. We first prove the convergence of this algorithm toward local maxima of the observed likelihood. Then, with the thought of the simulated annealing, we propose an instantiation of this general procedure to favor convergence toward global maxima: the tempering-SAEM.

À mes deux grand-mères,
Marthe Chevallier et Marie Rabot.

« La vie n'a pas besoin d'être facile,
à condition qu'elle ne soit pas vide. »

Lise Meitner [1]

[1] ***Lise Meitner*** *(1878–1968), physicienne autrichienne puis suédoise.* Ses travaux en physique nucléaire ont joué un rôle majeur dans la découverte de la fission nucléaire. En 1944, Otto Hahn reçoit le prix Nobel de chimie pour cette découverte et des travaux auxquels Lise Meitner avait largement contribué. Lise Meitner, elle, ne le recevra jamais.

Remerciements

C'est non sans une certaine émotion et beaucoup de reconnaissance que je voudrais profiter de cet espace pour remercier toutes les personnes ayant contribué de près ou de loin à la présente thèse.

Mes premières pensées vont bien évidement à Stéphanie Allassonnière, à qui je souhaite exprimer ma profonde gratitude pour avoir accepté d'encadrer cette thèse et d'y avoir toujours cru (parfois plus que moi !). Je tiens à la remercier pour sa gentillesse, toute la patience dont elle a fait preuve à mon égard et la confiance qu'elle m'a accordée, et ce dès le début de cette aventure, ainsi que pour la grande liberté dont j'ai joui durant ces trois années. Merci de m'avoir montré que les mathématiques passaient aussi par des discussions informelles, de m'avoir toujours incitée à exposer mes travaux, assister à des conférences. Pour tout cela je lui suis reconnaissante et je la remercie plus généralement de m'avoir guidée et accompagnée tout au long de ma thèse.

Je suis très heureuse de pouvoir remercier Jean-Michel Marin et Stefan Sommer d'avoir tout de suite accepté de rapporter cette thèse et d'y avoir prêté autant d'attention. Je vous remercie sincèrement pour votre relecture attentive et vos commentaires bienveillants. Mes remerciements vont également à Laure Fournier, Éric Moulines, Christian Robert et Nicolas Vayatis pour me faire l'honneur de participer à mon jury de soutenance. Merci de l'intérêt que vous avez porté à mes travaux.

Je remercie vivement toutes les personnes avec lesquelles j'ai eu la chance de collaborer. En premier lieu, Laure Fournier pour son expertise médicale et sa pédagogie lors de nos échanges. Je tiens également à remercier tout le groupe de travail de l'ICM qui fut pour moi un lieu d'échanges et d'enrichissements : Alexandre, Igor, Manon, Maxime, Raphaël et Stanley Durrleman. Enfin, je remercie Vianney pour sa disponibilité et sa réactivité. Je remercie aussi toutes les personnes qui m'ont accompagnée dans mes premiers pas d'enseignante, notamment Stéphanie Allassonnière pour m'avoir confié les TDs associés à son cours du MVA mais également mes étudiants et Flore Nabet avec qui j'ai eu la chance de travailler lors de ma troisième année de thèse.

Mes pensées vont également à mes collègues du *séminaire des doctorants·e·s* sans lesquels l'organisation du séminaire eût été plus pénible. Aymeric, merci pour ta pugnacité lorsqu'il s'agissait de trouver des orateurs motivés au CMLS. Céline, merci pour ta bonne humeur et ton maniement expert du logiciel de réservation des salles (avec une spécialité réservation de dernière minute, l'anticipation n'étant manifestement pas

le maître mot de notre organisation). Enfin, Corentin, merci pour tes talents de pâtissier et, plus sérieusement, ton grand dévouement dans l'organisation de ce séminaire.

Je tiens à remercier tous les professeurs que j'ai eus au cours de ma scolarité ou qui m'ont encadrée en stage. J'ai une pensée particulière pour Hervé Le Meur qui m'a, d'une certaine manière, initiée à la recherche mathématique en encadrant avec une extrême patience mes premiers « travaux » de recherche. J'ai gardé en mémoire de ces longues heures passées devant son tableau de précieux conseils qu'il m'arrive encore d'appliquer. Mes pensées vont également à Pierre-Guy Plamondon, Antoine Chambert-Loir, Christophe Giraud et Joan Glaunès pour leur présence, leur encouragements et leurs enseignements mathématiques qui sont à la genèse de ce manuscrit. En particulier, je n'oublie ni le précieux temps que m'a consacré Christophe Giraud afin de m'aider dans ma recherche de thèse ni la disponibilité de Joan Glaunès lors de mon stage recherche de master 2. De manière plus éloignée dans le temps, je tiens également à remercier mes professeurs de classes préparatoires, Étienne Mahé et Grégoire Taviot, pour m'avoir répété de nombreuses fois et sans montrer de signes manifestes d'agacement que j'avais le niveau pour poursuivre des études en mathématiques au magistère d'Orsay et *a fortiori* en doctorat. Enfin, tout ceci est un peu la faute de ma professeure de terminal, Sylvie Colesse, qui m'a faite entrapercevoir la richesse des mathématiques et que je ne remercierai jamais assez de m'avoir convaincue d'aller souffrir en maths sup (quoi que !).

Je profite de ces quelques lignes pour remercier toutes les personnes qui ont eu la même folle idée que moi de s'investir dans l'organisation de *La tournée de* π. En premier lieu Antoine sans qui je n'aurais sûrement pas rejoint l'aventure mais également Augustin, Joël, Luc, Martin, Maxence et Pierre.

Cette thèse a été marquée par la bonne ambiance qui règne dans le bureau 2010 et je voudrais remercier tous les (ex)-doctorants qui l'ont occupé. Antoine, merci pour ton sens de l'organisation sans faille, tes réponses à toutes les questions que l'on ne se posait pas au sujet de l'administratif et d'avoir veillé à la discipline du bureau ces trois années durant. Arthur, merci de nous avoir laissé tant d'espace en venant si peu. Plus sérieusement, merci pour nos discussions toujours animées mais rarement inintéressantes, que ce soit sur la politique ou les maths. Céline, merci pour ta chaleureuse présence quand la motivation venait à me manquer. Merci d'avoir permis à Antoine de nous rappeler à l'ordre en acceptant de répondre à mes sollicitations, merci pour les pique-niques au bord du lac, les services rendus et, surtout, merci pour ton soutien ces derniers mois. Fedor, merci pour les pirojkis, de m'avoir obligée à parler moins vite (même si ça n'a pas toujours été très efficace !) et pour tes conseils lors de mon voyage en Russie. Heythem, merci pour ces moments partagés, pas tant au bureau à proprement parler mais plutôt en salle café autour d'une tasse de tisane/café, au krav maga où j'ai pû jouir de la chance de te mettre des coups de pieds en toute impunité ou encore en dehors de ces murs. Merci d'apprécier les concerts bruyants et la bière. Enfin, un énorme merci et une grosse dose de gratitude pour tes talents de coursier lors de l'envoi du présent manuscrit à mes rapporteurs. Par contre, autant Céline qu'Heythem, je ne vous remercie pas de n'avoir jamais accepté de participer à une course d'obstacles avec moi ! Martin, merci pour les

discussions sur la musique et les pizzas partagées lors de nos trop longues réunions chez toi. Merci à Josué et Adrien qui sont arrivés sur la fin mais qui ont également contribué par leur présence à l'ambiance du bureau. Une mention spéciale à Josué pour ses non-recommandations cinématographiques. Je remercie également tous les doctorants que j'ai eu l'occasion de croiser au CMAP, qui n'ont pas eu la chance de faire partie du bureau 2010 mais qui ont participé à l'ambiance du laboratoire : Kevish pour le yoga et le krav mais également Alejandro, Belhal, Léa, Cheikh, Corentin, Florian, Frédéric, Geneviève, Hadrien, Jaouad, Jean-Bernard, Ludovic, Mathilde, Mathieu, Nicolas, Othmane, Perle, Raphaël, Rémi, Romain, Tristan, Vianney et j'en oublie sûrement... Enfin, je ne peux conclure ce paragraphe sans remercier Sirius Black de m'avoir si souvent rappelée à l'ordre et de participer à la décoration d'exception de ce bureau !

Outre les rencontres effectuées au CMAP, je tiens à remercier tous mes amis matheux, d'Orsay et d'ailleurs, qui ont suivi cette thèse de plus ou moins loin alors qu'ils n'en étaient en rien forcés ! En espérant n'en oublier aucun et après avoir longuement hésité entre ranger BenoîtS Tran et Robert à la lettre B ou aux lettres T et R respectivement, je tiens à adresser mes sincères remerciements à toutes les personnes qui ont subit le magistère d'Orsay (et parfois plus !) avec moi. Augustin, qui me supporte depuis si longtemps, merci pour nos discussions sur la diffusion des maths et (nos discussion sur) le rhum arrangé, merci pour les échanges un brin nostalgiques de la prépa et pour les cours de didjeridoo. Benjamin, qui fait office d'intrus dans la catégorie « anciens d'Orsay » mais qui a sû se faire une place. Tran, merci pour ton optimiste déconcertant dont il y a, selon moi, beaucoup à apprendre. Robert, merci pour les bons moments partagés, et ce malgré tes origines bretonnes (!), et pour nos conversations musicales. Céline encore, merci pour Marseille, le rosée et d'écouter de la musique aussi... Merci également pour tes talents de cuisinière et, en fait, la liste est trop longue ! Diane, merci de m'avoir convaincue de ne pas devenir prof par tes discours virulents. François, merci de m'avoir initiée aux bières trappistes, aux chemises à fleurs et de m'avoir faire redécouvrir Jacques Brel. Léo, merci pour ta programmation de qualité, bien que parfois quelques peu poussiéreuse, à ton « ciné-club » où tu as eu la gentillesse de toujours me convier alors que je n'étais pas des plus assidues. Perrine, merci pour ces précieux moments partagés bien que trop peu nombreux et pour la simplicité de nos échanges. Désolée d'être si peu disponible ! Merci Raphaël pour nos échanges mathématiques ou non, au détour d'un couloir, d'une rue d'Orsay ou de mon canapé, toujours appréciables. Enfin, Samuel, merci d'apporter toujours plus de politiquement non-correct à nos discussions et de cultiver chez moi un humour toujours plus douteux. Merci pour ces sorties improbables et ces discussions nocturnes. J'ai aussi une pensée pour Éléonore et Justine.

J'ai aussi des amis qui ont eu la présence d'esprit de ne pas s'engouffrer dans les mathématiques et je tiens à les remercier pour tous les moments de détente et leur indulgence envers mes manquements en matière de communication. En premier, un énorme merci à Mickaël de n'avoir jamais pris ombrage de mes longs silences. Merci pour la randonnée, les entrainements informels de self, les verres partagés quand je ne pouvais plus voir les maths en peinture, les après-midis au soleil dans ta maison au bout du monde, la nourriture pour poney et j'en passe... Dans la même veine, un gros merci à Sheley pour

sa motivation et sa bonne humeur. Un gros merci à Patrick pour son émerveillement crédule face aux mathématiques, les coquillages en fractale, les photos d'araignées alors qu'on est tous les deux arachnophobes et nos discussions scientifico-ésotériques sur la vie. Mes pensées vont également à Solène à qui je dois beaucoup de bons moments mais peu de nuits de sommeil ! Merci pour tes sourires pétillants et ton affection dans les moments difficiles (et même dans les moments où tout va bien d'ailleurs), merci pour les cours d'œnologie et nos discussions qui ne terminent qu'au petit matin. Je dois également énormément à Xavier qui a été des plus présents pour moi sur la première moitié de cette thèse, et je tiens à le remercier. Merci pour ta confiance et ta disponibilité, merci pour nos débats conflictuels, les promenades en bord de mer, les glaces de gros, les cours de géographie et ceux de mécanique, 1-3-4-2 ! Je remercie également Franck et Johanna d'être suffisamment fous pour me faire passer pour saine d'esprit à côté, merci d'être toujours de bonne compagnie et d'une (aussi) rare gentillesse. Un merci à Laurine pour sa douceur et sa prévenance ; un merci également à Nico et Katia, mes jeunes vieux mariés préférés. Un merci à Phil de n'avoir cessé de prendre de mes nouvelles même quand mes réponses se faisaient rares. Enfin, un merci à Raphaëlle, mon entraineuse de self-défense, pour sa patience nos entrainements durant et tous ses enseignements, que ce soit de l'art de mettre un coup de poing efficace ou de celui de la communication : verbale, para-verbale et surtout non-violente pour la gestion de conflits...y compris dans la vie de tous les jours ! De ces années de self, je retiendrai surtout que *la meilleure des défense est l'attaque* mais j'espère ne pas avoir à me servir de ce conseil lors de cette soutenance !

J'en viens maintenant aux remerciements plus personnels, à l'égard de ma famille et des personnes qui me sont les plus chères et sans qui rien de tout cela n'aurait pu voir le jour. Tout d'abord ma famille qui est trop nombreuse pour que je puisse l'énumérer en entier (et de toute façon ces remerciements sont déjà trop longs), mes oncles, tantes, cousins, cousines qui se sont toujours intéressés à ce que je faisais sans forcément comprendre grand chose. Une pensée un peu particulière pour Agnès, Bénédicte, Éva, Marielle et, surtout, pour ma grand-mère, Marie, qui, si elle ne comprend plus grand chose depuis bien longtemps, s'est toujours efforcée de trouver cela « formidable ». Je voudrais également remercier mes parents, Laurence et Thierry, qui me soutiennent dans mes choix et sont à mes côtés depuis le début. Merci de votre enthousiasme pour mes projets. Vous m'avez toujours écoutée avec attention parler de mes cours puis de ma recherche, allant jusqu'à relire certaines parties de cette thèse sans pour autant être vous-même scientifiques. J'adresse un énorme merci à mes deux sœurs, Camille et Lisa. Merci pour vos petites attentions qui me touchent à chaque fois et notre complicité inchangée malgré des parcours de vie différents : heureusement que je suis la plus vieille de nous trois pour n'avoir ni le permis, ni terminé mes études alors que toutes les deux oui ! Camille, merci de m'avoir toujours si bien reçue chez toi (y compris pour passer mes écrits de l'agrégation !), pour ta confiance et nos discussions toujours pleines de bienveillance et d'écoute. D'ailleurs, je suis fière d'être la première à avoir vu ton premier tatouage ! Un merci plus discret à Adrien, que j'ai toujours plaisir à croiser même s'il ne m'a toujours pas convaincue de me mettre au foot. Lisa, merci d'être aussi pleine d'énergie, impulsive

et un brin inconsciente. Merci pour nos fous-rires partagés et tes mots toujours gentils à mon égard. Je ne sais pas si je dois te remercier d'avoir ramené dans ma vie l'espèce de psycho-cat qui me sert de chat, le bien ou mal nommé Kafka, et qui n'a de cesse de me griffer et de semer ses poils dans tout mon appartement. En tout cas, par son biais, on peut dire que tu occupes une place de choix dans ma vie.

Il me reste à remercier les deux personnes qui ont par leur présence le plus contribué à cette thèse. Je ne sais comment te remercier Eric pour l'écoute dont tu as fait preuve à mon égard, et plus encore durant la période de rédaction. Merci pour ta disponibilité, ton soutien sans faille, tes nombreux câlins, même à plusieurs kilomètres d'écart et tes petits messages attentionnés. Merci de m'avoir fait suffisamment de place dans ton cœur et dans ta vie pour te soucier de moi quand tu avais mieux à faire. Merci de faire de l'informatique à Lyon et de vouloir mettre de la couleur dans ma vie (il y a encore pas mal de travail !). Merci pour notre virée en Suisse, les sorties entre b*tch*s et les semi-marathons. Pour tout ça et bien plus encore, un énorme merci.

Enfin, je ne saurais conclure ces remerciements sans évoquer Pierre qui m'a subit à ses côtés durant ces trois années. Merci de ta patience infinie à m'écouter et à me conseiller. Merci de m'avoir acceptée comme je suis et d'être qui tu es ; merci de ta gentillesse sans précédent et ta présence indéfectible à mes côtés. Merci d'avoir toujours cru en moi, de me soutenir, de m'encourager. Je ne te remercierai jamais assez pour tout ce que tu as fait pour moi : tu n'as pas compté ton temps et ton énergie, jusqu'à parfois te causer du tord. Les mots ne sont pas suffisant pour exprimer ma gratitude à ton égard : cette thèse te doit tellement et je t'en serais à jamais reconnaissante. Je te souhaite beaucoup de réussite dans ta carrière de chercheur mais aussi d'écrivain.

Glossaire

Notations

$[\![a, b]\!]$	The set of integers between a and b : $\{a, a+1, a+2, \ldots, b-2, b-1, b\}$.
$clos(X)$	Closure of the set X, *i.e.* the smallest closed set containing X.
$int(X)$	Interior of the set X, *i.e.* the largest open set contained in X.
$d(x, X)$	When X is a closed set, distance of x to X.
$Conv(X)$	Convex hull of the set X, *i.e.* the smallest convex set that contains X.
$\mathcal{D}iam(X)$	Diameter of the set X, *i.e.* $\sup\{d(x, y) \vert x, y \in X\}$.
$\mathscr{B}(X)$	Borel algebra of the set X, *i.e.* the smallest σ-algebra containing all open sets of X (or, equivalently, all closed sets of X).
$\mathcal{H}^d$	Hausdorff measure of dimension d.
$\mathscr{L}_d$	Lebesgue measure of dimension d.
$\mathscr{M}_{m,n}(\mathbb{R})$	Real matrices with m lines and n columns.
$\mathscr{S}_n^+(\mathbb{R})$	Real symmetric matrices of size n.
$\mathscr{C}^k(\mathbb{R}^d)$	Diffeomorphism of $\mathbb{R}^d$, which are k times differentiable, with derivative up to the k^{th} continue.
$\mathscr{C}_0^1(\mathbb{R}^d)$	Diffeomorphism of $\mathbb{R}^d$, which vanish at infinite and whose differential vanish at infinite.
$\mathbb{L}^p(X, \mu)$	Lebesgues spaces ; space of functions for which the p-th power of the absolute value is Lebesgue integrable. Equivalently, $\Vert f\Vert_p = \left(\int_X \vert f\vert^p \, \mathrm{d}\mu\right)^{1/p} < +\infty$.
$\langle x \,\vert\, y \rangle$	Scalar product between x and y. Unless otherwise specified, the canonical one regarding the considered set.
T_xM	Tangent space at x of the manifold M, *i.e.* the set of all tangent vector of M at x.
TM	Tangent bundle of the manifold M, *i.e.* the disjoint union of all the tangent spaces of M : $TM = \bigsqcup_{x \in M} T_xM$.

$\mathcal{E}xp_x(v)$	Riemannian exponential map at point x, of tangent vector v, *i.e.* value at time 1 of the unique geodesic passing through x at time 0 with speed v.
$P_{\gamma,t_0,t}(w)$	Parallel transport of the vector w, along the curve γ between the points $\gamma(t_0)$ and $\gamma(t)$.
$\Lambda^d E$	d-times exterior power of the euclidean space E, where E is of dimension $n \geqslant d$. $\Lambda^d E$ is a vector space of dimension $\binom{n}{d}$ spanned by the set of simple d-vectors $\xi_1 \wedge \ldots \wedge \xi_d$.
$\Omega_0^d(E)$	Continuous forms of dimension d on E, defined as above, vanishing at infinity *i.e.* $\Omega_0^d(E) = \mathscr{C}_0^0\left(E, \Lambda^d E^*\right)$.
$G_d(E)$	Grassmann manifold of dimension d in E, defined as above, *i.e.* the set of all d-dimensional subspaces of E.

Abbreviations

· **A** ·	**ABC**	Approximate Bayesian computation.
· ·	**ANOVA**	Analysis of variance.
· ·	**a.s.**	Almost surely.
· **B** ·	**BG-ICA**	Bernouilli-Gaussian ICA.
· **E** ·	**EM**	Expectation-Maximization.
· **G** ·	**GEM**	Generalized EM.
· ·	**GMM**	Gaussian mixture model.
· **H** ·	**HEGP**	Hôpital européen Georges Pompidou.
· ·	**HIV**	Human immunodeficiency viruses.
· **I** ·	**ICA**	Independent component analysis.
· ·	**IFA**	Independent factor analysis.
· ·	**i.i.d**	Independent and identically distributed.
· **K** ·	**KL**	Kullback-Leibler.
· **L** ·	**LDDMM**	Large deformation diffeomorphic metric mapping.
· **M** ·	**MAP**	Maximum a posteriori.
· ·	**MCEM**	Monte-Carlo EM.

· ·	**MCMC**	Markov chain Monte-Carlo.
· ·	**MLE**	Maximum likelihood estimation.
· ·	**MRI**	Magnetic resonance imaging.
· **P** ·	**PDE**	Partial differential equation.
· **R** ·	**RESCIST**	Response evaluation criteria in solid tumours.
· ·	**RKHS**	Reproducing kernel Hilbert space.
· **S** ·	**SAEM**	Stochastic Approximation of the EM.
· ·	**SEM**	Stochastic EM.
· **T** ·	**tmp-SAEM**	tempering-SAEM.
· **U** ·	**USPS**	U.S. Postal Service.

Table des matières

Table des figures

Chapter 5 – The piecewise-logistic and piecewise-geodesic shape models

Chapter 7 – A New Class of SAEM Algorithms. Escaping Local Minima and Handling Intractable Sampling

– Première partie –

Introduction

– Chapitre I –

Introduction en langue française

Le but de cette thèse est de proposer de nouvelles méthodes pour l'analyse statistique de données longitudinales à valeurs sur des variétés riemanniennes. L'étude de ce type de données étant généralement conduite par l'intermédiaire de modèles à effets mixtes, on commence par passer en revue la littérature classique sur ces modèles ainsi que les différentes techniques d'inférence statistique qui leurs sont associées.

Bien qu'efficients, ces modèles ne s'appliquent pas en l'état aux données à valeurs sur des variétés. Aussi s'attache-t-on à redéfinir le cadre d'étude classique pour l'étude des formes anatomiques, à savoir les espaces de forme. Sur la base de ce formalisme, on se propose d'effectuer une revue des différents modèles de régression géodésique pour l'étude de jeux de données longitudinales à valeurs sur des variétés. Ce type de modèle ne permet cependant pas d'expliquer la variabilité inter-individuelle.

Dans ce but et en se basant sur une notion de reparamétrage temporel, Schiratti et al. (2015, 2017) ont proposé un modèle générique pour l'étude de données longitudinales sur des variétés. Ce modèle fait l'objet du dernier paragraphe du présent chapitre.

Contents

Par delà les études transversales, étudier l'évolution temporelle de phénomènes connait un intérêt croissant. En effet, pour comprendre un phénomène, il semble plus adapté de comparer l'évolution des marqueurs de celui-ci au cours du temps plutôt que ceux-ci à un stade donné. Le suivi de maladies neuro-dégénératives, telles que les maladies d'Alzheimer, d'Huntington ou de Parkinson, s'effectue par exemple par le suivi de scores cognitifs au cours du temps. De fait, plus que la dégénérescence neuronale qui est une conséquence normale du vieillissement cérébral, le caractère pathologique des maladies dites neuro-génératives réside dans la rapidité de progression de cette dégénérescence. C'est également le cas pour le suivis de chimiothérapie : plus que par l'aspect ou le volume des tumeurs, les oncologues jugent que le traitement engagé est efficace dès lors qu'il induit une diminution du volume tumoral. L'étude de données longitudinales n'est pas cantonnée aux applications médicales et s'avère fructueuse dans des cadres d'applications variés tels que la vision par ordinateur, la détection automatique d'émotions sur un visage, les sciences sociales, *etc.*

L'étude de données longitudinales est généralement conduite par l'intermédiaire de modèles statistiques à effets mixtes. Des travaux récents ont notamment permis l'étude de données complexes, telles que des données anatomiques. Cependant, ces travaux supposent une progression unidirectionnelle et échouent à décrire des situations telles que la sclérose en plaques ou le suivi de chimiothérapie. En effet, pour ces pathologies, vont se succéder des phases de progression, de stabilisation et de remision de la maladie, induisant un changement de la dynamique d'évolution globale.

Le but de cette thèse est de développer des outils méthodologiques et algorithmiques pour l'analyse de données longitudinales, dans le cas de phénomènes dont la dynamique d'évolution est multiple et d'appliquer ces nouveaux outils pour le suivi du cancer du rein métastatique.

I. Modèles à effets mixtes pour l'étude de données longitudinales

Nous nous intéressons à l'analyse statistiques de mesures provenant de l'observation d'un phénomène de manière répétée dans le temps.

Les modèles à effets mixtes (Eisenhart, 1947; Fisher, 1919) permettent d'expliquer la variabilité d'une série d'observations par deux types d'effets : des *effets fixes* communs à l'ensemble des individus de la population, et des *effets aléatoires* propre à chacun des individus et traduisant la variabilité liée à chaque sujet. Ce type de modèles, de par leur nature mutli-échelle, est donc tout particulièrement adapté à l'étude de données

provenant d'une structure hiérarchique, dont l'analyse de données longitudinales est un cas particulier.

On distingue deux types de modèles à effets mixtes : les modèles linéaires, et les modèles non-linéaires. Les modèles à effets mixtes ont été considérablement étudiés, notamment du fait de leur grande applicabilité. Pour ne donner que deux exemples, dans le domaine médical, il se sont imposés comme outils de référence aussi bien dans le suivi de maladies neuro-dégénératives (Milliken et Edland, 2000; Ospina et al., 2012) que dans le suivi de chimio et radiothérapies (Ribba et al., 2014; Rios et al., 2017).

Dans Verbeke et al. (2014), on trouvera une revue claire de ce type de modèles et pour un aperçu plus complet, on peut par exemple se référer à Diggle et al. (2002) ou à Fitzmaurice et al. (2012).

I – 1. Modèles linéaires à effets mixtes

Les *modèles linéaires à effets mixtes* sont les modèles à effets mixtes les plus simples qui soient. Ils sont donc fréquemment utilisés pour l'analyse de données longitudinales. Un premier exemple historique et toujours très utilisé est l'analyse de variance ou ANOVA pour analysis of variance en anglais (Fisher, 2006; Scheffe, 1956). Cependant, l'analyse de variance suppose un jeu de données équilibré, *i.e.* un jeu de données dans lequel chaque individu est observé le même nombre de fois, ce qui est rarement le cas en pratique, en particulier dans le cas d'applications au domaine du médical.

En se basant sur les travaux d'Harville (1977), Laird et Ware (1982) proposent un modèle hiérarchique permettant de s'affranchir de cette contrainte. Étant données des observations multivariées $y_i \in \mathbb{R}^{k_i}$, Laird et Ware (1982) supposent qu'il existe, pour chacun des individus, deux matrices de design $H_i^\alpha \in \mathscr{M}_{k_i,p_\alpha}(\mathbb{R})$ et $H_i^\beta \in \mathscr{M}_{k_i,p_\beta}(\mathbb{R})$ relatives respectivement aux paramètres de population ou effets fixes $\alpha \in \mathbb{R}^p$ et aux paramètres individuels ou effets aléatoires $\beta_i \in \mathbb{R}^k$, telles que les observations soient une combinaison linéaire bruitée des effets fixes et aléatoires. Autrement dit, pour chaque sujet i, ils supposent que

$$y_i \;=\; H_i^\alpha \alpha + H_i^\beta \beta_i + \varepsilon_i$$

où ε_i est la réalisation d'une loi gaussienne multivariée, de dimension k_i, de moyenne nulle. De plus les effets aléatoires β_i sont supposés indépendants et identiquement distribués de loi normale. Ce modèle fournit donc un cadre d'étude très flexible et particulièrement adapté aux observations à données manquantes, comme cela est généralement le cas dans les sciences du vivant.

Un exemple de modèle-linéaire très utilisé pour l'analyse de données longitudinales scalaires est le *modèle avec pente et ordonnée à l'origine aléatoires* (Cohen et al., 1983). Soit un jeu de données obtenu par l'observation, pour chaque individu i, de k_i mesures $y_i = (y_{i,j})_{j\in[\![1,k_i]\!]}$ aux temps $t_i = (t_{i,j})_{j\in[\![1,k_i]\!]}$. Étant donné un temps de référence t_0, ce modèle permet d'estimer une droite traduisant la dynamique d'évolution de la population à l'échelle macroscopique ainsi que des droites d'évolutions individuelles. Plus

précisément, étant donnée la droite d'évolution moyenne $d\colon t \mapsto a(t-t_0)+b$, où a et b désignent respectivement la pente et l'ordonnée moyennes, on suppose que l'on observe des échantillons bruités des déformations affines de d, que sont, pour tout sujet i, les droites définies par $d_i\colon t \mapsto (a+a_i)(t-t_0)+(b+b_i)$:

$$\forall j \in [\![1, k_i]\!], \qquad y_{i,j} \;=\; (a+a_i)(t_{i,j}-t_0)+(b+b_i)+\varepsilon_{i,j}$$

où $\varepsilon_{i,j} \sim \mathcal{N}(0,\sigma)$ est un bruit blanc gaussien et où a_i et b_i sont les paramètres aléatoires du modèle et correspondent aux paramètres d'ajustements individuels de la pente et de l'ordonnée à l'origine. Ce modèle permet donc une certaine indépendance entre les différents individus, ce qui explique son succès dans les applications, par exemple en médecine personnalisée (Diaz et al., 2012) ou en écologie (Harrison et al., 2018).

Cependant, comme soulevé par Schiratti (2016) dans ses travaux de thèse, la nécessité d'un temps de référence t_0 en réduit considérablement le cadre d'application. En effet, dès lors qu'il n'existe pas de manière simple et sensée du point de vu de la modélisation d'imposer ce temps, celui-ci devra être estimé comme effet fixe du modèle, au même titre que les paramètres a et b. Problème : le modèle avec pente et ordonnée à l'origine devient alors non-identifiable. Il existe en effet une infinité de triplets (t_0, a, b) maximisant la vraisemblance du modèle, ce qui le rend inutilisable dans ce cas. Dans le cadre d'applications médicales, cette situation est en fait la plus courante : il n'existe en général pas de corrélation entre l'âge du patient et le stade de progression de la maladie étudiée. Pire, estimer ce temps se révèle être au cœur de la prise en charge médicale, ce dernier correspondant à un changement de l'état de santé du patient : échappement à un traitement en cas de chimiothérapie, entrée dans une nouvelle phase de progression d'une maladie, *etc.* Schiratti et al. (2015, 2017) ont proposé un modèle non-linéaire, dit *modèle générique spatio-temporel* pour pallier ce problème. Ce modèle est détaillé à la section III.2 du présent chapitre.

I – 2. Modèles non-linéaires à effets mixtes

La dépendance linéaire entre les effets et les observations peut bien évidement être levée, conduisant à l'ensemble des *modèles non-linéaires à effets mixtes*. Ces modèles ont historiquement été introduis par Sheiner et Beal (1980) et Bates et Watts (1988) dans le cadre de la modélisation pharmaco-cinétique. Toujours très utilisés dans ce cadre (Comets et al., 2010; Lavielle, 2014), ces modèles ont également prouvé leur applicabilité dans d'autres domaines tels que le suivi médical, comme expliqué en préambule. Ils font l'objet de recherches actives depuis les années 90 ; par soucis de concision, nous ne présentons ici que la version proposée par Lindstrom et Bates (1988).

De même que pour le cas linéaire, on suppose que l'on observe, pour chacun individu i, un couple (t_i, y_i) correspondant aux temps d'acquisition $t_i = (t_{i,j})_{j\in[\![1,k_i]\!]}$ et aux mesures $y_i = (y_{i,j})_{j\in[\![1,k_i]\!]}$ effectuées en ces temps. Soit f une fonction non-linéaire, et pour tout sujet i, $H_i^\alpha \in \mathscr{M}_{p_z,p_\alpha}(\mathbb{R})$ et $H_i^\beta \in \mathscr{M}_{p_z,p_\beta}(\mathbb{R})$ deux matrices de design relatives respectivement aux effets fixes $\alpha \in \mathbb{R}^{p_\alpha}$ et aléatoires $\beta_i \in \mathbb{R}^{p_\beta}$. On écrit alors, pour tout

individu i :

$$\forall j \in [\![1, k_i]\!], \quad y_{i,j} = f(z_i; t_{i,j}) + \varepsilon_{i,j} \qquad \text{où} \qquad z_i = H_i^{\alpha}\alpha + H_i^{\beta}\beta_i \in \mathbb{R}^{p_z},$$

où $\varepsilon_{i,j}$ est un bruit blanc gaussien et où les variables aléatoires β_i sont supposées indépendantes et identiquement distribuées selon une loi normale centrée. De manière évidente, les modèles linéaires décris ci-dessus sont des cas particuliers de modèles non-linéaires.

I – 3. Inférence statistique dans les modèles à effets mixtes

Outre la flexibilité du modèle introduit par Laird et Ware (1982), le succès des modèles à effets mixtes provient de leur grande faisabilité numérique due notamment au développement simultané des ordinateurs et d'algorithmes performants. En particulier, l'algorithme d'espérance-maximisation ou algorithme EM (Dempster et al., 1977) permet de maximiser la vraisemblance de modèles à données manquantes dans des cadres très généraux.

L'algorithme EM et ses variantes. Les modèles linéaires à effets mixtes sont un cadre d'application privilégié pour l'algorithme EM. Déjà, dans le papier séminal de Dempster et al. (1977) une attention particulière lui était consacrée et la tendance s'est poursuivie par la suite (Foulley, 2002; Laird et al., 1987; Laird et Ware, 1982; Meng et Van Dyk, 1997).

La convergence de l'algorithme EM vers un *minimum local* de la vraisemblance observée a été démontrée par Dempster et al. (1977) puis corrigée par Wu (1983). Cependant les hypothèses requises en première instance étaient difficiles à vérifier et le cadre introduit par Delyon et al. (1999) fournit des hypothèses plus raisonnables. L'algorithme itère deux étapes jusqu'à convergence : une étape d'espérance, l'étape E, dans laquelle on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées et une étape de maximisation, l'étape M, dans laquelle on estime le *maximum* de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E.

Outre des techniques pour accélérer sa convergence (McLachlan et Krishnan, 2007), plusieurs améliorations de l'algorithme EM ont été proposées. Globalement, on peut distinguer deux types d'améliorations : celles concernant l'étape d'espérance et celles concernant l'étape de maximisation. En ce qui concerne cette dernière étape, l'*EM généralisé* (Delyon et al., 1999) ne suppose plus une maximisation de l'espérance à chaque étape mais seulement un accroissement de celle-ci. Ainsi, on peut appliquer l'algorithme EM y compris en l'absence de solution analytique. Dans la version décrite par Lange (1995), l'étape de maximisation est réalisée par une méthode de Newton-Raphson.

Les solutions alternatives au calcul de l'espérance consistent en l'introduction d'une part de stochasticité au sein de la procédure d'estimation. Dans l'*EM stochastique* (SEM), Celeux et Diebolt (1985) proposent de remplacer le calcul de l'espérance par une évaluation numérique de celle-ci *via* une simulation des données manquantes. Wei et

Tanner (1990) généralisent cette idée et remplacent le calcul de l'espérance par une approximation de cette dernière par une méthode de Monte-Carlo, donnant lieu au *Monte-Carlo EM* ou MCEM. En modulant le nombre de tirages aléatoires dans la somme de Monte-Carlo (Celeux et al., 1995), on est en particulier à même de mimer un algorithme de type recuit simulé. Une approche alternative, développée par Delyon et al. (1999), consiste à remplacer le calcul de l'espérance par une approximation de type Robins-Monro (Robbins et Monro, 1951), que l'on sait converger vers l'espérance sous des hypothèses *ad hoc*. On parle alors d'*approximation stochastique de l'algorithme EM* ou d'algorithme SAEM. Enfin, contrairement à leurs homologues déterministes, ces variantes stochastiques de l'algorithme EM vont pouvoir s'extraire de *maxima* locaux. Ainsi la convergence vers des *maxima* globaux s'en trouve-t-elle favorisée.

Méthodes de Monte-Carlo par chaînes de Markov. Lorsque la simulation exacte des variables latentes n'est pas possible, recourir à une simulation approchée par une méthode de Monte-Carlo par chaînes de Markov, ou méthode MCMC pour Markov chain Monte Carlo en anglais, (Andrieu et al., 2003; Brooks et al., 2011; Robert et Casella, 1999) s'avère fructueux. L'idée des méthodes MCMC est de générer une chaîne de Markov convergeant vers la loi cible contre laquelle on veut réaliser la simulation. En particulier, on remplace la génération d'un échantillon selon cette loi compliquée par un nombre, potentiellement élevé, de simulations contre des distributions que l'on espère plus simples. Parmi ces échantillonneurs, le plus usité est sûrement l'*algorithme de Métropolis-Hastings*. D'abord introduit dans le cas particulier de la distribution de Boltzmann (Metropolis et Ulam, 1949; Metropolis et al., 1953), l'algorithme de Metropolis-Hastings est étendu au cas de n'importe quelle distribution par Hastings (1970). Une des forces de cet algorithme est qu'il ne nécessite la connaissance de la loi cible qu'à une constante multiplicative près uniquement. Ainsi, on s'évite le calcul de la constante de normalisation, calcul la plupart du temps impraticable dans les faits. L'algorithme de Metropolis-Hastings peut être vu comme une généralisation de la méthode de rejet : à chaque itération, et étant donné l'état de la chaîne à cette instant, on fait une proposition d'incrément et on accepte celui-ci s'il améliore la « vraisemblance ». Plus précisément, étant donné une loi cible π, un générateur (pseudo) aléatoire $q(\cdot; x_k)$ et x_k l'état de la chaine à l'itération k, on accepte la proposition $x^* \sim q(\cdot; x_k)$ avec probabilité

$$\alpha(x_k, x^*) \;=\; \frac{\pi(x^*)\, q(x_k, x^*)}{\pi(x_k)\, q(x^*, x_k)}\,.$$

De plus, cet échantillonneur peut-être incorporé dans un échantillonneur de Gibbs et est donc en particulier utilisable pour des données de grande dimension.

En se basant sur les travaux de Kuhn et Lavielle (2004) qui établissent la convergence du MCMC-SAEM dans le cas où les variables générées durant la procédure d'estimation restent bornées, Allassonnière et al. (2010) démontrent la convergence du MCMC-SAEM en toute généralité. Signalons également que la convergence de cet algorithme ne nécessite qu'une seule étape de MCMC, ce qui le rend très compétitif sur le plan numérique. Cet algorithme fait l'objet du logiciel Monolix et a démontré sa grande applicabilité,

notamment pour des modèles de pharmaco-cinétiques (Chan et al., 2011; Lavielle et Mentré, 2007).

Malgré la flexibilité des modèles décris ici (linéaires en I.1 et non-linéaires en I.2), de par leur écriture, ceux-ci ne s'appliquent qu'à des données scalaires. Or, les applications, et particulièrement les applications aux sciences médicales que nous avons à cœur, suggèrent l'usage de données hautement structurées telles que des scanners, images, tenseurs ou formes anatomiques 3D. Il apparait donc comme nécessaire de proposer un cadre d'étude statistique compatible avec ces données à la fois massives et hétérogènes.

II. De l'usage de la géométrie riemannienne pour l'étude de données longitudinales

Dans cette section, et sauf mention contraire, nous appellerons de manière générique *forme* toute donnée structurée. Ainsi une forme peut désigner aussi bien une image, qu'un maillage, un tenseur, une sous-variété, *etc.*

La géométrie riemannienne se révèle un outil particulièrement adapté à la modélisation mathématique des formes. En effet, plutôt que d'analyser les formes en elle-mêmes de manière individuelle, il semble plus efficient de considérer des ensembles ou populations de formes et d'essayer de les appréhender comme des espaces au sens mathématique du terme (Trouvé et Younes, 2015). De par leur construction, ces espaces vont hériter naturellement d'une structure de variété riemannienne.

II – 1. Espaces de forme

Le premier exemple d'*espace de forme* (bien que non-conceptualisé comme tel à l'époque) remonte aux travaux de D'Arcy Thompson (1942). Son idée est alors de comparer non pas les formes anatomiques de manière intrinsèque mais plutôt de quantifier les déformations nécessaires pour passer d'une forme anatomique à une autre comme illustré à la figure 1.1. Cette idée va être reprise et formalisée par Grenander (1993), donnant ainsi naissance à l'anatomie computationelle.

L'idée est la suivante : étant donné un ensemble M de formes vivant dans $\mathbb{R}^n$, on va construire son espace de forme correspondant en faisant agir (de manière transitive) un *groupe de déformations* $\mathcal{G}$ sur l'ensemble M et en considérant l'unique[1] orbite de cette action $\mathcal{G} \cdot x_0$ où $x_0 \in M$. Ainsi toutes les formes d'un espace de forme peuvent être obtenues en déformant x_0 via un élément de $\mathcal{G}$. On appelle x_0 forme modèle ou

[1] L'unicité de l'orbite provenant de la transitivité de l'action. Sinon, on peut se restreindre à une orbite particulière.

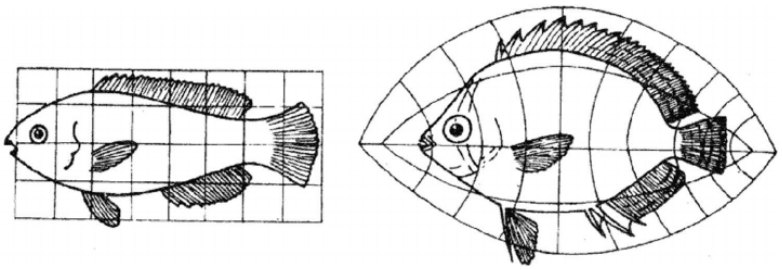


Figure 1.1 – *Espace de forme de D'Arcy Thompson.*
Déformation géométrique de deux poissons. Illustration tirée du livre *On Growth and Form* de D'Arcy Thompson (1942).

template. En fait de groupe de déformations $\mathcal{G}$, on choisira le plus souvent le groupe $\mathscr{C}^1(\mathbb{R}^n)$ des $\mathscr{C}^1$-difféomorphismes de $\mathbb{R}^n$, mais on pourrait se restreindre à n'importe quel sous-groupe des bijections de $\mathbb{R}^n$ dans lui-même. L'exemple le plus simple d'espaces de forme est celui des points de repères ou *landmarks*

$$M \;=\; \{x = (x_1, \dots, x_p) \in (\mathbb{R}^n)^p \mid \forall i, j \in [\![1, p]\!],\; i \neq j,\; x_i \neq x_j\}$$

équipé du groupe $\mathcal{G} = \mathscr{C}^1(\mathbb{R}^n)$ et où l'action de $\mathcal{G}$ sur M est donnée par

$$\forall\, g \in \mathcal{G}, \quad \forall\, x \in M, \qquad g \cdot x = \big(g(x_1), \dots, g(x_p)\big)\,.$$

Cet espace a été introduit et étudié par Kendall (1984) principalement. Un autre exemple classique d'espaces de forme est celui des images, à savoir $M = \mathbb{L}^2(\Omega, \mathbb{R})$ où $\Omega \subset \mathbb{R}^n$.

Modèles de déformations. En faisant agir un tel groupe sur la variété riemannienne des formes, on peut munir cette dernière d'une distance permettant de quantifier le coût de déformation d'une forme à une autre (Younes, 2010). Plus précisément, si on suppose $\mathcal{G}$ muni d'une distance invariante à droite $d_{\mathcal{G}}$, alors on peut munir M de la pseudo-distance définie par

$$\forall\, x, y \in M, \quad d_M(x, y) \;=\; \inf_{g \in \mathcal{G}} \{d_{\mathcal{G}}(Id, g) \mid g \cdot x = y\}\,.$$

Ainsi, pour munir M d'une distance à même de quantifier le coût de déformation, il « suffit » de munir $\mathcal{G}$ d'une distance invariante à droite. Dans ce but, Dupuis et al. (1998) et Beg et al. (2005) ont introduit le concept de *grandes déformations* ou LDDMM pour large deformation diffeomorphic metric mapping en anglais. L'idée des grandes déformations est de permettre, comme leur nom l'indique, des déformations de l'espace grandes mais en gardant un contrôle, notamment en terme de régularité, sur celles-ci. Pour cela, on va restreindre le groupe des déformations $\mathcal{G}$ à un sous groupe $\mathcal{G}_V \subset \mathcal{G}$ de déformations qui se comportent bien, en un sens à préciser. On rappelle ici très brièvement et sans démonstration les prémices du cadre des grandes déformations. On en trouvera une construction claire et détaillée dans Younes (2010) et Glaunès (2005) par exemple.

Soit V un ensemble de champs de vecteurs sur $\mathbb{R}^n$, dont la norme définit le coût de déformation. On suppose d'une part que V peut être muni d'une structure hilbertienne et qu'il s'injecte continûment dans l'espace $\mathscr{C}^1_0(\mathbb{R}^n)$ des difféomorphismes v nuls et de différentielle $\mathrm{d}v$ nulle à l'infini. Intuitivement, on voudrait quantifier le coût d'appariement entre deux formes x et y *via* le champ de déplacement $v - Id$ tel que $v(x) = y$ avec l'idée que le morphisme Id correspond à la situation idéale où aucune déformation n'est nécessaire. Cependant, dès lors que la déformation devient trop importante, *i.e.* dès lors que ϕ devient grand en norme devant Id, un tel morphisme n'est plus inversible. Or, pour des questions de symétries évidentes, on souhaite que ces transformations soient inversibles. Afin de contrôler l'amplitude des déformations, l'idée de Beg et al. (2005) est de procéder infinitésimalement. Soit $\mathbb{L}^2_V = \mathbb{L}^2([0,1], V)$ l'ensemble des champs de vecteurs $v = (v_t)_{t\in[0,1]}$ paramétrés par le temps t et $\mathbb{L}^2$-intégrables par rapport à t, *i.e.* tels que

$$\int_0^1 \|v_t\|^2_{1,\infty}\,\mathrm{d}t \;=\; \int_0^1 \left(\|v_t\|_\infty + \|\mathrm{d}v_t\|_\infty\right)^2 \mathrm{d}t \;<\; +\infty\,.$$

On pose alors (et on vérifie que l'on définit bien un groupe (Younes, 2010))

$$\mathcal{G}_V \;=\; \{\,\phi_1^v \,|\, v \in \mathbb{L}^2_V\,\}$$

où ϕ_1^v est le flot au temps $t = 1$ associé à l'équation différentielle

$$\partial_t \phi_t^v \;=\; v_t \circ \phi_t^v \qquad ; \qquad \phi_0^v \;=\; Id\,. \tag{1.1}$$

Cette équation est le pendant infinitésimal du champ de déplacement $v - Id$. Ainsi, en minimisant la norme des différents champs de vecteurs v_t, on minimise à chaque instant la déformation et donc la déformation globale, qui est donnée par ϕ_v^1. On attire l'attention du lecteur sur le fait que l'application $t \mapsto \phi_t^v$ n'est *a priori* pas à valeurs dans $\mathbb{R}^d$ mais dans la variété riemannienne des formes M. Aussi lui préférera-t-on le plus souvent son écriture intégrale où $\int$ désigne l'intégrale de Bochner (Bochner, 1933) qui généralise aux espaces de Banach la notion d'intégrale de Lebesgue :

$$\forall t \in \mathbb{R}, \quad \forall x \in M, \quad \phi_t^v(x) \;=\; x + \int_0^t v_t \circ \phi_s^v(x))\,\mathrm{d}s\,.$$

À noter que $\phi_1^v \in \mathscr{C}^1(\mathbb{R}^n)$ et que l'équation différentielle ci-dessus admet une unique solution pour toute condition initiale (Younes, 2010). Enfin, on pose pour tout $\phi, \phi' \in \mathcal{G}_V$

$$d(Id, \phi) \;=\; \inf_{v\in\mathbb{L}^2_V} \left\{ \left(\int_0^1 \|v_t\|^2_V\,\mathrm{d}t\right)^{1/2} \,\middle|\, \phi_1^v = \phi \right\} \quad \text{et} \quad d_{\mathcal{G}}(\phi, \phi') \;=\; d(Id, \phi' \circ \phi^{-1})\,.$$

On peut montrer que ce minimum est atteint pour un certain v de $\mathbb{L}^2_V$, *i.e.* qu'il existe un champ de vecteurs v permettant de transporter une configuration de l'espace de forme sur une autre et pour lequel le coup de déformation est minimal (Younes, 2010). De plus, la distance $d_{\mathcal{G}}$ est invariante à droite et, finalement, le coût de déformation pour passer d'une forme x à une forme y est donné par

$$\forall\, x, y \in M, \quad d_M(x, y) \;=\; \inf_{v\in\mathbb{L}^2_V} \left\{ \left(\int_0^1 \|v_t\|^2_V\,\mathrm{d}t\right)^{1/2} \,\middle|\, \phi_1^v \cdot x = y \right\}.$$

Appariement de formes. Ainsi, du fait de la discussion précédente, dès lors que l'on sait recaler deux formes l'une sur l'autre, on est à même de quantifier la différence entre elles. En fait de recalage exact, on lui préfère un recalage inexact car plus réaliste dans le cadre de l'application à l'anatomie computationnelle. Autrement dit l'appariement de deux formes est réalisé *via* la minimisation du coût de déformation explicité ci-dessus, pénalisé par un terme d'attache aux données A. Formellement, le problème d'appariement pour A consiste à minimiser sur $\mathbb{L}^2_V$ la fonctionnelle définie par

$$J\colon v \;\mapsto\; \frac{1}{2}\int_0^1 \|v_t\|_V^2 \,\mathrm{d}t \;+\; \lambda\, A(\phi_1^v) \tag{1.2}$$

où $\lambda > 0$. L'existence d'un tel minimiseur a été démontré par Glaunès (2005) dans ses travaux de thèse.

Par suite, étant donné le formalisme des grandes déformations, dès lors que l'on est à même de définir une représentation adaptée pour les-dites formes et un terme d'attache aux données, on va pouvoir quantifier la différence entre deux formes. Dans le cas des landmarks où les points sont labellisés, le terme d'attache aux données consiste en une simple distance $\mathbb{L}^2$:

$$\forall\, x = (x_1, \ldots, x_p),\ y = (y_1, \ldots, y_p) \in M\,, \quad A\,(x, y) = \sum_{i=1}^{p} \|x_i - y_i\|_2^2$$

où $\|\cdot\|_2$ désigne la norme 2 de $\mathbb{R}^n$. De nombreux travaux, que l'on ne détaille pas ici mais sur lesquels on revient partiellement au chapitre 5, ont été réalisés dans ce sens. De manière non-exhaustive, on peut par exemple citer les travaux de Glaunès et al. (2004) pour l'appariement de points non-labellisés *via* l'appariement de mesures discrètes, de Glaunès et al. (2008) pour l'appariement de courbes, de Vaillant et Glaunès (2005) et Durrleman (2010) pour l'appariement de surfaces orientées *via* l'appariement de courants, de Charon et Trouvé (2013) pour l'appariement de formes non-orientées *via* les varifolds, de Roussillon et Glaunès (2016) pour l'appariement de formes non-orientées également mais *via* les cycles normaux ou, enfin, de Charlier et al. (2017) pour l'analyse de formes fonctionelles, *i.e.* de fonctions ou de signaux définis sur des supports géométriques variables.

Principe de réduction de dimension pour les problèmes d'appariement. Finalement, le dernier point qu'il nous reste à expliciter concerne le choix d'une norme appropriée pour l'espace de Hilbert V. Pour ce faire, nous allons munir V d'une structure d'espace de Hilbert à noyau reproduisant ou RKHS pour reproducing kernel Hilbert space en anglais. Les RKHS, introduits par (Aronszajn, 1950), ont en effet su prouver leur utilité pratique ces dernières années, notamment en apprentissage automatique (Friedman et al., 2001).

Soit k_V un noyau reproduisant sur V. Alors, on munit V d'une norme en posant

$$\forall v \in V, \qquad \|v\|_V^2 \;=\; \int_{\mathbb{R}^n} \Big\langle Lv(x) \,\Big|\, v(x) \Big\rangle \,\mathrm{d}x$$

où $L\colon V \to V^*$ est un opérateur différentiel tel que $K = L^{-1}$ avec

$$\forall v \in V, \quad \forall x \in M, \qquad (Kv)(x) = \int_{\mathbb{R}^n} k_V(x,y)\, v(y)\, \mathrm{d}y\,.$$

On cherche à apparier deux formes x et y. En se basant sur les travaux de Joshi et Miller (2000) pour l'appariement de landmarks, Durrleman et al. (2011) ont proposé une représentation parcimonieuse du groupe des déformations *via* l'introduction de points de contrôle. Supposons que l'on peut associer à chacune des deux formes une configuration de N_c points distincts, que l'on note abusivement x et y. Soit un ensemble $c(t) = \{c_1(t), \ldots, c_{N_c}(t)\}$ composé de N_c points distincts sur M, dit *points de contrôle* et variant au cours du temps. Soit un terme d'attache aux données A qui mesure l'attache à la forme y et soit la famille de champs de vecteurs $(v_t)_t$ régissant la dynamique de $t \mapsto c(t)$, *i.e.* telle que

$$\forall i \in [\![1, N_c]\!], \qquad \frac{\mathrm{d}c_i}{\mathrm{d}t} = v_t(c_i) \quad \text{et} \quad c_i(0) = x_i\,. \tag{1.3}$$

Alors (Joshi et Miller, 2000), il existe une famille de N_c vecteurs $\alpha(t)$ dépendant du temps telle que la famille de champs de vecteurs réalisant l'appariement entre x et y, à savoir minimisant la fonctionnelle J définie en (1.2), est donnée par

$$\forall t \in \mathbb{R}, \quad \forall x \in M, \qquad v_t(x) \;=\; \sum_{j=1}^{N_c} k_V\big(x, c_j(t)\big)\, \alpha_j(t)\,.$$

Les vecteurs $\alpha(t)$ sont appelés *moments* associés aux points de contrôle $c(t)$. En particulier, en se basant sur des résultats de mécanique hamiltonienne et en utilisant que pour tout i on a $c_i(0) = x_i$, J se ré-écrit (Allassonnière et al., 2005; Miller et al., 2006)

$$J\big(\alpha(0)\big) \;=\; \sum_{i=1}^{N_c}\sum_{j=1}^{N_c} \Big\langle k_V(x_i, x_j)\, \alpha_j(0) \,\Big|\, \alpha_i(0) \Big\rangle + \lambda\, A\big(c(1)\big)\big) \tag{1.4}$$

et on a remplacé l'estimation d'une famille de champs de vecteurs par l'estimation d'un nombre fini de moments initiaux. En particulier, $c(0)$ étant fixé, on interprète $c(1)$ comme une fonction du moment initial $\alpha(0)$.

Finalement, la minimisation de la fonctionnelle d'énergie J est réalisée par *tirs géodésiques* (Allassonnière et al., 2005; Miller et al., 2006) : jusqu'à convergence de l'algorithme, se succèdent une étape d'intégration de l'équation (1.3) à moments initiaux $\alpha(0)$ fixés et une étape de minimisation par descente de gradient des moments initiaux $\alpha(0)$ à points de contrôle finaux $c(1)$ fixés. Une interprétation schématique de cet algorithme est présenté à la figure 1.2.

On dispose ainsi d'un cadre mathématique flexible et rigoureux pour l'étude des formes : la géométrie riemannienne. Comme expliqué précédemment, les modèles à effets mixtes tels que décrits à la section I ne s'appliquent pas en l'état sur des variétés riemanniennes (non euclidiennes). Des généralisations de ces modèles ont toutefois été proposées dans la littérature et font l'objet du paragraphe suivant.

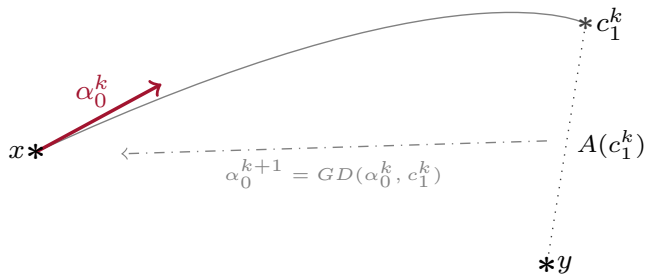


Figure 1.2 – *Algorithme de tirs géodésiques.*
Dans cette figure, on note α_0^k les moments initiaux $\alpha(0)$ à l'itération k, et c_1^k la configuration de points de contrôle finale. L'algorithme de tirs géodésiques alterne une étape de descente de gradient et une étape d'intégration jusqu'à convergence.

II – 2. Modèles de régression géodésique

On a vu précédemment, au paragraphe I.1, que le modèle de pente et ordonnée à l'origine aléatoires était très utilisé dans le cadre euclidien. Dans ce modèle, on suppose que l'évolution de la population à l'échelle macroscopique peut s'expliquer par une droite et que les évolutions individuelles sont des déformations (linéaires) de cette droite représentative. Si on munit $\mathbb{R}^n$ de son produit scalaire usuel, on suppose en fait que l'évolution moyenne de la population suit une trajectoire géodésique.

Approche par les formes. Dans sa forme la plus simple, *i.e.* si on annule tous les effets aléatoires, on retrouve le classique modèle de régression linéaire, également très usité. Fletcher (2011) propose de généraliser *in extenso* ce modèle en supposant une trajectoire moyenne géodésique, conduisant à un modèle de *régression géodésique*. Plus précisément, étant donné un point $p \in M$ de la variété riemannienne et un vecteur $v \in TM$ du fibré tangent de M, on écrit pour toute observation $y_i \in M$ associée au scalaire $x_i \in \mathbb{R}$:

$$y_i \;=\; \mathcal{E}xp\left(\mathcal{E}xp\left(p\,;x_i v\right);\,\varepsilon\right)$$

où ε est une variable aléatoire à valeurs dans l'espace tangent à M en le point $\mathcal{E}xp\,(p, x_i v)$ et où $\mathcal{E}xp\,(p, v) = \mathcal{E}xp_p(v)$ désigne l'exponentielle riemannienne au point p, de vecteur tangent v, *i.e.* la valeur au temps 1 de l'unique géodésique passant par p au temps 0 avec pour vitesse initiale v (Gallot et al., 2004; Jost, 2002). Ce modèle ne possède cependant pas d'expression explicite de la vraisemblance. L'estimation des paramètres est donc réalisée via la minimisation d'un critère de moindres carrés. Kim et al. (2014) proposent une généralisation de ce modèle pour de la régression multivariée.

En se basant sur le modèle de régression géodésique et le modèle de Laird et Ware (1982), Muralidharan et Fletcher (2012) introduisent un *modèle génératif et hiérarchique* pour l'analyse de données longitudinales à valeurs sur des variétés riemanniennes. Comme

illustré à la figure 1.3, chaque individu est supposé suivre une trajectoire géodésique, elle-même considérée comme une perturbation aléatoire d'une trajectoire moyenne également géodésique. Plus précisément, à supposer que l'on observe pour chaque individu le couple (x_i, y_i) où $x_i = (x_{i,j})_{j\in[\![1,k_i]\!]} \in \mathbb{R}^{k_i}$ et $y_i = (y_{i,j})_{j\in[\![1,k_i]\!]} \in M^{k_i}$,

$$\forall j \in [\![1, k_i]\!], \quad y_{i,j} \;=\; \mathcal{E}xp\left(\mathcal{E}xp\left(p_i\,;x_{i,j}v_i\right);\varepsilon_{i,j}\right)$$

où on a gardé les mêmes notations que précédemment. Les effets aléatoires (p_i, v_i) sont alors vus comme une perturbation aléatoire d'une courbe géodésique sur le fibré tangent de M. Pour ce faire, Muralidharan et Fletcher (2012) proposent de munir le fibré tangent de M d'une métrique riemannienne : la métrique de Sasaki. Soit les effets fixes $(\alpha, \beta) \in TM$ et une famille de vecteurs tangents à TM en ce point (q_i, w_i). Alors, le modèle proposé s'écrit

$$(p_i, v_i) \;=\; \mathcal{E}xp_{\mathcal{S}}\left((\alpha,\beta)\,;(q_i,w_i)\right)$$

où $\mathcal{E}xp_{\mathcal{S}}$ désigne l'exponentielle riemannienne associée à la métrique de Sasaki, sur le fibré tangent TM de M. L'estimation des paramètres est, comme pour la régression géodésique, réalisée par moindres carrés.

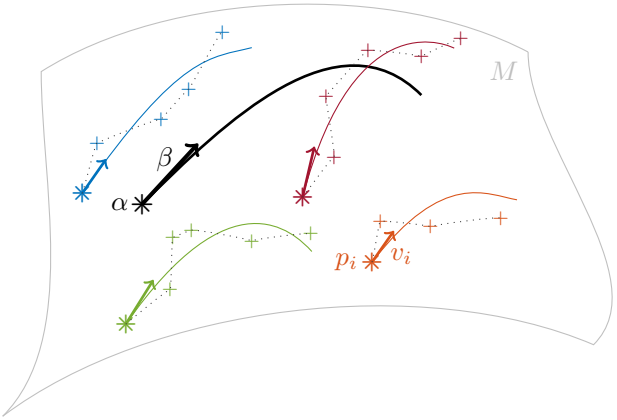


Figure *1.3* – *Modèle hiérarchique géodésique de Muralidharan et Fletcher (2012).* On observe des échantillons bruités le long de trajectoires géodésiques, construites comme des perturbations aléatoires d'une trajectoire représentative géodésique.

Approche par les difféomorphismes. Au paragraphe II.1 nous avons expliqué le lien étroit entre groupes de déformations et espaces de forme. En s'appuyant sur ce lien, Singh et al. (2013, 2014) introduisent un *modèle hiérarchique géodésique pour les difféomorphismes*. Ce modèle repose sur l'estimation d'une trajectoire géodésique de difféomorphismes à l'échelle de la population. On peut en effet munir le groupe des $\mathscr{C}^1$-difféomorphismes de $\mathbb{R}^n$ d'une métrique adaptée, le rendant ainsi riemannien (Arnold, 1966; Miller et al., 2006). Au niveau individuel, les trajectoires consistent en des portions

de géodésiques sur la variété des formes M. Ces courbes étant supposées géodésiques, elles sont entièrement paramétrées par leur valeur $p_i(0)$ et vitesse $v_i(0)$ au premier temps d'observation (Gallot et al., 2004). Ainsi, les effets aléatoires du modèle proposé résident en le couple $(p_i(0), v_i(0))$ et on les estime par une méthode de moindres carrés en imposant que $p_i(t) = \phi_i(t) \cdot p_i(0)$ soit proche des observations correspondantes et où $\phi_i(t)$ est un difféomorphisme émanant de la trajectoire moyenne. En particulier, les effets aléatoires pour ce modèle dépendent fortement du premier temps d'observation et modifier ce temps modifie d'autant la valeur estimée pour ces effets. Pour autant, le premier temps d'acquisition n'a généralement que peu de sens en terme de modélisation et un modèle robuste aux changements d'origine temporelle semble plus adapté en pratique.

Ce modèle a montré son efficacité sur des données images (Singh et al., 2016). Plus généralement, l'amélioration des techniques d'imagerie médicale a suscité un vif intérêt ces dix dernières années et l'établissement de modèles permettant de traiter efficacement ce type de données a été au cœur de nombreux travaux. On peut par exemple citer ceux de Datar et al. (2009), Durrleman et al. (2013) ou Fitzmaurice et al. (2012).

Par delà les images, des modèles pour permettre l'étude de formes 2D et 3D encodées par des landmarks (*cf.* paragraphe II.1) ont également été développés. Par exemple, Trouvé et Vialard (2012) ont proposé un modèle de régression non-paramétrique par spline, reposant sur l'introduction d'une perturbation aléatoire dans les équations hamiltoniennes régissant le flot géodésique. En se basant sur les travaux de Cates et al. (2007), Datar et al. (2012) ont développé une nouvelle manière de dériver des modèles pour l'étude de formes qui se repose sur l'alternance de deux étapes de minimisation : minimisation d'une fonctionnelle de coût quantifiant l'entropie du système afin de maximiser la correspondance entre les formes et estimation des paramètres d'un modèle linéaire à effets mixtes. Ces dernier travaux ont spécifiquement donné lieu au développement du logiciel open source Seg3D, attestant de leur applicabilité.

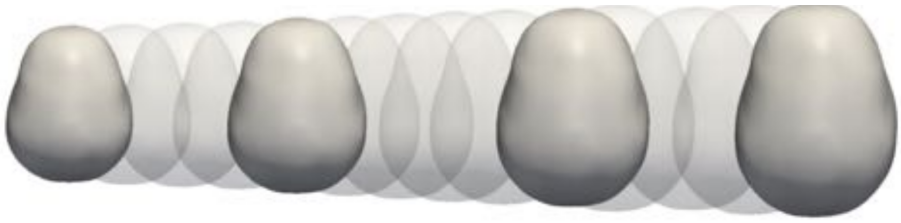


Figure *1.4* – *Un exemple de régression de formes.*
Illustration tirée de Fishbaugh et al. (2017). Trajectoire (en transparent) obtenue par régression des quatre formes (en traits pleins) observées.

Bien que performants, les modèles sus-cités sont peu versatiles : ils ne peuvent traiter qu'un seul type de données ; en l'occurrence des images ou des landmarks. Afin de pallier cette limitation, Fishbaugh et al. (2017) ont proposé un modèle de *régression géodésique de formes* dans le cadre général des grandes déformations et en se basant sur la représentation parcimonieuse du groupe des déformations, *via* les points de contrôle,

introduite par Durrleman et al. (2011, 2013). Soit k_V un noyau reproduisant sur V. On reprend les notations du paragraphe traitant du principe de réduction, page 13.

Le but de ce modèle est d'estimer une trajectoire continue, et géodésique, de formes à partir d'une séquence finie de formes $(O_{t_i})_i$ observées en les temps $t_i \in [t_0, T]$. On interprète cette trajectoire comme la déformation d'une forme template X_0 par un difféomorphisme ϕ_t suffisamment régulier et satisfaisant l'équation de flot géodésique (1.1) : à chaque instant, $X_t = \phi_t(X_0)$. Cette trajectoire étant supposée géodésique, le morphisme ϕ_t minimise en particulier la fonctionnelle d'énergie J définie en (1.2). Le principe de réduction vu précédemment nous assure que l'on peut paramétriser le flot ϕ_t par un nombre fini de points de contrôle et de moments, à savoir les points de contrôle et moments initiaux $(c(0), \alpha(0))$ associés à l'équation de tirs géodésiques (1.3). De plus, ϕ_t satisfait l'équation de flot géodésique (1.1) où l'on peut exprimer v_t en fonction de k_V (1.4). Ainsi, pour résumer, il existe deux fonctions F et G qui dépendent du noyau k_V et telles que

$$\big(\dot{c}(t), \dot{\alpha}(t)\big) \;=\; F\big(c(t), \alpha(t)\big) \qquad \text{et} \qquad \dot{X}(t) \;=\; G\big(X(t), c(t), \alpha(t)\big)\,.$$

De plus, en imposant une contrainte de régularité sur ϕ_t ne dépendant que de $(c(0), \alpha(0))$, on peut ré-écrire la fonctionnelle J comme une fonction dépendant de $(X(0), c(0), \alpha(0))$ uniquement. On est alors ramené à la même situation que pour le tir géodésique, *i.e.* à savoir celle de minimiser une fonctionnelle en alternant des étapes d'intégration (éventuellement dans le passé) et de descente de gradient. L'algorithmique détaillée est présentée dans Fishbaugh et al. (2017) et appliquée à la régression de formes anatomiques 3D.

En particulier, dans la construction décrite ci-dessus, on ne fait aucune hypothèse sur la nature de l'espace de forme et le modèle proposé par Fishbaugh et al. (2017) permet de traiter des formes de nature variée (courants, varifolds, cycles normaux, *etc.*) comme escompté. De plus, l'introduction des points de contrôle réduit le nombre de paramètres à estimer, tout en continuant à autoriser des déformations importantes. D'un point de vue statistique, l'intérêt est grand : en règle générale, on a accès à un nombre réduit d'observations et il convient donc de réduire d'autant la dimension de l'espace des paramètres que l'on souhaite estimer pour ne pas subir trop intensément les effets du fléau de la dimension (Giraud, 2014).

Le modèle décrit ci-dessus fait parti du logiciel open-source Deformetrica, dont on peut trouver une description unifiée dans Bône et al. (2018). De plus, nous présentons à la figure 1.4, tirée de Fishbaugh et al. (2017), ce que l'on entend par régression de formes, à savoir l'estimation d'une trajectoire *continue* de formes qui permet d'expliquer au mieux les formes observées.

Ceux de Muralidharan et Fletcher (2012) et Singh et al. (2013, 2014) mis à part, les modèles présentés ici ne permettent malheureusement pas d'expliquer la variabilité entre les différents individus. Or, dans l'optique de faire de la prédiction à partir des données collectées, cette variabilité se doit d'être expliquée. Autrement dit, on ne peut

pas se contenter de modèle de régression géodésique pour expliquer complètement un jeu de données longitudinales.

De plus, on souhaite donner un sens aux paramètres estimés. Notamment, et comme expliqué au paragraphe I.1, le rythme de progression est informatif dans la pratique et tout particulièrement en cas d'application médicale. Ainsi, on veut pouvoir permettre à chacun des sujets d'emprunter sa courbe d'évolution à son propre rythme, *i.e.* à une vitesse et avec temps de référence qui lui sont propres. Autrement dit, dans un but applicatif, chaque sujet devrait pouvoir emprunter sa trajectoire d'évolution avec sa propre paramétrisation temporelle. Les modèles de Muralidharan et Fletcher (2012) et Singh et al. (2013, 2014) en forçant une dépendance forte des trajectoires sujets-spécifiques en le premier temps d'acquisition n'autorisent pas de telles reparamétrisations.

III. Modèles spatio-temporels pour l'étude de données longitudinales

En pratique, les différents sujets ne sont presque jamais observés le même nombre de fois et encore moins aux mêmes âges. Ainsi, on a besoin de modèles robustes à ce type de limitations.

III – 1. Déformations spatio-temporelles

Pour palier ces limitations, Durrleman et al. (2009) décident de discriminer les déformations spatiales, liées à la géométrie intrinsèque des formes observées, des déformations temporelles, liées aux contraintes d'acquisition. Pour ce faire, ils introduisent la notion d'*atlas spatio-temporel* que l'on décrit ci-après.

Soit le jeu de données consistant en l'observation de N sujets $(S^i)_{i\in[\![1,N]\!]}$ aux temps $(t^i_j)_{i\in[\![1,N]\!],j\in[\![1,k_i]\!]}$. L'idée de Durrleman et al. (2009) est d'estimer une trajectoire continue $M_t = \chi_t(M_0)$ de formes telle que chacune des observations corresponde à l'évaluation en un temps donné d'une déformation spatio-temporelle de ce scénario moyen. La trajectoire moyenne est obtenue par régression à partir d'une forme template M_0 sur laquelle est imposée des conditions de régularité notées $\mathrm{Reg}(\chi)$. De plus, pour chaque individu, on se donne un couple de déformations (ψ^i, ϕ^i), l'une représentant une déformation de type spatiale et l'autre de type temporelle, de sorte à mettre en correspondance $\phi^i\left(M(\psi^i(t^i_j))\right)$ et $S^i(t^i_j)$, à savoir la déformation spatio-temporelle de la trajectoire

moyenne M au temps t_j^i et l'observation $S^i(t_j^i)$. Par suite, on pose :

$$J(\psi^i, \phi^i, \chi, M_0) \; = \\ \sum_{i=1}^{N} \left\{ \sum_{t_j^i} d\left(\phi^i \left(\chi_{\psi^i(t_j^i)}(M_0) \right) \, , \, S^i(t_j^i) \right)^2 + \gamma^{\phi} \mathrm{Reg}(\phi^i) + \gamma^{\psi} \mathrm{Reg}(\psi^i) + \gamma^{\chi} \mathrm{Reg}(\chi) \right\} \quad (1.5)$$

où $\mathrm{Reg}(\psi^i)$ et $\mathrm{Reg}(\phi^i)$ représentent des contraintes de régularité sur ψ^i et ϕ^i. Pour minimiser cette fonctionnelle, Durrleman et al. (2009) procèdent comme dans le modèle de grandes déformations et interprètent les morphismes ϕ^i et ψ^i comme les flots d'équations différentielles *ad hoc* et dont les champs de vecteurs sous-jacents appartiennent à un RKHS. Ainsi, ils remplacent l'estimation de morphismes par l'estimation d'une famille de moments associés à des noyaux reproduisants.

Ce modèle est le premier à notre connaissance à découpler déformations spatiales et temporelles. De plus, il permet d'expliquer fidèlement des données massives et hétérogènes telles que des formes anatomiques. Cependant, c'est un modèle uniquement descriptif et non-explicatif : en effet, les déformations considérées ne dérivant pas d'un modèle statistique, les connaissances apprises au cours de la procédure d'optimisation ne nous permettent ni de prédire l'évolution d'un sujet donné dans le futur, ni de déduire les trajectoires d'évolution des sujets les unes à partir des autres. Dans la même idée, Yang et al. (2011) et Delor et al. (2013) introduisent la notion de décalages temporels. Ces modèles étant tous deux non-génératifs, ils souffrent néanmoins des mêmes limitations que le modèle de Durrleman et al. (2009).

Lorenzi et al. (2015) proposent une approche un peu différente dans le cadre de l'étude du vieillissement du cerveau : leur idée est de proposer un modèle capable de différencier un vieillissement normal dû à l'age « anatomique » du patient d'un vieillissement anormal et donc pathologique. Cependant, comme pour les modèles sus-cités, les paramètres encodant le décalage entre vieillissement naturel et vieillissement pathologique ne dérivent pas d'un modèle statistique. Ce modèle est donc uniquement descriptif. De plus, cette idée est très spécifique au contexte du vieillissement cérébral et ne s'exporte pas facilement à d'autres cadres d'application.

Dans le but d'expliquer la dépendances des trajectoires individuelles les unes aux autres, Durrleman et al. (2013) généralisent le modèle de Durrleman et al. (2009) pour en faire un modèle *génératif*. Plus précisément et avec les mêmes notations que précédemment, ils supposent que la formulation du terme d'attache aux données de J peut être comprise comme $-\log p(S^i(t_j^i))$ où la loi $p(S^i(t_j^i))$ serait une loi normale centrée en $\phi^i \left(\chi_{\psi^i(t_j^i)}(M_0) \right)$ sur la variété, *i.e.* de la forme

$$\frac{1}{2} \exp\left(-\frac{1}{2} d\left(\phi^i \left(\chi_{\psi^i(t_j^i)}(M_0) \right) \, , \, S^i(t_j^i) \right) \right) .$$

La fonctionnelle à estimer reste inchangée et les auteurs utilisent la même technique de passage par les RKHS pour l'estimation des différents morphismes.

La difficulté d'estimation de la paramétrisation temporelle introduite par Durrleman et al. (2013) provient notamment du caractère non-paramétrique de cette dernière. Afin de lever cette limitation, Hong et al. (2014) introduisent des déformations temporelles paramétriques, en l'occurrence logistiques, donnant lieu à un modèle de *régression géodésique avec recalage temporel paramétrique*. L'algorithmique s'en trouve ainsi considérablement simplifiée et consiste en l'alternance d'une étape de régression pour déterminer la trajectoire d'évolution moyenne et d'une étape de descente de gradient pour estimer les paramètres encodant la déformation temporelle. Malheureusement, ce modèle ne traite que de la régression géodésique et se généralise mal à l'étude de données longitudinales.

Enfin, Su et al. (2014) introduisent également une notion de paramétrisation temporelle mais dans un but tout autre : ici, la paramétrisation sert à aligner géométriquement des trajectoires. Pour cela, les auteurs supposent que les différentes trajectoires observées disposent d'une paramétrisation intrinsèque et non-observée qu'ils vont chercher à apprendre au cours de la procédure d'estimation. Ce modèle permet une comparaison efficace des trajectoires en tant qu'objets géométriques. Su et al. (2014) en proposent une application à la reconnaissance visuelle de la parole. Pour autant, cette paramétrisation n'a pas de sens en terme de modélisation est n'est donc pas interprétable.

Finalement, si on résume la discussion précédente, un modèle performant tant sur l'aspect modélisation que numérique doit être génératif et reposer sur une notion de déformation temporelle paramétrique, faisant sens comme une paramétrisation temporelle de la trajectoire, *i.e.* interprétable dans un cadre applicatif.

De plus, procéder comme Durrleman et al. (2009, 2013), à savoir estimer la trajectoire template par un algorithme « max-max » en minimisant la somme des carrées des distances après recalage entre les données et le template candidat, *i.e.* en minimisant une fonctionnelle d'appariement semblable à celle introduite à l'équation (1.5) n'est pas satisfaisant. En effet, Devilliers et al. (2017,) ont démontré que cet algorithme était inconsistant à cause du bruit. Plus précisément, ils ont obtenus un équivalent du biais de consistance en fonction du niveau de bruit. Ainsi l'inconsistance est-elle inévitable quand le niveau de bruit est suffisamment grand, ce qui est la situation usuelle en pratique.

III – 2. Modèles spatio-temporels à effets mixtes

Schiratti et al. (2015, 2017) ont proposé un modèle hiérarchique pour l'étude de données longitudinales, dit *modèle générique spatio-temporel*. Ce modèle repose sur la notion de variations parallèles d'une courbe sur une variété riemannienne. Aussi commence-t-on par rappeler brièvement cette notion. En effet, ce modèle ne suppose plus que les indivi-

dus suivent une trajectoire d'évolution géodésique mais plutôt une trajectoire parallèle à la courbe représentative, elle supposée géodésique.

Formellement, on souhaite généraliser au cadre riemannien la notion de droites parallèles. Soit M une variété riemannienne géodésiquement complète. Soit $\gamma\colon I \subset \mathbb{R} \to M$ une courbe différentiable sur M, $t_0 \in I$ et $w_0 \in T_{\gamma(t_0)}M$ un vecteur tangent. Alors, on appelle *variation parallèle* de γ dans la direction de w la courbe $\eta^w(\gamma;\cdot)\colon I \to M$ définie par

$$\eta^w(\gamma;\cdot)\colon t \;\mapsto\; \mathcal{E}xp_{\gamma(t)}(P_{\gamma,t_0,t}(w))$$

où $P_{\gamma,t_0,t}(w)$ désigne le transport parallèle du vecteur w, le long de la courbe γ entre les points $\gamma(t_0)$ et $\gamma(t)$. Si dans le cas euclidien les courbes sont formellement identiques, dans le cas riemannien, cela peut ne pas être le cas : autrement dit, du fait de la courbure de la variété, la variation parallèle d'une géodésique n'a *a priori* aucune raison de rester géodésique. Pour s'en convaincre, on peut par exemple visualiser la terre, son équateur et ses différents parallèles dont les deux tropiques.

En particulier, pour calculer la variation parallèle d'une courbe donnée, on doit d'abord transporter le vecteur de direction w parallèlement le long de la courbe γ, et ce pour tout temps t. Numériquement, ce calcul peut être couteux ; des schéma numériques ont récemment été proposés afin d'en réduire la complexité (Louis et al., 2017, 2018).

Le modèle proposé par Schiratti et al. (2015, 2017) est à mettre en parallèle avec le modèle développé par Durrleman et al. (2009) dans sa construction. Soit le jeu de données longitudinales (y_i, t_i) où pour chaque sujet i, on observe $y_i = (y_{i,j})_{j\in\in\mathbb{R}^{k_i}}$ aux temps $t_i = (t_{i,j})_{j\in[\![1,k_i]\!]}$. L'idée de Schiratti et al. (2015, 2017) est d'interpréter ces observations comme des échantillons bruités le long de trajectoires sujet-spécifiques, elles-mêmes dérivant d'une trajectoire représentative de l'évolution globale de l'ensemble de la population par des déformations spatio-temporelles. De plus, comme expliqué en préambule et pour des raisons numériques, ils souhaitent construire un modèle paramétrique.

Pour ce faire, ils imposent une trajectoire représentative géodésique. De fait, la trajectoire représentative est alors naturellement paramétrée par le triplet (t_0, p_0, v_0) où $t_0 \in \mathbb{R}$ désigne un temps de référence, $p_0 \in M$ la valeur de la courbe en ce point et $v_0 \in T_{p_0}M$ la valeur de son vecteur vitesse (Gallot et al., 2004). Finalement, la courbe représentative $\gamma_0\colon \mathbb{R} \to M$ s'écrit

$$\gamma_0\colon\; t \mapsto \mathcal{E}xp_{p_0,t_0}(v_0)(t)$$

où $\mathcal{E}xp_{p_0,t_0}(v_0)$ désigne l'exponentielle riemannienne passant par p_0, à vitesse v_0, au temps t_0. À l'échelle individuelle, les trajectoires sont vues comme les variations parallèles, reparamétrées en temps, de cette courbe γ_0 ; à savoir, pour tout individu i, étant donné un vecteur $w_i \in T_{p_0}M$ et le couple $(\alpha_i, \tau_i) \in \mathbb{R}^2$,

$$\gamma_i\colon\; t \mapsto \eta^{w_i}(\gamma_0; \psi_i(t)) \qquad \text{où} \qquad \psi_i\colon\; t \mapsto \alpha_i(t - t_0 - \tau_i) + t_0$$

et où les effets aléatoires (α_i, τ_i) s'interprètent comme des facteurs d'accélération et des décalages temporels individuels. Autrement dit, Schiratti et al. (2015, 2017) autorisent chacun des sujets à emprunter sa trajectoire spatiale d'évolution avec sa propre vitesse α_i et sa propre avance (ou retard selon le signe) de τ_i par rapport à la trajectoire moyenne. La figure 1.5 illustre cette construction. Finalement, le modèle statistique est donné par

$$y_{i,j} \;=\; \eta^{w_i}\Big(\mathcal{E}xp_{p_0,t_0}(v_0)\;;\;\alpha_i(t_{i,j}-t_0-\tau_i)+t_0\Big)\;+\;\varepsilon_{i,j}$$

avec $\varepsilon_{i,j}$ un bruit blanc gaussien *i.i.d.* Afin de rendre le modèle identifiable, les w_i doivent être choisis perpendiculaires à la trajectoire γ_0. De plus, pour réduire la dimension de l'espace des paramètres à estimer (Giraud, 2014), ils sont supposés dériver d'un modèle d'analyse en composantes indépendantes (Hyvärinen et al., 2004). Autrement dit, Schiratti et al. (2015, 2017) supposent que les w_i sont chacun combinaison linéaire de sources indépendantes et, en fait d'estimer w_i, proposent d'estimer une matrice de design A et des sources s_i telles que $w_i = As_i$. L'estimation des paramètres est réalisée *via* un *maximum a posteriori*, dont Schiratti (2016) démontre l'existence dans ses travaux de thèse, à l'aide de l'algorithme MCMC-SAEM introduit au paragraphe I.3.

Ce modèle a notamment été appliqué à la détection précoce de la maladie d'Alzheimer où il a montré son efficacité. L'étude de Bilgel et al. (2016) sur le dépôt cortical de β-amyloïde, une des caractéristiques pré-cliniques de la maladie d'Alzheimer, est une autre illustration de l'applicabilité de ce modèle.

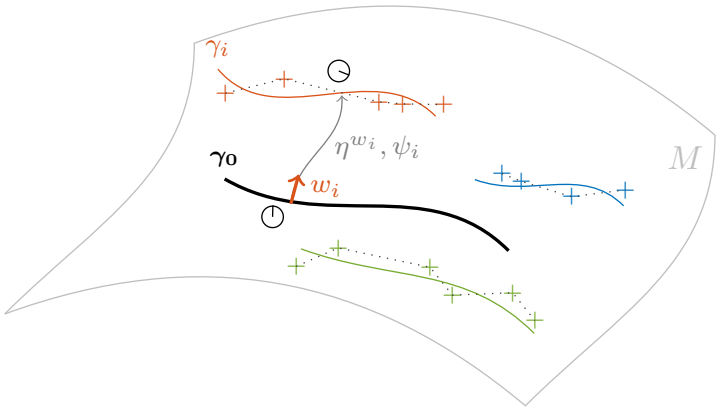


Figure *1.5* – *Modèle générique spatio-temporel de Schiratti et al. (2015, 2017).* On observe des échantillons bruités le long de trajectoires sujet-spécifiques, construites comme des déformations spatio-temporelles d'une trajectoire représentative géodésique. En particulier, chaque trajectoire γ_i évolue avec sa propre temporalité, donnée par la reparamétrisation ψ_i et sa propre géométrie, donnée par le vecteur de translation w_i.

Au paragraphe I.1, nous avons argumenté la nécessité de pouvoir estimer le temps t_0. Vérifions que dans le cas univarié où $M = \mathbb{R}$ le modèle de Schiratti et al. (2015, 2017) diffère effectivement du modèle avec pente et ordonnée à l'origine aléatoires. Supposons

$\mathbb{R}$ muni de sa métrique usuelle. Dans ce cas, le modèle générique s'écrit, pour tout sujet i et toute observation j,

$$y_{i,j} \;=\; p_0 + v_0\,\alpha_i\,(t_{i,j} - t_0 - \tau_i) + \varepsilon_{i,j} \;=\; v_0\,\alpha_i\,(t_{i,j} - t_0) + p_0 - v_0\,\alpha_i\,\tau_i + \varepsilon_{i,j}\,.$$

Ce modèle est donc non linéaire du fait de la présence d'une multiplication entre les effets aléatoires α_i et τ_i. À l'inverse, le modèle avec pente et ordonnée à l'origine aléatoires, en prenant des notations consistantes avec celles de Schiratti et al. (2015, 2017) se ré-écrit

$$y_{i,j} \;=\; (v_0 + \alpha_i)(t_{i,j} - t_0) + (p_0 + \beta_i) + \varepsilon_{i,j}$$

et les deux modèles différent donc bien l'un de l'autre. Plus précisément, on peut interpréter ces deux modèles comme suit : si le modèle avec pente et ordonnée compare la distribution des observations par rapport à un temps de référence t_0, celui de Schiratti et al. (2015, 2017) compare les temps d'observations par rapport à une valeur mesure de référence p_0 et permet donc l'estimation du temps t_0.

Dans la même idée, Kim et al. (2017) ont également proposé un modèle à effets mixtes pour l'étude de données longitudinales : le *modèle riemannien non-linéaire à effets mixtes*. Dans sa construction, il apparait comme une forme de généralisation du modèle hiérarchique géodésique introduit par Muralidharan et Fletcher (2012) et décrit page 15. Plus précisément, étant donné un jeu de données (y_i, t_i) provenant de l'observation des $y_i = (y_{i,j})_{j\in[\![1,k_i]\!]}$ aux temps $t_i = (t_{i,j})_{j\in[\![1,k_i]\!]}$, le modèle s'écrit

$$\begin{cases} y_{i,j} \;=\; \mathcal{E}xp\Big(\,\mathcal{E}xp(b_i\,;\,\Gamma_{b,b_i}(v)\alpha_i(t_{i,j} - \tau_i - t_0))\,;\,\varepsilon_{i,j}\,\Big) \\ b_i \;=\; \mathcal{E}xp\,(b\,;\,u_i) \end{cases}$$

où $\Gamma_{b,b_i}(v) \in T_{b_i}M$ désigne le transport parallèle du vecteur $v \in T_bM$ le long d'une géodésique reliant b et b_i. Autrement dit, si on note γ_i une telle géodésique, $\Gamma_{b,b_i} = P_{\gamma_i,0,1}$ avec les notations introduites précédemment. Le modèle de Kim et al. (2017) partage avec Schiratti et al. (2015, 2017) une reparamétrisation temporelle affine. Cependant et contrairement au modèle de Schiratti et al. (2015, 2017), du fait de la complexité de leur modèle, les paramètres ne peuvent être estimés de manière exacte. Les auteurs introduisent donc une technique d'estimation approchée pour palier ce problème. Enfin, ce modèle est également appliqué à l'étude de l'atrophie corticale.

Le modèle de Schiratti et al. (2015, 2017) a fait l'objet de développements récents pour accroitre son applicabilité. En effet, le modèle tel que nous l'avons décrit induit des paramètres potentiellement en très grande dimension, ce qui rend leur estimation complexe. Koval et al. (2017) proposent une première amélioration de ce modèle pour l'étude des réseaux, *i.e.* de mesures variant au cours du temps sur un graphe fixe, et l'appliquent à l'étude de l'atrophie corticale (Koval et al., 2018).

Bône et al. (2018) proposent une instanciation du modèle de Schiratti et al. (2015, 2017) dans le cas des formes en se plaçant dans le cadre des grandes déformations. Comme

expliqué page 13, ce modèle repose sur un représentation parcimonieuse du groupe des déformations *via* l'introduction de points de contrôle (Durrleman et al., 2011, 2013). Soit $c_0 \in \mathbb{R}^{n_{cp}d}$ un ensemble de n_{cp} points de contrôle, une forme $y_0 \in M \subset \mathbb{R}^d$ et un moment $m_0 \in \mathbb{R}^{n_{cp}d}$. La trajectoire représentative s'écrit alors $\gamma_0 \colon t \mapsto \mathcal{E}xp_{c_0,t_0,t}(m_0) \circ y_0$ et les trajectoires individuelles sont obtenues en prenant la parallélisation reparamétrée de cette courbe représentative, autrement dit :

$$\gamma_i \colon t \mapsto \eta^{w_i}(\psi_i(t)) \circ y_0 \qquad \text{où} \qquad \eta^{w_i} \colon t \mapsto \mathcal{E}xp_{c(t),0,1}\left(P_t(w_i)\right)$$

où, pour tout vecteur moment w, $P_t(w)$ désigne le transport parallèle du vecteur w, le long de la courbe γ_0 entre les instants t_0 et t et où $c(t) \in \mathbb{R}^{n_{cp}d}$ désigne les points de contrôle au temps t, à savoir $c(t) = \mathcal{E}xp_{c_0,t_0,t}(m_0) \circ c_0$. De même que dans le cadre générique, les vecteurs w_i sont supposés être combinaison linéaire de sources indépendantes et les différents paramètres du modèle sont estimés *via* un MCMC-SAEM.

Enfin, une dernière amélioration du modèle générique a été proposée par Bône et al. (2019) en se basant sur des réseaux de neurones, dont l'utilisation permet notamment de traiter des données de très grande dimension. Le modèle générique fait toujours l'objet de recherches actives ; on peut par exemple citer les travaux en cours de Debavelaere et al. (2019) visant à inclure ce modèle générique dans un modèle de mélange afin de pouvoir classer les individus d'une même cohorte en plusieurs sous-populations et, par exemple, de distinguer la progression des patients en bonne santé de celle des patients malades, *etc.*

IV. Plan de la thèse

Le modèle que nous proposons dans cette dissertation se base sur le modèle générique et vise à le généraliser à des situations dans lesquelles la dynamique d'évolution n'est pas unidirectionnelle. En effet, il s'agit de la situation standard dans les applications. C'est le cas par exemple du suivi de chimiothérapie auquel nous consacrons un chapitre : le chapitre 5. En cas de mise en place d'un nouveau traitement, se succèdent généralement trois phases distinctes de progression de la maladie : une phase de réponse au traitement dans laquelle la taille des tumeurs diminue, une phase dite de stabilité et, le plus souvent, une phase d'échappement au traitement lors de laquelle la taille des tumeurs croît de nouveau. Ainsi, modéliser la trajectoire représentative de l'évolution de la population par une courbe géodésique n'est absolument pas réaliste. On propose de s'affranchir de cette contrainte en construisant une trajectoire représentative géodésique par morceau. Ainsi, la progression pourra connaître plusieurs phases d'évolution distinctes : en l'occurrence, une par portion de courbe géodésique dans la courbe représentative.

De plus, malgré une approche novatrice qui a ouvert de nouveaux champs de recherche, le modèle générique développé par Schiratti et al. (2015, 2017) souffre d'un manque de garanties théoriques, ce qui ne permet pas de conclure quant à la fiabilité des

résultats qu'ils obtiennent autrement que par une validation graphique. Aussi, nous nous sommes attelés à démontrer la consistance de notre modèle, démontrant du même coup la consistance du modèle de Schiratti et al. (2015, 2017) et des extensions présentées au paragraphe précédent.

En suivant les méthodes usuelles pour l'inférence statistique dans les modèle non-linéaires à effets mixtes, les paramètres de notre modèle sont estimés *via* l'algorithme MCMC-SAEM. Pour autant, nos expériences numériques se sont heurtées à des limitations techniques dues notamment à une trop grande sensibilité de l'algorithme SAEM à ses conditions initiales. Nous avons donc travaillé à une possible amélioration de cet algorithme en se basant sur des techniques de type recuit simulé ou de « tempering ».

Finalement, les travaux présentés dans ce manuscrit se trouvent à la croisée de ce qui nous semble être le fondement de la modélisation mathématique pour la médecine : l'application avec une collaboration, toujours en cours, avec des oncologues et radiologues de l'Hôpital Européen Georges Pompidou (HEGP), la théorie en tant que garante de la fiabilité des modèles proposés et le développement de techniques numériques performantes afin de permettre l'analyse de jeux de données complexes, hétérogènes et de plus en plus massifs du fait de l'amélioration des techniques d'imagerie médicale.

IV – 1. Organisation des chapitres

Le présent manuscrit est découpé en deux parties : la première consiste en la mise en place d'un cadre d'étude cohérent pour l'analyse de données longitudinales à valeurs sur des variétés riemanniennes ; la deuxième partie sur concentre sur l'aspect algorithmique en proposant une nouvelle classe d'approximations stochastiques de l'algorithme EM. Plus précisément, les différents chapitres du présent manuscrit s'organisent comme suit.

Par la suite, on distingue par le symbole ♢ les chapitres qui correspondent à une contribution de notre part.

Chapitre 3. ♢ Ce premier chapitre est dédié à l'établissement d'un cadre d'étude cohérent pour l'analyse statistique de données longitudinales à valeurs sur des variétés riemanniennes. Dans ce but, nous proposons une généralisation du modèle introduit par Schiratti et al. (2015, 2017) pour l'étude de données longitudinales à dynamiques d'évolution multiples. Ce modèle repose sur la discrimination de déformations temporelles, liées à l'acquisition des données et au rythme de progression du phénomène observé, de déformations spatiales, liées à la géométrie intrinsèque des formes observées.

Pour ce faire, nous proposons un modèle non-linéaire à effets mixtes dans lequel les trajectoires individuelles d'évolution sont vues comme des déformations spatio-temporelles d'une trajectoire moyenne représentative de l'évolution de la population à l'échelle macroscopique. Nous présentons ce modèle sous des hypothèses très génériques afin d'englober une grande classe de modèles plus spécifiques.

L'estimation des paramètres du modèle géométrique est réalisée par un *maximum*

a posteriori dont nous démontrons l'existence et la consistance ; autrement dit, nous démontrons que l'estimateur du MAP s'éloigne d'autant moins de la vraie valeur des paramètres que la taille de l'échantillon est grande. Ce dernier résultat est d'autant plus important que le modèle proposé dans ce chapitre englobe le modèle de Schiratti et al. (2015, 2017), lui même appliqué à la détection précoce de la maladie d'Alzheimer. Ainsi, on fournit des garanties théoriques sur des études pré-cliniques déjà en cours.

Ce chapitre est issu de Chevallier et al. (2017) et Chevallier et al. (2019).

Chapitre 4. Afin de permettre l'étude de modèles pour les formes anatomiques en 3 dimension, on redonne dans le présent chapitre les fondements mathématiques de deux espaces de forme très utilisés dans ce contexte : les courants (Vaillant et Glaunès, 2005) et les varifolds (Charon et Trouvé, 2013).

Chapitre 5. ♢ Ainsi le modèle proposé au chapitre 3 est-il à même de quantifier le rythme de progression d'un processus continu. Nous appliquons ce modèle au suivi de chimiothérapie et plus particulièrement au suivi du cancer métastatique du rein. En effet, dans ce contexte, la compréhension du rythme de progression du cancer est au cœur de la prise en charge médicale.

La première application concerne le suivi de scores RECIST. Ces scores étant des données scalaires, on réalise une instantiation du modèle générique pour des données réelles bornées en se plaçant sur le segment $[0, 1]$ munit de la métrique logistique. Ce modèle a été élaboré en collaboration avec des oncologues et radiologues de l'hôpital européen Georges Pompidou (HEGP). Des expériences numériques sur données synthétiques et réelles en valident la pertinence.

La seconde application porte sur le suivi de formes anatomiques 3D, toujours pour l'évaluation de la réponse tumorale. Ce modèle repose sur la notion de grandes déformations que nous avons discutée en introduction et s'applique aussi bien aux courants (Vaillant et Glaunès, 2005) qu'aux varifolds (Charon et Trouvé, 2013), qui sont des espaces de forme standards pour l'analyse de formes anatomiques et dont on a rappelé les fondements mathématiques au Chapitre 4. On propose également des expériences numériques sur données synthétiques.

Ce chapitre est partiellement issu de Chevallier et al. (2019).

Chapitre 6. Numériquement, l'estimation des paramètres est réalisée *via* une approximation stochastique de l'algorithme EM, à savoir l'algorithme SAEM. Avant d'étudier plus en profondeur cet algorithme, on présente dans ce chapitre la littérature classique concernant l'algorithme EM et ses variantes usuelles. Nous portons une attention particulière au papier séminal de Delyon et al. (1999).

Chapitre 7. ♢ Malgré la performance numérique de l'algorithme SAEM, du fait de la complexité de notre modèle, nos propres expériences se sont heurtées aux limites de ce

dernier. En particulier, l'algorithme SAEM est très sensible à ses conditions initiales et, malgré la stochasticité de la procédure induite par l'approximation stochastique, peut rester piégé dans des minima locaux. De plus, l'algorithme SAEM suppose que l'on est à même de simuler la loi conditionnelle des variables latentes sachant les observations (avec la terminologie des modèles à variables latentes), éventuellement par une méthode de type MCMC, ce qui n'est pas toujours le cas.

Nous proposons dans ce chapitre une nouvelle classe d'algorithmes SAEM : les algorithmes SAEM approchés, ou approximated-SAEM en anglais, dont on démontre la convergence vers des maxima locaux sous des hypothèses standards. Cette classe repose sur la simulation par une loi approchée, en un sens à définir, de la vraie loi conditionnelle dans l'étape de simulation. En particulier, on englobe des algorithmes pré-existants tel que l'ABC-SAEM (Picchini et Samson, 2018) dont l'efficacité numérique avait été établie mais dont la convergence théorique n'avait pas été démontrée.

Enfin, en se basant sur des techniques de recuit simulé, on propose une version tempérée de l'algorithme SAEM afin de favoriser sa convergence vers des minima globaux. Dans cette version, on approche la loi conditionnelle en la tempérant suivant un schéma de températures sinusoïdal amorti. Nous appliquons cette méthode à l'estimation des paramètres dans les modèles de mélange gaussien et en illustrons ainsi la supériorité numérique sur l'algorithme SAEM. Cet algorithme est également appliqué à la séparation de sources *via* l'analyse en facteurs indépendants.

Ce chapitre est issu de Allassonnière et Chevallier (2019).

IV – 2. Liste des publications

Ce manuscrit a donné lieu aux publications suivantes.

a. Acte de conférence

- Chevallier et al. (2017) : **Learning spatio-temporal piecewise-geodesic trajectories from longitudinal manifold-valued data**. Juliette Chevallier, Stéphane Oudard et Stéphanie Allassonnière. Paru dans *Advances in Neural Information Processing Systems* 30, Décembre 2017.

b. Prépublications

- Allassonnière et Chevallier (2019) : **A new class of EM algorithms. Escaping local minima and handling intractable sampling**. Juliette Chevallier et Stéphanie Allassonnière. Juin 2019, Hal-02044722.
- Chevallier et al. (2019) : **A coherent framework for learning spatio-temporal piecewise-geodesic trajectories from longitudinal manifold-valued data**.

Juliette Chevallier, Vianney Debavelaere et Stéphanie Allassonnière. Mai 2019, Hal-01646298.

Reconnaissance.
Ce travail bénéficie d'un financement public Investissement d'avenir, réference ANR-11 LABX-0056-LMH.

– Chapter II –

Introduction in English Language

The aim of this thesis is to propose new methods for the statistical analysis of longitudinal Riemannian manifold-valued data. Since the study of such data is generally conducted using mixed effects models, we first review the classical literature on these models and the different statistical inference techniques associated with them.

Although efficient, these models do not currently apply to manifold-valued data. So we start by redefining the classical study framework for the study of anatomical shapes, i.e. shape spaces. Based on this formalism, we then review the different geodesic regression models for the study of longitudinal manifold-valued data. However, this type of model does not explain inter-individual variability.

For this purpose and based on a notion of temporal reparameterization, Schiratti et al. (2015, 2017) proposed a generic model for the study of longitudinal manifold-valued data. This model is discussed in the last paragraph of this chapter.

Contents

Beyond transversal studies, temporal evolution of phenomena is a field of growing interest. Indeed, for the purpose of understanding a phenomenon, it appears more suitable to compare the evolution of its markers over time than to do so at a given stage. The follow-up of neurodegenerative disorders, such as Alzheimer's disease or Parkinson's disease, is carried out *via* the monitoring of cognitive scores over time. Actually, rather than neuronal degeneration itself which is a natural consequence of cerebral aging, the pathological nature of such diseases lies in the rate of progression of this senescence. The same applies for chemotherapy monitoring: rather than tumors aspect or size, oncologists asses that a given treatment is efficient from the moment it results in a decrease of tumor volume. The study of longitudinal data is not restricted to medical applications and proves successful in various fields of application such as computer vision, automatic detection of facial emotions, social sciences, *etc.*

The study of longitudinal data is usually achieved through mixed effects models. Recent works of Schiratti et al. (2015, 2017) allowed for instance the study of complex data, such as anatomical data. However, these works assume an unidirectional dynamic and fail to encompass situations like multiple sclerosis or chemotherapy monitoring. Indeed, such diseases follow a chronic courses, with phases of worsening, stabilization and improvement, inducing changes in the global dynamic.

The purpose of this PhD work is to develop methodological tools and algorithms suited for the analysis of longitudinal data arising from phenomena that undergo multiple dynamics and to apply these new tools to chemotherapy monitoring of metastatic kidney cancer.

I. Mixed Effects Models for Longitudinal Data

We focus on the statistical analysis of measurements arising from repeated observations of a phenomenon over time.

Mixed effects models (Eisenhart, 1947; Fisher, 1919) help explain a series of observations through two different types of effects: the *fixed effects* shared by all of the individuals in the population and the *random effects* specific to each individual which convey the variability peculiar to each subject. This type of models, by its very multiple-scale nature, is particularly suited to the study of data that come from a hierarchical structure, whose longitudinal data analysis is a particular case.

Two types of mixed effects models can be distinguished: linear mixed models and nonlinear mixed models. Mixed effects models where intensively studied, mainly due to their broad applicability. To give just two examples in the medical field, they became popular as reference tools in both follow-up of neurodegenerative diseases (Milliken and

Edland, 2000; Ospina et al., 2012) and chemotherapy monitoring (Ribba et al., 2014; Rios et al., 2017).

In Verbeke et al. (2014), one can find a comprehensive review of mixed-effects models for longitudinal data ; for a more complete overview, one can refer to Diggle et al. (2002) or Fitzmaurice et al. (2012).

I – 1. Linear Mixed Effects Models

Linear mixed effects models are the simplest ones. They are therefore frequently used for studying longitudinal data. A first historical example, which is still widely used, is the analysis of variance or ANOVA method (Fisher, 2006; Scheffe, 1956). However, variance analysis assume that the dataset is balanced, *i.e.* a data set in which each individual is observed the same number of times, which is rarely the case in practice, especially in the medical context.

Building on the work of Harville (1977), Laird and Ware (1982) proposed a hierarchical model allowing to overcome this constraint. Given multivariate observations $y_i \in \mathbb{R}^{k_i}$, Laird and Ware (1982) assume that there exists for each individual two design matrices $H_i^\alpha \in \mathscr{M}_{k_i,p_\alpha}(\mathbb{R})$ and $H_i^\beta \in \mathscr{M}_{k_i,p_\beta}(\mathbb{R})$ relating respectively to population parameters or fixed effects $\alpha \in \mathbb{R}^p$ and to individual parameters or random effects $\beta_i \in \mathbb{R}^k$, such that the observations are a noisy linear combination of fixed and random effects. In other words, for each subject $i \in [\![1, n]\!]$, we assume

$$y_i \;=\; H_i^\alpha \alpha + H_i^\beta \beta_i + \varepsilon_i \,,$$

where ε_i is drawn from a n_i-dimensional multivariate Gaussian distribution, with mean equal to zero. Moreover, random effects β_i are assumed to be normally distributed and independent of each other. This model thus provides a flexible framework particularly suited for latent data observations, which is a standard issue in the living sciences.

An example of linear model widely used for the analysis of scalar longitudinal data is the *random slope and intercept model* (Cohen et al., 1983). Consider a dataset arising from the observation, for each individual i, of k_i measurements $y_i = (y_{i,j})_{j\in[\![1,k_i]\!]}$ at the times $t_i = (t_{i,j})_{j\in[\![1,k_i]\!]}$. Given a reference time t_0, this model allows to estimate a line reflecting the evolution dynamic of the population at macroscopic scale, as well as individual lines reflecting individual evolutions. More precisely, given the average evolution slope $d\colon t \mapsto a(t-t_0)+b$, where a and b are respectively the mean gradient and the mean intercept, we assume that we observe noisy samples of affine transformations of d, which are, for any individual i, the lines defined by $d_i\colon t \mapsto (a+a_i)(t-t_0)+(b+b_i)$:

$$\forall j \in [\![1, k_i]\!], \qquad y_{i,j} \;=\; (a + a_i)(t_{i,j} - t_0) + (b + b_i) + \varepsilon_{i,j} \,,$$

where $\varepsilon_{i,j} \;\sim\; \mathcal{N}(0, \sigma)$ is a white Gaussian noise and where a_i and b_i are the random parameters of the model and correspond to the individual adjustment parameters of

the slope and the y-intersect. This model thus provides some kind of independence between individuals, which explains its success in applications, for example in personalized medicine (Diaz et al., 2012) or in ecology (Harrison et al., 2018).

However, as Schiratti (2016) raised in his PhD work, the requirement of a reference time t_0 considerably reduces its scope of application. Indeed, as soon as there does not exist a simple and sound way from modeling point of view of fixing this reference time, one will have to estimate it as a fixed effect of the model, at the same level as parameters a and b. Problem : the random slope and intercept model then becomes non-identifiable. There exists an infinite quantity of triplets (a, b, t_0) that maximize the likelihood of the model; that makes it unusable in this case. This situation is actually the most frequent in medical applications: there is no general correlation between the age of the patients and the current stage of development of their disease. Worse still, estimating this time is a key issue of an appropriate medical care as it corresponds to a change in the condition of the patient: escape from the treatment in chemotherapy monitoring, entry into a new phase of progression of a disease, *etc.* Schiratti et al. (2015, 2017) proposed a nonlinear mixed model, called *generic spatio-temporal model* in order to cope with this issue. This model is detailed in the section III.2 of this chapter.

I – 2. Nonlinear Mixed Effects Models

The linear dependency between effects and observations can of course be dismissed, leading to *nonlinear mixed effects models*. These models were historically introduced by Sheiner and Beal (1980) and Bates and Watts (1988) in the context of pharmacokinetics studies. While still broadly used in this context (Comets et al., 2010; Lavielle, 2014), these models also proved to be applicable in other domains such as medical follow-up, as explained in the preamble. They are subjects of intensive research since the 90's; for the sake of concision, we only present here the version proposed by Lindstrom and Bates (1988).

As in the linear case, we assume that we observe, for each individual i, a couple (t_i, y_i) corresponding to measurements $y_i = (y_{i,j})_{j\in[\![1,k_i]\!]}$ and their associated times $t_i = (t_{i,j})_{j\in[\![1,k_i]\!]}$. Let f a nonlinear function, and for each subject i, let $H_i^\alpha \in \mathscr{M}_{p_z,p_\alpha}(\mathbb{R})$ and $H_i^\beta \in \mathscr{M}_{p_z,p_\beta}(\mathbb{R})$ two design matrices relating respectively to fixed effects $\alpha \in \mathbb{R}^{p_\alpha}$ and random effects $\beta_i \in \mathbb{R}^{p_\beta}$. We then write for each individual i:

$$\forall j \in [\![1, k_i]\!], \quad y_{i,j} \;=\; f(z_i; t_{i,j}) + \varepsilon_{i,j}\,, \qquad \text{where} \qquad z_i \;=\; H_i^\alpha \alpha + H_i^\beta \beta_i \in \mathbb{R}^{p_z}\,,$$

where $\varepsilon_{i,j}$ is a white Gaussian noise, and where the random variables β_i are assumed to be normally distributed and independent of each other. Obviously, the linear models described above are particular cases of nonlinear models.

I – 3. Statistical Inference for Nonlinear Mixed Effects Models

Besides the flexibility of the models introduced by Laird and Ware (1982), the success of mixed effects models is due to their great numerical tractability and to the joint developpement of efficient algorithms and powerful computers. In particular, the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is a very popular algorithm which allows for maximizing the likelihood of latent data models in very general frameworks.

The EM Algorithm and its Variants. Linear mixed effects models are an ideal application framework for the EM algorithm. Already, in the seminal paper of Dempster et al. (1977) a specific attention were given to it and the trend continued thereafter (Foulley, 2002; Laird et al., 1987; Laird and Ware, 1982; Meng and Van Dyk, 1997).

Convergence of the EM algorithm toward a *local minimum* of the observed likelihood was proved by Dempster et al. (1977) with revisions by Wu (1983). However, the assumptions required at first instance were difficult to check and the framework introduced by Delyon et al. (1999) provides more reasonable ones. The algorithm iterates two steps until convergence: an expectation step, the E-step, in which we compute the the conditional expected log-likelihood, taking into account the last observed variables, and a maximization step, the M-step, in which we estimate a *maximum* of the likelihood of the parameters by maximizing the likelihood found through the E-step.

Besides upgrades to the speed of convergence (McLachlan and Krishnan, 2007), many other variants were proposed for the EM algorithm. We can distinguish two types of upgrades: the ones concerning the expectation step and the ones concerning the maximization step. For the later, the *Generalized EM* (Delyon et al., 1999) no longer requires a maximization of the expectation at each step but only an increase in it. This way, one can apply the EM algorithm even without any analytic solution to the M-step. In Lange (1995) version, the maximization step is performed by use of a Newton-Raphson method for instance.

Alternatives to the computation of the expectation involve the introduction of stochasticity in the estimation procedure. With the *stochastic EM* algorithm (SEM), Celeux and Diebolt (1985) proposed to replace the computation of the expectation by a numerical estimation of it *via* a simulation of the latent data. Wei and Tanner (1990) generalized this idea by replacing the computation of the expectation by a Monte-Carlo approximation of it, leading to the *Monte-Carlo EM* or MCEM. By adjusting the number of random samplings in the Monte-Carlo summation, we are able to mimic the behavior of a simulated annealing algorithm (Celeux et al., 1995). An alternative approach developed by Delyon et al. (1999) consists in replacing the computation of the expectation by an approximation of Robins-Monro type (Robbins and Monro, 1951), which is known to converge toward the expectation under *ad hoc* hypotheses. This procedure is referred to as *stochastic approximation EM* algorithm or SAEM algorithm. Finally, unlike their deterministic counterparts, these stochastic variants of the EM algorithm are able to get

away from local *maxima*. Therefore the convergence toward global *maxima* is favored.

Markov Chain Monte Carlo Methods. When exact simulation of latent variables is not tractable, rely on approximate sampling through Markov chain Monte Carlo method, or MCMC method, (Andrieu et al., 2003; Brooks et al., 2011; Robert and Casella, 1999) proved to be successful. The idea of MCMC methods is to generate a Markov chain converging toward the law we want to draw variables from. More specifically, we replace the simulation of one sample from this complicated law by the generation of a potentially high number of samples from (hopefully) simpler distributions. Among these samplers, the most frequently used is probably the *Metropolis-Hastings algorithm*. First introduced in the particular case of the Boltzmann distribution (Metropolis and Ulam, 1949; Metropolis et al., 1953), the Metropolis-Hastings algorithm was generalized to any distribution by Hastings (1970). A strength of this algorithm is that it only requires the knowledge of the target distribution up to a multiplicative constant. Then, it allows to avoid computing the normalization constant, which is often an intractable computation. The Metropolis-Hastings algorithm can be seen as a generalization of rejection sampling: at each iterative step, and given the current state of the Markov chain, we make a proposal for an increment and we accept it as soon as it improves the "likelihood". More precisely, given a target law π, a (pseudo)random generator $q(\cdot; x_k)$ and x_k the current state of the chain at iteration k, we accept the proposal $x^* \sim q(\cdot; x_k)$ with probability

$$\alpha(x_k, x^*) \;=\; \frac{\pi(x^*)\, q(x_k, x^*)}{\pi(x_k)\, q(x^*, x_k)}\,.$$

Moreover, this sampler can be incorporated into a Gibbs sampler and is thus particularly suitable for high dimensional data.

Building on the work of Kuhn and Lavielle (2004) which proves the convergence of the MCMC-SAEM algorithm in the case where the variables generated along the procedure remain bounded, Allassonnière et al. (2010) proved the convergence of the MCMC-SAEM algorithm in greater generality. Note that the convergence of this algorithm only requires a single step of MCMC, which makes it very competitive computationally speaking. This algorithm is used in the Monolix software and proves to be widely applicable, especially for pharmacokinetics models (Chan et al., 2011; Lavielle and Mentré, 2007).

Despite the flexibility of the models we described (linear models in I.1 and nonlinear models in I.2), and as a result of how their are written, they can be applied only to scalar data. But the applications, especially in the medical sciences we are interested in, involve highly structured data such as scanners, images, tensors or 3D anatomical shapes. It is thus deemed necessary to propose a statistical framework suited from these data, both massive and heterogeneous.

II. The Use of Riemannian Geometry for the Study of Longitudinal Data

In this section, and unless otherwise stated, we will refer generically to any structured data as a *shape*. Thus, a shape may name an image, as well as a mesh, a tensor, a submanifold, *etc.*

Riemannian geometry is a particularly suitable tool for the mathematical modeling of shapes. Indeed, rather than analyzing shapes individually, it seems more efficient to consider sets or populations of shapes and to try to understand them as spaces in the mathematical sense (Trouvé and Younes, 2015). By construction, these spaces will naturally inherit a Riemannian manifold structure.

II – 1. Shape spaces

The first historical example of *shape spaces* (although not considered as such at that time) has its origins in the work of D'Arcy Thompson (1942). His idea was to compare not the anatomical shapes in a intrinsic way but rather to quantify the necessary deformations allowing to go from an anatomical shape to another, as illustrated in figure 2.1. This idea will be taken up and formalized by Grenander (1993), giving birth to the field of computational anatomy.

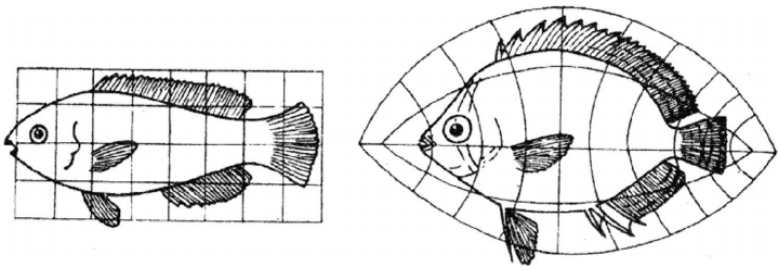


***Figure 2.1** – D'Arcy Thompson's shapes space.*
Geometric deformations of two fishes. Illustration taken from the book *On Growth and Form* of D'Arcy Thompson (1942).

This idea is as follows: given a set M of shapes living in $\mathbb{R}^n$, we are going to construct the corresponding shape space by having a *group of deformations* $\mathcal{G}$ acts (transitively) on the set M and by considering the unique[1] orbit of this action $\mathcal{G} \cdot x_0$, where $x_0 \in M$. Then all the shapes in a shape space can be obtained by deformation of x_0 *via* an element of $\mathcal{G}$. We call x_0 a template. We will often choose for the group of deformations $\mathcal{G}$ the group $\mathscr{C}^1(\mathbb{R}^n)$ of the $\mathscr{C}^1$-diffeomorphisms of $\mathbb{R}^n$, but any sub-group of the group of bijections

[1] uniqueness of the orbit comes from the action of group being transitive. Otherwise, we can restrict ourselves to a particular orbit.

of $\mathbb{R}^n$ could be used. The simplest example of a shape space is the space of *landmarks*

$$M \;=\; \{x = (x_1, \dots, x_p) \in (\mathbb{R}^n)^p \mid \forall i, j \in [\![1, p]\!],\; i \neq j,\; x_i \neq x_j\}$$

endowed with the group $\mathcal{G} = \mathscr{C}^1(\mathbb{R}^n)$, where the action of $\mathcal{G}$ on M is given by

$$\forall\, g \in \mathcal{G}, \quad \forall\, x \in M, \qquad g \cdot x = \big(g(x_1), \dots, g(x_p)\big)\,.$$

This space was introduced and studied mainly by Kendall (1984). Another classical example of a shape space is the space of images: $M = \mathbb{L}^2(\Omega, \mathbb{R})$ with $\Omega \subset \mathbb{R}^n$.

Deformation Metric Mapping. Making such a group act on the Riemannian manifold of shapes, we can endow this manifold with a distance that quantifies the cost of deformation from a shape to an other (Younes, 2010). More precisely, if we assume that $\mathcal{G}$ is endowed with a right-invariant metric $d_{\mathcal{G}}$, then we can endow M with the pseudo-metric defined by

$$\forall\, x, y \in M, \quad d_M(x, y) \;=\; \inf_{g \in \mathcal{G}} \{d_{\mathcal{G}}(Id, g) \mid g \cdot x = y\}\,.$$

Thus, in order to endow M with a distance suited to quantify the cost of deformation, it is "enough" to endow $\mathcal{G}$ with a right-invariant metric. For that purpose, Dupuis et al. (1998) and Beg et al. (2005) introduced the concept of *large deformation diffeomorphic metric mapping* or LDDMM. The main idea of LDDMM is to allow for, accordingly to their name, large deformations of space while keeping control over them, particularly through smoothness. To do so, we restrict the group of deformations $\mathcal{G}$ to a sub-group $\mathcal{G}_V \subset \mathcal{G}$ of well-behaving deformations, in a sense we are going to specify. We briefly cover the mathematical foundations of LDDMM. One can find a comprehensive and detailed construction in Younes (2010) and Glaunès (2005) for instance.

Let V a set of vector fields over $\mathbb{R}^n$, whose norms correspond to the cost of deformation. We assume that V can be endowed with a Hilbert space structure and is continuously embedded into the space $\mathscr{C}^1_0(\mathbb{R}^n)$ of the diffeomorphisms v that decay at infinity and whose differentials $\mathrm{d}v$ decay at infinity. Intuitively, we would want to quantify the cost of matching two forms x and y *via* the displacement field $v - Id$ such that $v(x) = y$ while having in mind that the morphism Id corresponds to an ideal situation where no deformation is needed. However, as soon as the deformation becomes too important, *i.e.* as soon as ϕ becomes large in norm with respect to Id, such a morphism cannot be invertible. But, because of obvious requirements of symmetry, we want these deformations to be invertible. In order to control their amplitude, Beg et al. (2005) had the idea of proceeding infinitesimally. Let $\mathbb{L}^2_V = \mathbb{L}^2([0,1], V)$ the set of all time-dependent vector fields $v = (v_t)_{t \in [0,1]}$ which are $\mathbb{L}^2$-integrable with respect to t, *i.e.* such that

$$\int_0^1 \|v_t\|^2_{1,\infty}\, \mathrm{d}t \;=\; \int_0^1 \left(\|v_t\|_\infty + \|\mathrm{d}v_t\|_\infty\right)^2 \mathrm{d}t \;<\; +\infty\,.$$

We then denote (and one can check that it rightfully defines a group (Younes, 2010))

$$\mathcal{G}_V \;=\; \{\, \phi_1^v \,|\, v \in \mathbb{L}^2_V \,\}\,,$$

where ϕ_1^v is the flow at time $t = 1$ associated with the differential equation

$$\partial_t \phi_t^v \;=\; v_t \circ \phi_t^v \qquad ; \qquad \phi_0^v \;=\; Id\,. \tag{2.1}$$

This equation is the infinitesimal counterpart of the displacement field $v - Id$. While minimizing the norm of these vector fields v_t, we minimize at each moment the deformation and thus the global deformation given by ϕ_v^1. We want to emphasize here that the application $t \mapsto \phi_t^v$ does not *a priori* have its values in $\mathbb{R}^d$ but rather in the Riemannian manifold of shapes M. For this reason, we will often use a better-defined integral formulation, where $\int$ is the Bochner integral (Bochner, 1933) for Banach spaces that generalize Lebesgue's integral :

$$\forall t \in \mathbb{R}, \quad \forall x \in M, \quad \phi_t^v(x) \;=\; x + \int_0^t v_t \circ \phi_s^v(x))\,\mathrm{d}s\,.$$

Note that $\phi_1^v \in \mathscr{C}^1(\mathbb{R}^n)$ and that the differential equation above admits an unique solution for each initial condition (Younes, 2010). Lastly, we denote for all $\phi, \phi' \in \mathcal{G}_V$

$$d(Id, \phi) \;=\; \inf_{v \in \mathbb{L}^2_V} \left\{ \left(\int_0^1 \|v_t\|_V^2 \,\mathrm{d}t \right)^{1/2} \,\middle|\, \phi_1^v = \phi \right\} \quad \text{and} \quad d_{\mathcal{G}}(\phi, \phi') \;=\; d(Id, \phi' \circ \phi^{-1})\,.$$

It is possible to prove that this *minimum* is reached for a certain v in $\mathbb{L}^2_V$, *i.e.* that there exists a vector field v able to transport a configuration of the shape space into another while minimizing the cost of deformation (Younes, 2010). Moreover, the metric $d_{\mathcal{G}}$ is right-invariant and, finally, the cost of deformation from a shape x to an other shape y is given by

$$\forall\, x, y \in M, \quad d_M(x, y) \;=\; \inf_{v \in \mathbb{L}^2_V} \left\{ \left(\int_0^1 \|v_t\|_V^2 \,\mathrm{d}t \right)^{1/2} \,\middle|\, \phi_1^v \cdot x = y \right\}\,.$$

Matching via Large Deformation Diffeomorphisms. As the previous discussion indicates, if we are able to match two shapes, we are then able to quantify the difference between them. We will in fact favor inexact matching instead of exact matching, the former being more realistic in the case of computational anatomy. In other words, the matching of two shapes is performed through the minimization of the cost of deformation described above penalized by a data attachment term A. The matching problem with A consists in minimizing on $\mathbb{L}^2_V$ the functional defined by

$$J\colon v \;\mapsto\; \frac{1}{2} \int_0^1 \|v_t\|_V^2 \,\mathrm{d}t \;+\; \lambda\, A(\phi_1^v)\,, \tag{2.2}$$

where $\lambda > 0$. The existence of such a minimizer was proved by Glaunès (2005) in his doctoral work.

Given the large deformation framework, and given both an appropriate representation of shapes and a suitable data attachment term, we are able to quantify the difference

between any two shapes. In the case of landmarks, where points are labeled, the data attachment term is nothing but a $\mathbb{L}^2$ type metric:

$$\forall\, x = (x_1, \dots, x_p),\ y = (y_1, \dots, y_p) \in M, \quad A(x, y) = \sum_{i=1}^{p} \|x_i - y_i\|_2^2 ,$$

where $\|\cdot\|_2$ is for the 2-norm of $\mathbb{R}^n$. Numerous works were conducted on that matter which we will not detail here but which we partially review in chapter 5. Albeit non-exhaustively, we cite the work of Glaunès et al. (2004) for non-labeled points matching through discrete measures matching, the work of Glaunès et al. (2008) for curves matching, of Vaillant and Glaunès (2005) and Durrleman (2010) for orientated surfaces matching *via* currents matching, of Charon and Trouvé (2013) for non-orientated shapes matching *via* the use of of varifolds, of Roussillon and Glaunès (2016) for non-orientated shapes matching again, but *via* the use of normal cycles, and lastly the work of Charlier et al. (2017) for the analysis of functional shapes, *i.e.* of functions and signals defined on variable geometric supports.

Sparse Parameterization of Diffeomorphisms. Finally, the last aspect that need to be elucidated is the choice of an appropriate norm for the Hilbert space V. For that purpose, we are going to endow V with a reproducing kernel Hilbert space structure, or RKHS structure. RKHS, which were introduced by Aronszajn (1950), are broadly used, notably in machine learning (Friedman et al., 2001), and proved to be widely applicable these last decades.

Let k_V be a reproducting kernel on V. Then, we endow V with the norm defined by

$$\forall v \in V, \qquad \|v\|_V^2 \;=\; \int_{\mathbb{R}^n} \Big\langle Lv(x) \,\Big|\, v(x) \Big\rangle \,\mathrm{d}x \,,$$

where $L\colon V \to V^*$ is a differential operator such that $K = L^{-1}$ with

$$\forall v \in V, \quad \forall x \in M, \qquad (Kv)(x) = \int_{\mathbb{R}^n} k_V(x, y)\, v(y)\, \mathrm{d}y \,.$$

We are trying to match two shapes x and y. Building on the work of Joshi and Miller (2000) for landmarks matching, Durrleman et al. (2011) proposed a sparse representation of the group of deformations through the introduction of control points. Assume that we can associate to each shape a configuration of N_c distinct points, that we denote abusively x and y. Assume $c(t) = \{c_1(t), \dots, c_{N_c}(t)\}$ is composed of N_c distinct points on M, called *control points* and changing over time. Let A be a data attachment term that measures the attach to the shape y and let $(v_t)_t$ be the family of vector fields directing the dynamics of $t \mapsto c(t)$, *i.e.* such that

$$\forall i \in [\![1, N_c]\!], \qquad \frac{\mathrm{d}c_i}{\mathrm{d}t} = v_t(c_i) \quad \text{and} \quad c_i(0) = x_i \,. \tag{2.3}$$

Then (Joshi and Miller, 2000), there exists a family of N_c vectors $\alpha(t)$ depending on time such that the family of vector fields that realize the matching between x and y,

that is to say which minimize the functional J defined in (2.2), is given by

$$\forall t \in \mathbb{R}, \quad \forall x \in M, \qquad v_t(x) \;=\; \sum_{j=1}^{N_c} k_V\big(x, c_j(t)\big)\, \alpha_j(t)\,.$$

The vectors $\alpha(t)$ are called *momenta* associated to the control points $c(t)$. Using classical Hamiltonian mechanics results, and since for each i we have $c_i(0) = x_i$, J also writes (Allassonnière et al., 2005; Miller et al., 2006)

$$J(\alpha(0)) \;=\; \sum_{i=1}^{N_c}\sum_{j=1}^{N_c} \Big\langle k_V(x_i, x_j)\, \alpha_j(0) \,\Big|\, \alpha_i(0) \Big\rangle + \lambda\, A(c(1))) \tag{2.4}$$

and we replaced the estimation of a family of vector fields by the estimation of a finite number of initial momenta. In particular, with $c(0)$ fixed, we interpret $c(1)$ as a function of the initial momentum $\alpha(0)$.

The minimization of the energy functional J is performed through a *geodesic shooting* algorithm (Allassonnière et al., 2005; Miller et al., 2006): until convergence, it iterates steps of integration of the equation de l'équation (2.3) with fixed initial momenta $\alpha(0)$ and a step of minimization by gradient descent from initial momenta $\alpha(0)$ to fixed final control points $c(1)$. A schematic explanation of this algorithm is showed in figure 2.2.

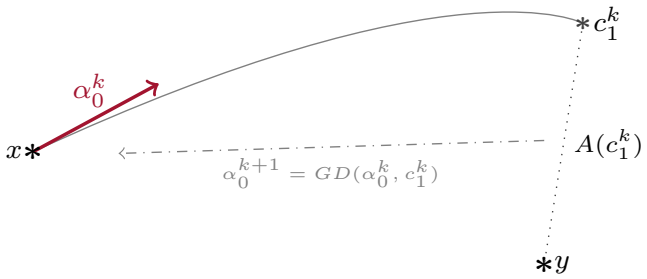


***Figure 2.2** – Geodesic shooting algorithm.*
In this figure, we denote α_0^k the initial momenta $\alpha(0)$ at iteration k, and c_1^k the final control points configuration at this step. The geodesic shooting algorithm alternates between a gradient descent step and an integration step until convergence occurs.

This provides a flexible and rigorous framework for the study of shapes through Riemannian geometry. As explained above, mixed effects models we described in Section I do not apply as such to Riemannian manifold (non euclidean ones, for instance). Nevertheless, generalizations of these models were proposed in the scientific literature, which we deal with in the next section.

II – 2. Geodesic Regression on Riemannian Manifolds

As we said in Section I.1, the random slope and intercept model is commonly used in the euclidean case. In this model, it is assumed that the population can be described at a macroscopic scale by a straight line and that individual evolutions are linear deformations of this template line. If we endow $\mathbb{R}^n$ with the usual scalar product, we are actually assuming the mean evolution of the population is undergoing a geodesic trajectory.

Shapes-Based Approach. In its simplest form, *i.e.* if we remove all random effects, we get back the classical linear regression model, also broadly used. Fletcher (2011) propose to generalize *in extenso* this model, assuming the trajectory is geodesic for some metric, which leads to a *geodesic regression model*. More precisely, given a point $p \in M$ on the Riemannian manifold and a vector $v \in TM$ of the tangent bundle of M, we write for each observation $y_i \in M$ associated to the scalar $x_i \in \mathbb{R}$:

$$y_i \;=\; \mathcal{E}xp\left(\mathcal{E}xp\left(p\,;x_i v\right);\,\varepsilon\right),$$

where ε is a random variable with values in the tangent space to M at point $\mathcal{E}xp\,(p, x_i v)$, and where $\mathcal{E}xp\,(p, v) = \mathcal{E}xp_p(v)$ is the (Riemannian) exponential map at point p, of tangent vector v, *i.e.* the value at time 1 of the unique geodesic passing through p at time 0 with initial speed v (Gallot et al., 2004; Jost, 2002). However, there is no explicit expression for the likelihood of this model. Estimation of the parameters is thus performed through minimization of least squares criterion. Kim et al. (2014) propose a generalization of this model for multivariate regression.

Building on the geodesic regression model and on Laird and Ware (1982) model, Muralidharan and Fletcher (2012) introduced a *generative and hierarchical model* for the analysis of longitudinal data with values on Riemannian manifold. As illustrated in figure 2.3, each individual is assumed to follow a geodesic trajectory, itself a random perturbation of a mean geodesic trajectory. More precisely, assuming we observe the couple (x_i, y_i) for each subject i, where $x_i = (x_{i,j})_{j\in[\![1,k_i]\!]} \in \mathbb{R}^{k_i}$ and $y_i = (y_{i,j})_{j\in[\![1,k_i]\!]} \in M^{k_i}$,

$$\forall j \in [\![1, k_i]\!], \quad y_{i,j} \;=\; \mathcal{E}xp\left(\mathcal{E}xp\left(p_i\,;x_{i,j} v_i\right);\,\varepsilon_{i,j}\right),$$

where we kept the previous notations. The random effects (p_i, v_i) are then seen as a random perturbation of a geodesic curve on the tangent bundle of M. To do so, Muralidharan and Fletcher (2012) propose to endow the tangent bundle of M with a Riemannian metric: the Sasaki's metric. Let $(\alpha, \beta) \in TM$ be the fixed effects and let a family of tangent vectors in TM at that point (q_i, w_i). Then, the proposed model writes

$$(p_i, v_i) \;=\; \mathcal{E}xp_{\mathcal{S}}\left((\alpha, \beta)\,;\,(q_i, w_i)\right),$$

where $\mathcal{E}xp_{\mathcal{S}}$ is the (Riemannian) exponential map associated with Sasaki's metric on the tangent bundle TM of M. Estimation of the parameters is, like in geodesic regression, performed through a least squares algorithm.

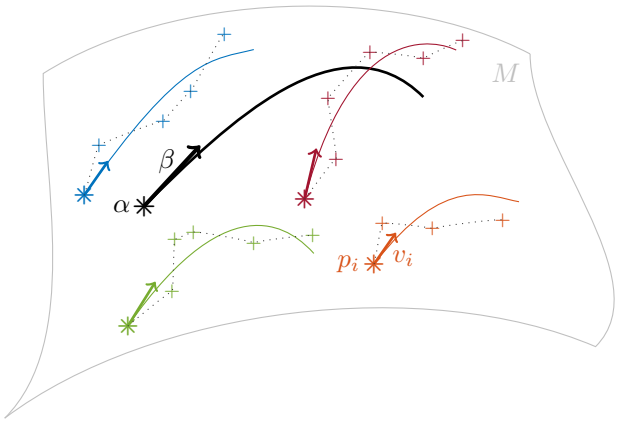


***Figure 2.3** – Geodesic hierarchical model of Muralidharan and Fletcher (2012).* We observe noisy samples along geodesic trajectories, seen as random perturbations of a representative geodesic trajectory.

Deformations-Based Approach. We have explained in Section II.1 the close link between groups of deformations and shape spaces. Building on this link, Singh et al. (2013, 2014) introduced a *geodesic hierarchical model for diffeomorphisms*. This model consists in estimating a geodesic trajectory of diffeomorphisms at the population scale. We can indeed endow the group of $\mathscr{C}^1$-diffeomorphisms of $\mathbb{R}^n$ with a suitable metric, making it a Riemannian manifold (Arnold, 1966; Miller et al., 2006). At the individual scale, trajectories consist in pieces of geodesics on the manifold of shapes M. These curves being assumed to be geodesic, they are completely parametrized by their value $p_i(0)$ and speed $v_i(0)$ at the first time of observation (Gallot et al., 2004). Then, the random effects of the model are enclosed in the couple $(p_i(0), v_i(0))$ and we estimate them with a least squares method, while requiring that $p_i(t) = \phi_i(t) \cdot p_i(0)$ is close enough to the corresponding observations, and where $\phi_i(t)$ is a diffeomorphism arising from the mean trajectory. In particular, random effects in this model depends heavily on the first time of observation, and thus a change in this first time involves a corresponding change in the estimated value for these effects. The first time of acquisition does not have insofar any meaningful place in the modeling process and there is thus a need for a model robust to changes in the time origin.

This model proved to be efficient on images (Singh et al., 2016). More generally, improvement of imaging techniques aroused lively interest the last decade, and the search for models well suited for this type of data was at the heart of many works. We can cite for example Datar et al. (2009), Durrleman et al. (2013) and Fitzmaurice et al. (2012).

Beyond images,models adapted to 2D et 3D shapes encoded with landmarks (*cf.* Section II.1) were also developed. For example, Trouvé and Vialard (2012) proposed a non-parametric spline regression model relying on a random perturbation in the Hamiltonian equations that determine the geodesic flow. Building on the work of Cates et al. (2007), Datar et al. (2012) developed a new way to create models for the study of shapes,

relying on the alternation of two steps of minimization: minimization of a cost functional corresponding to the entropy of the system in order to maximize the correspondence between the shapes and estimation of the parameters of a linear mixed effects model. These works led to the development of the open source software Seg3D, which speak to the relevance of this approach.

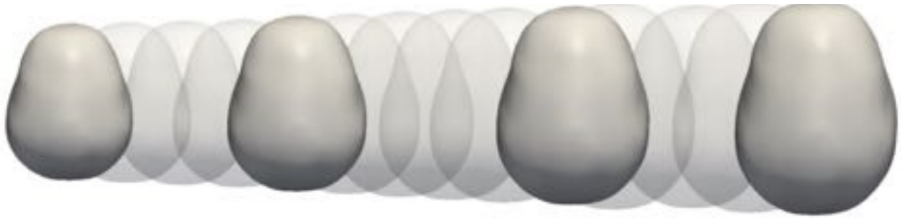


Figure 2.4 *– An example of shapes regression.*
Illustration taken from Fishbaugh et al. (2017). The trajectories (transparent surfaces) were obtained through regression of the four observed shapes (bold surfaces).

While efficient, the models we described are not versatile enough: they can only handle one type of data; images or landmarks for instance. In order to overcome this constraint, Fishbaugh et al. (2017) proposed a *geodesic regression of shapes model* in the general framework of LDDMM, doing so with the help of the sparse representation of the group of deformations *via* the control points introduced by Durrleman et al. (2011, 2013). Let k_V be a reproducing kernel on V. We use the notations from the paragraph dealing with the principle of reduction, page 41.

The purpose of this model is to estimate a continuous geodesic trajectory of shapes from a finite sequence of shapes $(O_{t_i})_i$ observed at times $t_i \in [t_0, T]$. We interpret this trajectory as the deformation of a template shape X_0 by a diffeomorphism ϕ_t smooth enough and satisfying the geodesic flow equation (2.1): at each time, $X_t = \phi_t(X_0)$. This trajectory being assumed to be geodesic, the morphism ϕ_t minimize in particular the energy functional J defined in (2.2). The reduction principle we described above ensures that we can parameterize the flow ϕ_t with a finite number of control points and momenta, namely the initial control points and momenta $(c(0), \alpha(0))$ associated with the geodesic shooting equation (2.3). Moreover, ϕ_t satisfies the geodesic flow equation (2.1), where we can express v_t in terms of k_V (2.4). In brief, there exist two functions F and G depending on the kernel k_V such that

$$\big(\dot{c}(t), \dot{\alpha}(t)\big) \;=\; F\big(c(t), \alpha(t)\big) \qquad \text{and} \qquad \dot{X}(t) \;=\; G\big(X(t), c(t), \alpha(t)\big)\,.$$

If we also put a smoothness constraint on ϕ_t that only depends on $(c(0), \alpha(0))$, we can rewrite the functional J as a function depending only on $(X(0), c(0), \alpha(0))$. We are then reduced to the geodesic shooting situation, *i.e.* the minimization of a function by alterning steps of integration (possibly in the past) and gradient descent. The detailed algorithm is presented in Fishbaugh et al. (2017) and applied therein to regression of 3D anatomical shapes.

In the construction described above, we do not make any assumption on the nature of the shape space and the model proposed by Fishbaugh et al. (2017) allows to treat shapes of various nature (currents, varifolds, normal cycles, *etc.*), as expected. Moreover, the use of control points reduces the number of parameters one needs to estimate while allowing important deformations. It is of a great importance from a statistical point of view: practically speaking, we have access only to a small number of observations and we have to reduce the dimension of the parameters space as far as possible in order to circumvent the curse of dimensionality (Giraud, 2014).

The model we described is a part of the open source software Deformetrica, itself described in detail in Bône et al. (2018). Moreover, we specify in figure 2.4, taken from Fishbaugh et al. (2017) what we mean by shapes regression, namely the estimation of a *continuous* trajectory of shapes that explains at best the observed shapes.

Aside from the ones of Muralidharan and Fletcher (2012) and Singh et al. (2013, 2014), the models we described are unfortunately not able to explain de variability between individuals. But still, in the aim of making predictions from datasets, this variability is of tremendous importance. In other words, the geodesic regression model is not enough to fully explain a longitudinal dataset.

We also want to give some meaning to the estimated parameters. As explained in Section I.1, the rate of progression provides information in practice, especially in medical applications. We want to allow each subject to follow its own evolution curve with its own pace, *i.e.* with specific speed and reference times. The models of Muralidharan and Fletcher (2012) and Singh et al. (2013, 2014) require a strong dependence of the individual trajectory on the first time of acquisition and thus do not allow for such reparametrizations.

III. Spatio-Temporal Models for the Study of Longitudinal Data

In practice, the different individuals are seldom observed the same number of times and even less at the same age. Therefore, we need models able to overcome this issue.

III – 1. Spatio-Temporal Transformations

In order to cope with this matter of facts, Durrleman et al. (2009) choose to differentiate spatial deformations, linked with the intrinsic geometry of observed shapes, from temporal deformations, linked with acquisition constraints. To do so, they introduce the notion of *spatio-temporal atlas* described bellow.

Let a dataset consisting in the observation of N subjects $(S^i)_{i\in[\![1,N]\!]}$ at the corresponding times $(t_j^i)_{i\in[\![1,N]\!],j\in[\![1,k_i]\!]}$. The idea of Durrleman et al. (2009) is to estimate a continuous trajectory $M_t = \chi_t(M_0)$ of shapes such that each observation corresponds to the evaluation at a given time of a spatio-temporal deformation of this template scenario. The mean trajectory is obtained by regression from a template shape M_0 on which smoothness conditions are imposed; we denote $\mathrm{Reg}(\chi)$ these conditions. Moreover, for each individual, we define a couple of deformations (ψ^i, ϕ^i), one corresponding to a spatial deformation and one corresponding to a temporal deformation, in order to match $\phi^i\left(M(\psi^i(t_j^i))\right)$ and $S^i(t_j^i)$, namely the spatio-temporal deformation of the mean trajectory M at time t_j^i and the observation $S^i(t_j^i)$. Then, we denote:

$$J(\psi^i, \phi^i, \chi, M_0) \;=$$
$$\sum_{i=1}^{N}\left\{\sum_{t_j^i} d\left(\phi^i\left(\chi_{\psi^i(t_j^i)}(M_0)\right)\,,\,S^i(t_j^i)\right)^2 + \gamma^{\phi}\mathrm{Reg}(\phi^i) + \gamma^{\psi}\mathrm{Reg}(\psi^i) + \gamma^{\chi}\mathrm{Reg}(\chi)\right\}, \quad (2.5)$$

where $\mathrm{Reg}(\psi^i)$ and $\mathrm{Reg}(\phi^i)$ are smoothness constraints on ψ^i and ϕ^i. In order to minimize this functional, Durrleman et al. (2009) proceed as in the LDDMM framework and thus interpret the morphisms ϕ^i and ψ^i as the flows of *ad hoc* differential equations, whose underlying vector fields belong to a RKHS. Doing so, they only need to estimate a family of momenta associated to reproducing kernels rather than a family of morphisms.

To our knowledge, this model is the first to break spatial and temporal deformations up. Moreover, it is able to handle massive heterogeneous data like anatomic shapes. However, this model is only descriptive and not explicative: the considered deformations do not arise from a statistical model and the knowledge we get through the optimization procedure cannot help us predict the evolution of a given subject in the future nor to deduct the evolution trajectory of the individual from each other. In the same vein, Yang et al. (2011) and Delor et al. (2013) introduce the notion of time warps. Both these models being non-generative, they suffer the same limitations as the one of Durrleman et al. (2009) nevertheless.

Lorenzi et al. (2015) propose a slightly different approach while studying brain senescence: their model is designed to be able to distinguish natural aging due to "anatomical" age from pathological aging. However, like in the models described above, the parameters involved do not arise from a statistical model and this model is thus purely descriptive. Moreover, this idea relies on the very specificity of brain senescence processes and is not easy to extend to other scope of application.

For the purpose of explaining the dependency between individual trajectories, Durrleman et al. (2013) generalized the model of Durrleman et al. (2009) in order to obtain a *generative* model. More precisely and with the previous notations, they assume that the data attachment term can be seen as $-\log p(S^i(t_j^i))$, where the distribution $p(S^i(t_j^i))$

would be a manifold-valued normal distribution centered in $\phi^i\left(\chi_{\psi^i(t_j^i)}(M_0)\right)$. In other words, they assume that data attachment term can be written as:

$$\frac{1}{2}\exp\left(-\frac{1}{2}\,d\left(\phi^i\left(\chi_{\psi^i(t_j^i)}(M_0)\right)\,,\,S^i(t_j^i)\right)\right).$$

The estimated functional remains unchanged and the authors use the same RKHS range of techniques while estimating the relevant morphisms.

As the time parametrization introduced in Durrleman et al. (2013) is statistically non-parametric, it is difficult to estimate. In order to remove this constraint, Hong et al. (2014) introduced parametric temporal deformations, logistic ones in this instance, leading to a *geodesic regression with parametric time warp model*. The algorithmic is then considerably simplified and is nothing but the alternation between a regression step that determines the mean evolution trajectory and a gradient descent step aiming for the estimation of the parameters which describe the temporal deformation. Unfortunately, this model only covers geodesic regression and do not generalize well to the study of longitudinal data.

Then, Su et al. (2014) also introduce a notion of time warp but toward a different end: their parametrization is used to geometrically align the trajectories. To do so, the authors assume that the observed trajectories possess an intrinsic unobserved parametrization they will try to learn through the estimation procedure. This model allows for an efficient comparison of the trajectories conceived as geometrical objects. Su et al. (2014) propose an application of it to visual recognition of speech. But still, this parametrization does not make sense from a modeling perspective and is therefore not interpretable.

Finally, summarizing the previous discussion, a model which claims to be efficient both from modeling and numerical perspectives has to be generative and to rely on a parametric notion of temporal deformation that makes sense as a time warp of the trajectory and is thus open to interpretation in an applicative framework.

In addition, proceed as Durrleman et al. (2009, 2013), namely estimate the template trajectory by a "max-max" algorithm that minimizes the sum of the squares of the distances after realignment between the data and the candidate template, *i.e.* that minimizes a matching functional similar to the one introduced in equation (2.5) is not satisfactory. Indeed, Devilliers et al. (2017,) proved that this algorithm was inconsistent because of noise. More precisely, they obtained an equivalent of the consistency bias as a function of the noise level. Thus, inconsistency is inevitable when the noise level is high enough, which is the usual situation in practice.

III – 2. Spatio-Temporal Mixed Effects Models

Schiratti et al. (2015, 2017) proposed a hierarchical model for the study of longitudinal

data, called *generic spatio-temporal model*. This model relies on the notion of parallel variations of a curve in a Riemannian manifold. We thus start by recalling briefly this notion. Indeed, this model no longer assumes that individuals follow a geodesic evolutionary trajectory but rather a trajectory parallel to the representative curve, which is still supposed to be geodesic.

We want to generalize to the Riemannian manifold case the notion of parallel lines. Let M a geodesically complete Riemannian manifold. Let $\gamma\colon I \subset \mathbb{R} \to M$ a differentiable curve on M, $t_0 \in I$ and $w_0 \in T_{\gamma(t_0)}M$ a tangent vector. Then, we call *parallel variation* of γ along the direction w the curve $\eta^w(\gamma;\cdot)\colon I \to M$ defined by

$$\eta^w(\gamma;\cdot)\colon t \;\mapsto\; \mathcal{E}xp_{\gamma(t)}\big(P_{\gamma,t_0,t}(w)\big)\,,$$

where $P_{\gamma,t_0,t}(w)$ is the parallel transport of the vector w, along the curve γ between the points $\gamma(t_0)$ and $\gamma(t)$. If in the euclidean case these curves are identical, in the Riemannian case this may not be the case: in other words, due to the curvature of the manifold, the parallel variation of a geodesic *a priori* does not have any reason to remain a geodesic. One only needs to look at the Earth, its equator and its different parallels including the two tropics to be convinced.

In particular, in order to compute the parallel variation of a given curve, we first have to transport the direction vector w along the curve γ, for all time t. Numerically, it may be very expensive; suitable numerical schemes were recently designed with the purpose of reducing the computational cost (Louis et al., 2017, 2018).

The model proposed by Schiratti et al. (2015, 2017) is to be compared to the model developed in Durrleman et al. (2009). Let the longitudinal dataset (y_i, t_i), where for each subject i, we observe $y_i = (y_{i,j})_{j\in\in\mathbb{R}^{k_i}}$ at times $t_i = (t_{i,j})_{j\in[\![1,k_i]\!]}$. The idea of Schiratti et al. (2015, 2017) is to see these observations as noisy samples along individual-specific trajectories, themselves arising from a representative path of the global evolution of the population through spatio-temporal deformations. Moreover, as explained previously and for numerical reasons, the authors want to build a parametric model.

To do so, they require the representative trajectory to be geodesic. This trajectory is then naturally parameterized with the triplet (t_0, p_0, v_0), where $t_0 \in \mathbb{R}$ is a reference time, $p_0 \in M$ the value of the curve at that time and $v_0 \in T_{p_0}M$ the value of its velocity vector (Gallot et al., 2004). Finally, the representative curve $\gamma_0\colon \mathbb{R} \to M$ writes

$$\gamma_0\colon\; t \mapsto \mathcal{E}xp_{p_0,t_0}(v_0)(t)\,,$$

where $\mathcal{E}xp_{p_0,t_0}(v_0)$ is the exponential map passing through p_0, at velocity v_0, at time t_0. At the individual scale, these trajectories are seen as parallel variations, with time warps, of this curve γ_0; that is to say, for each individual i, and given a vector $w_i \in T_{p_0}M$ and the couple $(\alpha_i, \tau_i) \in \mathbb{R}^2$,

$$\gamma_i\colon\; t \mapsto \eta^{w_i}\big(\gamma_0;\psi_i(t)\big)\,, \qquad \text{where} \qquad \psi_i\colon\; t \mapsto \alpha_i(t - t_0 - \tau_i) + t_0\,,$$

and where the random effects (α_i, τ_i) are to be interpreted as acceleration factors and individual time warps. In other words, Schiratti et al. (2015, 2017) allow each subject to follow its own spatial trajectory of evolution at its own pace α_i while being potentially ahead in time (or on the contrary late in time, according to its sign) of a value τ_i with respect to the mean trajectory. Figure 2.5 illustrates this construction. The statistical model is then given by

$$y_{i,j} \;=\; \eta^{w_i}\Big(\mathcal{E}xp_{p_0,t_0}(v_0)\;;\;\alpha_i(t_{i,j}-t_0-\tau_i)+t_0\Big)\;+\;\varepsilon_{i,j}$$

with $\varepsilon_{i,j}$ a Gaussian white noise independent and identically distributed. In order to have an identifiable model, the vectors w_i have to be chosen orthogonal to the trajectory γ_0. Moreover, in order to reduce the dimension of the estimated parameters space (Giraud, 2014), they are required to arise from a independent components statistical model (Hyvärinen et al., 2004). In other words, Schiratti et al. (2015, 2017) assume that the w_i are all linear combinations of independent sources and, instead of directly estimating w_i, they propose to estimate a design matrix A and some sources s_i such that $w_i = As_i$. Estimation of the parameters is performed through a well-defined *a posteriori maximum*, whose existence is proved by Schiratti (2016) in his PhD work, with the use of the MCMC-SAEM algorithm we introduced in Section I.3.

This model was notably applied to early detection of Alzheimer's disease and proved thereat to be efficient. The study of Bilgel et al. (2016) on β-amyloïde plaques, one of the hallmarks of Alzheimer's diseases, is another illustration of the applicability of this model.

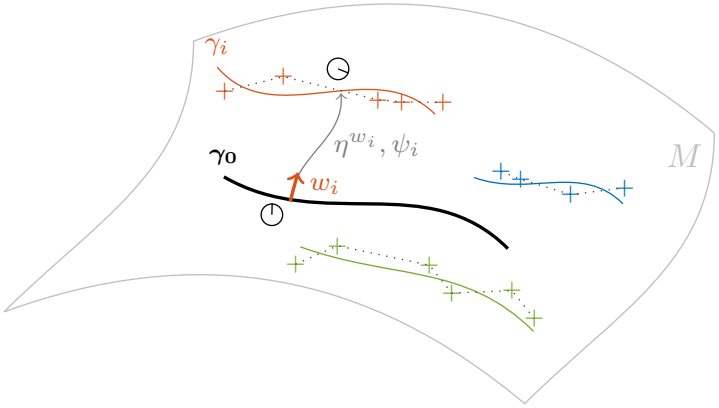


Figure 2.5 *– Generic spatio-temporel model of Schiratti et al. (2015, 2017).* We observe noisy samples along individual-specific trajectories, constructed as spatio-temporal deformations of a geodesic representative trajectory. In particular, each trajectory γ_i evolves with its own pace, given by the time warp ψ_i and its own geometry, given by the translation vector w_i.

In Section I.1, we debated the need to be able to estimate the time t_0. Let us check that in the unidirectional case, *i.e.* with $M = \mathbb{R}$, the model of Schiratti et al. (2015,

2017) actually differs from the random slope and intercept model. Assume that $\mathbb{R}$ is endowed with its usual metric. Then, the generic model writes, for each subject i and each observation j,

$$y_{i,j} \;=\; p_0 + v_0\,\alpha_i\,(t_{i,j} - t_0 - \tau_i) + \varepsilon_{i,j} \;=\; v_0\,\alpha_i\,(t_{i,j} - t_0) + p_0 - v_0\,\alpha_i\,\tau_i + \varepsilon_{i,j}\,.$$

This model is nonlinear for there is a product of the random effects α_i and τ_i. On the other side, the random slope and intercept model, with notations consistent with the ones of Schiratti et al. (2015, 2017), can be written in the form

$$y_{i,j} \;=\; (v_0 + \alpha_i)(t_{i,j} - t_0) + (p_0 + \beta_i) + \varepsilon_{i,j}\,.$$

More precisely, we can interpret these models as follows: the random slope and intercept model compare the observation distribution with a reference time t_0, while the model of Schiratti et al. (2015, 2017) compare the observations with respect to a reference measure value p_0 and thus allows for the estimation of time t_0.

In the same vein, Kim et al. (2017) also proposed a mixed effects model for the study of longitudinal data: the *Riemannian nonlinear mixed effects model*. It appears as a kind of generalization of the geodesic hierarchical model introduced by Muralidharan and Fletcher (2012) and described page 43. More precisely, given a dataset (y_i, t_i) arising from the observation of $y_i = (y_{i,j})_{j\in[\![1,k_i]\!]}$ at times $t_i = (t_{i,j})_{j\in[\![1,k_i]\!]}$, the model writes

$$\begin{cases} y_{i,j} \;=\; \mathcal{E}xp\Big(\,\mathcal{E}xp(b_i\,;\,\Gamma_{b,b_i}(v)\alpha_i(t_{i,j} - \tau_i - t_0))\,;\,\varepsilon_{i,j}\,\Big) \\ b_i \;=\; \mathcal{E}xp\,(b\,;\,u_i) \end{cases},$$

where $\Gamma_{b,b_i}(v) \in T_{b_i}M$ is the parallel transport of vector $v \in T_bM$ along a trajectory connecting b and b_i. In other words, if we denote γ_i such a geodesic, $\Gamma_{b,b_i} = P_{\gamma_i,0,1}$ with the previous notations. The model of Kim et al. (2017) shares an affine time warp with Schiratti et al. (2015, 2017). However and unlike the model of Schiratti et al. (2015, 2017), it is impossible to estimate exactly the parameters, because of the high complexity of the model. The authors thus introduce a technique of approximate estimation to cope with this issue. Lastly, this model was also applied to cortical atrophy.

The model of Schiratti et al. (2015, 2017) was recently improved in order to increase its applicability. Indeed, in the form we described the model may use high dimensional parameters, making their estimation complex. Koval et al. (2017) propose a first upgrade of the model for the study of networks, *i.e.* of measures varying over time on a fixed graph, and apply it to cortical atrophy (Koval et al., 2018).

Bône et al. (2018) propose an instance of the model of Schiratti et al. (2015, 2017) in the case of shapes in the LDDMM framework. As explained page 41, this model relies on a sparse representation of the group of deformations *via* control points (Durrleman et al., 2011, 2013). Let $c_0 \in \mathbb{R}^{n_{cp}d}$ a set of n_{cp} control points, a shape $y_0 \in M \subset \mathbb{R}^d$ and

a momentum $m_0 \in \mathbb{R}^{n_{cp}d}$. The representative path then writes $\gamma_0 \colon t \mapsto \mathcal{E}xp_{c_0,t_0,t}(m_0) \circ y_0$ and the individual trajectories are obtained through parameterized parallels of this representative curve; in other words:

$$\gamma_i \colon t \mapsto \eta^{w_i}(\psi_i(t)) \circ y_0 \,, \qquad \text{where} \qquad \eta^{w_i} \colon t \mapsto \mathcal{E}xp_{c(t),0,1}\left(P_t(w_i)\right),$$

and where, for all momentum vector w, $P_t(w)$ is the parallel transport of vector w, along the curve γ_0 between times t_0 and t. We denote $c(t) \in \mathbb{R}^{n_{cp}d}$ the control points at time t, namely $c(t) = \mathcal{E}xp_{c_0,t_0,t}(m_0) \circ c_0$. Like in the generic case, vectors w_i are assumed to be linear combinations of independent sources and the parameters of the model are estimated *via* a MCMC-SAEM algorithm.

Finally, a last upgrade of the generic model was proposed by Bône et al. (2019) with neural networks, whose use allows to handle very high dimensional data. The generic model is still the subject of intensive research; for example, the work of Debavelaere et al. (2019) seeks to include the generic model in a mixture model in order to classify individuals from a same cohort in different sub-populations. It permits, for example, to distinguish the evolution of healthy patients from the evolution of ill patients, *etc.*

IV. Thesis Outline

The model we propose in this manuscript builds on the generic model and seeks to generalize it to situations where the evolution dynamic is not univariate. Indeed, it is the standard situation in most applications. This is particularly the case in chemotherapy monitoring; we devote a full chapter to this situation: Chapter 5. When giving a patient a new treatment against cancer, he typically undergoes three phases: a decrease of tumors sizes while as the patient responds to therapy, then a stable phase and, most of the time, an escape from the treatments through an new increase in tumors size. Therefore, it is not realistic nor reasonable to model the representative trajectory of the population with a geodesic curve. We want to overcome this constraint with the use of a piecewise-geodesic representative path. Then, progression of the disease can have distincts evolution phases: one for each geodesic chunk in the representative curve.

Furthermore, despite being an innovative approach which grounded a new research fields, the generic model developed by Schiratti et al. (2015, 2017) suffers from a lack of theoretical guarantees, making it impossible to conclude about the reliability of their results otherwise than through graphical validation. Hence, we focused on proving consistency of our model, getting as a byproduct a consistency result for the model of Schiratti et al. (2015, 2017) and its extensions we described in the previous section.

Following standard methods for statistical inference in nonlinear mixed effects models, the parameters of our model are estimated through a MCMC-SAEM algorithm. However, our numerical experiments suffered from technical limitations such as a high sensibility of the SAEM algorithm regarding its initial conditions. We then worked

on a potential upgrade of this algorithm building on simulated annealing or tempering techniques.

Finally, the work presented in this document is at a crossroad in what we believe is a basis of mathematical modeling in medicine: a combination of clinical applications, with for instance an ongoing collaboration with oncologists and radiologists at the Georges Pompidou European hospital (HEGP for hôpital européen Gorges Pompidou in French), mathematical theory as a guarantor of proposed models reliability and the development of efficient numerical tools allowing for the analysis of complex heterogeneous datasets which become increasingly massive as imaging techniques improve.

IV – 1. Overview of the Chapters

This manuscript is divided in two parts: the first one consists in the establishment of a coherent framework for the analysis of manifold-valued longitudinal data; the second one focuses on the algorithmic aspect by proposing a new class of stochastic approximations of the EM algorithm. More precisely, the chapters are organized as follows.

Thereafter, the chapter with symbol ♢ correspond to our contribution.

Chapter 3. ♢ The first chapter is dedicated to the establishment of a coherent framework for the statistical analysis of manifold-valued longitudinal data. For this purpose, we propose a generalization of the model introduced by Schiratti et al. (2015, 2017) in order to encompass the case of longitudinal data undergoing multivariate evolution dynamics. This model relies on the distinction of temporal deformations, linked with the acquisition process of the data and the pace of evolution of the observed phenomenon, versus space deformations, linked with intrinsic geometry of observed shapes.

To do so, we propose a nonlinear mixed effects model whose individual evolution trajectories are seen as spatio-temporal deformations of an average path representative of the evolution at the macroscopic population scale. We present this model under fairly generic hypotheses in order to encompass a wide class of more specific models.

Estimation of the parameters of the geometric model is performed through a *maximum a posteriori* whose existence and consistency are proved; in other words, we prove that the larger the sample, the closer the estimator of the *maximum a posteriori* is from the real values of the parameters. This last result is all the more important given that the proposed model include the model of Schiratti et al. (2015, 2017) as a particular case. The later being applied to early detection of Alzheimer's disease, we thus provide theoretical guarantees on ongoing pre-clinic studies.

This chapter is derived from Chevallier et al. (2017) and Chevallier et al. (2019).

Chapter 4. In order to allow for models dealing with 3-dimensional anatomical shapes, this chapter provides the mathematical foundations of two shape spaces widely used in this context: the currents (Vaillant and Glaunès, 2005) and the varifolds (Charon and

Trouvé, 2013).

Chapter 5. ♢ Then, the model proposed in Chapter 3 is able to quantify the pace of progression of a continuous process. We apply this model to chemotherapy monitoring and especially to follow-up of metastatic kidney cancer. Indeed, in this context, understanding the pace of progression of the cancer is at the heart of medical care.

The first application consists in monitoring RECIST scores. These scores being scalar data, we realize an instantiation of the generic model for real bounded data on segment $[0, 1]$ endowed with logistic metric. This model was elaborated in collaboration with oncologists and radiologists of the Georges Pompidou European hospital (HEGP for hôpital européen Gorges Pompidou in French). Numerical experiments on synthetic and real data validate its relevance.

The second application consists in follow-up of 3D anatomical shapes, again for evaluation of tumors response. This model relies on the notion of large deformations we discussed in the introduction and can be applied to currents (Vaillant and Glaunès, 2005) and varifolds (Charon and Trouvé, 2013), which are standard shape spaces for the analysis of anatomical data and whose mathematical background was recalled at Chapter 4. We also propose numerical experiments on synthetic data.

This chapter is partly derived from Chevallier et al. (2019).

Chapter 6. Numerically, estimation of the parameters is performed through a stochastic approximation of the EM algorithm, namely the SAEM algorithm. Before studying in depht this algorithm, we present in this chapter the classical literature on the EM algorithm and its usual variants. We pay particular attention to the seminal paper of Delyon et al. (1999).

Chapter 7. ♢ Despite numerical performance of the SAEM algorithm, and because of the complexity of our model, our numerical experiments suffered from limitations. In particular, the SAEM algorithm is very sensitive to its initial condition, and, even with the additional stochasticity associated to its approximation feature, it can be trapped in local minima. Moreover, the SAEM assumes that we are able to draw from the conditional distribution of latent variables given the observations (with the latent variable models terminology), possibly with MCMC-type method, which is not always the case.

We propose in this chapter a new class of SAEM algorithms: the approximated-SAEM algorithms, for which we prove convergence toward local maxima under standard assumptions. This class relies on sampling from an approximation, in a sense to be specified, of the real conditional law in the simulation step. In particular, we encompass pre-existent algorithms like the ABC-SAEM (Picchini and Samson, 2018) whose numerical efficiency was empirically established but whose theoretical convergence was not proved.

Lastly, building on simulated annealing techniques, we propose a tempered version of

the SAEM algorithm in order to favor convergence toward global minima. In this version, we approximate the conditional law by tempering it following a damped oscillatory temperature scheme. This method is applied to the estimation of parameters in Gaussian mixture models and we illustrate its numerical superiority over SAEM algorithm. This algorithm is also applied to blind source separation through independent factor analysis.

This chapter is derived from Allassonnière and Chevallier (2019).

IV – 2. List of Publications

This manuscript led to the following publications.

a. Conference Proceeding

- Chevallier et al. (2017): **Learning spatio-temporal piecewise-geodesic trajectories from longitudinal manifold-valued data**. Juliette Chevallier, Stéphane Oudard and Stéphanie Allassonnière. In *Advances in Neural Information Processing Systems* 30, December 2017.

b. Preprints

- Allassonnière and Chevallier (2019): **A new class of EM algorithms. Escaping local minima and handling intractable sampling**. Juliette Chevallier and Stéphanie Allassonnière. June 2019, HAL-02044722.
- Chevallier et al. (2019): **A coherent framework for learning spatio-temporal piecewise-geodesic trajectories from longitudinal manifold-valued data**. Juliette Chevallier, Vianney Debavelaere and Stéphanie Allassonnière. May 2019, HAL-01646298.

Acknowledgments.

This work was supported by a public grant as part of the Investissement d'avenir, project reference ANR-11-LABX-0056-LMH.

– Deuxième partie –

A Coherent Framework for Longitudinal Observations on a Riemannian Manifold

Motivations and State of the Art

Longitudinal studies are powerful tools to achieve a better understanding of temporal progressions of biological or natural phenomenons. For instance, longitudinal psychometric data are often used to explore differences in the progression of Alzheimer's and more generally neurodegenerative diseases. Other important applications such as pattern recognition, disease and treatment monitoring, study of face expression dynamics, *etc.* come also from longitudinal studies.

Moreover, efforts in medicine and medical follow-up rely more and more on the understanding of the global disease progression and not only on punctual states of health, often with the help of medical images. In order to provide everyone the best possible treatment, there is a need for prediction methods that allow to grasp quickly the efficiency of a possible treatment without doing invasive measures, such as biopsies. Hence, designing models that deal with medical images and more generally extracted features and shapes from this images is very important for application-related uses.

Reaction-diffusion based tumor growth models have demonstrated their efficiency for cancer monitoring (Konukoglu et al., 2010; Rekik et al., 2013). However, such methods cover images but not shapes or whatever else type of data, e.g. scores. Moreover, even for images, as these models rely on reaction-diffusion equations, they can only apply for situations in which the observed phenomenon is linked to diffusion dynamics. As a consequence, these models can apply to untreated cancer but not to treated-one nor to multiple sclerosis monitoring, neurodegenerative disease follow-up, like Alzheimer's or Parkinson's diseases, or more complicated framework. Likewise, coupling nonlinear partial differential equation (PDE) models and optimization is efficient to make prediction when patients are only under monitoring, *i.e.* without treatment (Colin et al., 2010; Saut et al., 2014). The idea of these methods is to model the tumor growth with a "simple" PDE model, involving few parameters which are estimated from series of CT-scan or MRI. For therapy planing purpose, such methods can reinforce the decision to wait without specific treatment, in case of slow progression of the tumor, thus preventing heavy treatment. However, they cannot predict the response to a given molecule and so help in the choice of the appropriate one, nor in the choice of an appropriate sequence of molecules.

Anatomical data – and most of structured data – have been successfully modeled as points on a Riemannian manifold, *i.e.* as points on a smooth manifold equipped with a Riemannian metric tensor. These spaces are often called *shape spaces*. The choice of the shape space and its metric tensor is driven by the type of data in the study (Charlier et al., 2017; Charon and Trouvé, 2013; Klassen et al., 2004; Trouvé and Younes, 2015; Vaillant and Glaunès, 2005). Geometrical properties of shape manifolds have been properly defined over the last decades. Moreover, according to the Whitney embedding theorem (Gallot et al., 2004), as the shape spaces are second-countable, they are always embedded in a real d-dimensional Euclidean space, the space of measurements, which leads us to consider the shape manifold as a submanifold of this Euclidean space. Therefore, the temporal evolution of empirical data may be modeled as a parametric curve in the space of measurements and more precisely as a noisy version of an underlying parametric curve living on the Riemannian shape submanifold. Given a cohort of individuals followed over a time period, we thus observe discrete samples of such a curve for each subject. We call this set of observations a longitudinal data set. Figure 3.1 illustrates this understanding of the data.

Mixed effects models have proved their efficiency in the study of longitudinal data sets (Laird and Ware, 1982), especially for medical purposes (Milliken and Edland, 2000; Ribba et al., 2014). Indeed, mixed effects models provide a general and flexible framework to study complex data which depends on unobserved variables, such as longitudinal data sets. They consist of two parts: *fixed effects* which describe the data at the population level and *random effects* which are associated with individual experimental units. In the framework of longitudinal analysis, through mixed effects models, one can explain in two steps, both a representative path of the evolution of the whole population and individual-specific progressions. Given a longitudinal data set, a representative trajectory and its variability are first estimated. Then, we can define subject-specific trajectories according to global progression.

The temporal alignment in longitudinal data analysis is an efficient way to compare trajectories (Su et al., 2014,). Here, the authors propose to use the temporal registration to align the different trajectories. Despite good results for comparing trajectories, the interpretation of the temporal parameterization is lost. However, within medical applications, the time parametrization reveals information on the data patient's state of health and has to be considered.

The recent generic approach of Schiratti et al. (2015, 2017) to align patients takes all this into account. This model was built with the aim of granting *temporal* and *spatial* inter-subject variability through individual variations of a common time-line and parallel shifting of a representative trajectory. Each individual trajectory has its own intrinsic geometric pattern through spatial variability and its own time parametrization through time variability. In term of modeling, the time variability allows some individuals to follow the same progression path but at a different age and with possibly a different

pace. Schiratti et al. (2015, 2017) have made a strong hypothesis to build their model as they assume the characteristic evolution to be geodesic, *i.e.* that the characteristic trajectory is the shortest path between the initial representative state and the final one. This hypothesis helps for the parametrization of the model which becomes both generic, *i.e.* allowing for many different types of data, and numerically estimable. However, such an assumption significantly reduces the effective framework of their model. Like the described-above PDEs based models, such a model can be applied to situation with a unique dynamic like neurodegenerative disease, but not to situations in which the dynamic can fluctuate. For instance, this model cannot be used for multiple sclerosis monitoring in which the disease progression is accompanied by recession nor for monitoring tumor regression or recurrence in response to treatments.

Here, we will relax this assumption to make the model applicable to a wider variety of situations and data sets: we address each situation in which the evolution can fluctuate several times.

Estimation is formulated as a well-defined *Maximum A Posteriori* (MAP) problem which we prove to be consistent under mild assumptions. Numerically, the MAP estimation of the parameters is performed through the expectation-maximization (EM) algorithm (Dempster et al., 1977). However, as our model is strongly nonlinear and we have to use a stochastic version of the the EM algorithm, namely the Markov chain Monte Carlo stochastic approximation expectation-maximization (MCMC-SAEM) algorithm (Lavielle, 2014). Theoretical results regarding its convergence have been proved by Delyon et al. (1999) and Allassonnière et al. (2010) and its numerical efficiency has been demonstrated for these types of models (Schiratti et al. (2015, 2017), Monolix – `http://lixoft.com/`).

Due to the versatility of the Riemannian geometry, the proposed model provides a comprehensive support for a wide range of practical situations, from unidimensional data to shape analysis. Moreover, the same algorithm can be used in all these situations.

This part is based on Chevallier et al. (2019) and organized as follows: Chapter 3 is dedicated to our generic nonlinear mixed effects model for piecewise-geodesically distributed data and Chapter 5 to instantiations of this generic framework.

In Section I of Chapter 3, we detail the construction of our generic model. Riemannian geometry allows us to derive a method that makes light assumptions about the data and applications we are able to deal with. In Section II and III of the same chapter, we explain how to use the MCMC-SAEM algorithm to produce MAP and prove a consistency theorem.

In order to allow for 3-dimensional shapes models in Chapter 5 we briefly cover the mathematical foundations of currents and varifolds in Chapter 4.

Last, in Chapter 5, we then make the generic formulation explicit for one-dimension manifolds and piecewise logistically distributed data in Section I and for shape analysis in Section II. These two particular cases are built in the target of chemotherapy monitoring.

In Section III of Chapter 5, some experiments are performed for the piecewise logistic model and for the piecewise-geodesic shape model: both on synthetic and on real data from the hôpital européen Georges Pompidou (HEGP, Georges Pompidou European hospital) for the piecewise logistic model. These experiments highlight the robustness of our model to noise and its performance in understanding individual paths of progression.

– Chapter III –

Learning Spatio-temporal Piecewise-Geodesic Trajectories from Longitudinal Manifold-Valued Data

Ce premier chapitre est dédié à l'établissement d'un cadre d'étude cohérent pour l'analyse statistique de données longitudinales à valeurs sur des variétés riemanniennes. Dans ce but, nous proposons une généralisation du modèle introduit par Schiratti et al. (2015, 2017) pour l'étude de données longitudinales à dynamiques d'évolution multiples. Ce modèle repose sur la discrimination de déformations temporelles, liées à l'acquisition des données et au rythme de progression du phénomène observé, de déformations spatiales, liées à la géométrie intrinsèque des formes observées.

Pour ce faire, nous proposons un modèle non-linéaire à effets mixtes dans lequel les trajectoires individuelles d'évolution sont vues comme des déformations spatio-temporelles d'une trajectoire moyenne représentative de l'évolution de la population à l'échelle macroscopique. Nous présentons ce modèle sous des hypothèses très génériques afin d'englober une grande classe de modèles plus spécifiques.

L'estimation des paramètres du modèle géométrique est réalisée par un *maximum a posteriori* dont on démontre l'existence et la consistance ; autrement dit, on démontre que l'estimateur du *maximum a posteriori* s'éloigne d'autant moins de la vraie valeur des paramètres que la taille de l'échantillon est grande. Ce dernier résultat est d'autant plus important que le modèle proposé dans ce chapitre englobe le modèle de Schiratti et al. (2015, 2017), lui même appliqué à la détection précoce de la maladie d'Alzheimer. Ainsi, on fournit des garanties théoriques sur des études pré-cliniques déjà en cours.

Contents

In the following, we describe a *generic* method to build mixed effects models for piecewise-geodesically distributed data. This leads us to a large variety of possible situations that we will be able to deal with within the same framework. This model has been first introduced in Chevallier et al. (2017) and extended in Chevallier et al. (2019). We then prove the existence of the *maximum a posteriori* estimate and its asymptotic consistency under reasonable assumptions.

I. Generic Mixed Effects Model for Piecewise-Geodesically Distributed Data

We consider a longitudinal data set $\boldsymbol{y}$ obtained by repeating multivariate measurements of $n \in \mathbb{N}^*$ individuals. Each individual $i \in [\![1, n]\!]$ is observed $k_i \in \mathbb{N}^*$ times, at the time points $\boldsymbol{t}_i = (t_{i,j})_{j \in [\![1,k_i]\!]}$, and we denote $\boldsymbol{y}_i = (y_{i,j})_{j \in [\![1,k_i]\!]}$ the sequence of observations for this individual. We also denote $k = \sum_{i=1}^{n} k_i$ the total numbers of observations and assume that each observation $y_{i,j}$ is a point of $\mathbb{R}^d$, where $d \in \mathbb{N}$ can be considered as the dimension of the problem. Thus, our observed data consists in a sequence in $\mathbb{R}^{kd}$, $\boldsymbol{y} = (y_{i,j})_{(i,j) \in [\![1,n]\!] \times [\![1,k_i]\!]}$, where $[\![1, n]\!] \times [\![1, k_i]\!]$ denotes for compactness the set $\{(i, j) | i \in [\![1, n]\!] \wedge j \in [\![1, k_i]\!]\}$.

We generalize the idea of Schiratti et al. (2015, 2017) and build our model in a hierarchical way. Our data points are seen as noisy samples along trajectories and we suppose that each individual trajectory derives from a group-representative scenario through spatio-temporal transformations. Key to our model is that the group-representative trajectory in no longer assumed to be geodesic but piecewise-geodesic. Thus, the characteristic trajectory is no more the shortest path between the initial and the final representative states but a concatenation of shortest paths between several intermediate states. In particular, this allows us to consider situations in which the evolution can fluctuate. We present at Figure 3.1 an example of situation our generic model can address.

To ensure that the optimization of those trajectories can be computationally performed in a reasonable amount of time, we build a parametric model. That is to say that the trajectories depend on a finite number of variables. In the following (see Subsection I.3), we will denote $\boldsymbol{z}_{\text{pop}}$ the variables driving the group-representative scenario and $\boldsymbol{z}_i$ those associated to the individual $i \in [\![1, n]\!]$. For the sake of clarity, we first detail the construction of the trajectories from a geometrical point of view. Then, we state our generative model in a statistical perspective.

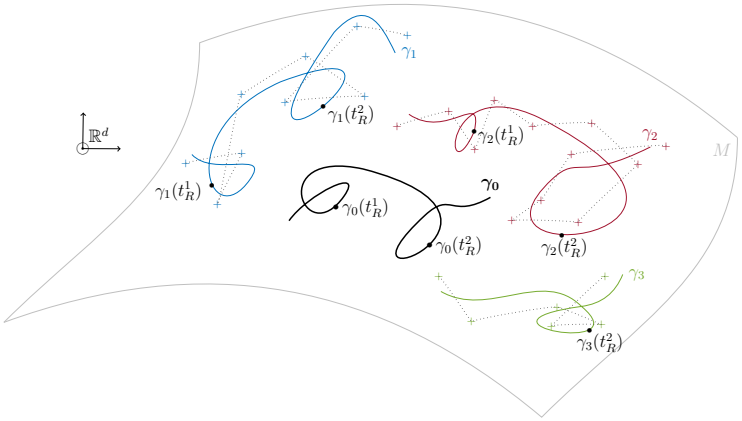


***Figure 3.1** – The generic piecewise-geodesic curve model.*
The observed data (crosses) consist in noisy samples along manifold-valued trajectories. Each individual path γ_i (solid colored lines) is a spatio-temporal variation of a piecewise-geodesic representative trajectory γ_0 (bold black line). In particular, the individual trajectories are not necessarily piecewise-geodesic.

I – 1. The Group-Representative Trajectory

Let $m \in \mathbb{N}^*$ and $\boldsymbol{t_R} = \left(-\infty < t_R^1 < \ldots < t_R^{m-1} < +\infty\right)$ a subdivision of $\mathbb{R}$, called the *breaking-up times* sequence. In order the representative trajectory γ_0 to be geodesic on each of the m sub-intervals of $\boldsymbol{t_R}$, we build γ_0 component by component.

a. A Piecewise-Geodesic Curve

In this context, let M_0 be a geodesically complete submanifold of $\mathbb{R}^d$, $\left(\bar{\gamma}_0^\ell\right)_{\ell\in[\![1,m]\!]}$ a family of geodesics on M_0 and $\left(\phi_0^\ell\right)_{\ell\in[\![1,m]\!]}$ a family of isometries defined on M_0. For all $\ell \in [\![1,m]\!]$, we set $M_0^\ell = \phi_0^\ell(M_0)$ and $\gamma_0^\ell = \phi_0^\ell \circ \bar{\gamma_0}^\ell$. The isometric nature of the mapping ϕ_0^ℓ ensures that the manifolds $(M_0^\ell)_{\ell\in[\![1,m]\!]}$ remain Riemannian and that the curves $\gamma_0^\ell \colon \mathbb{R} \to M_0^\ell$ remain geodesic. In particular, each γ_0^ℓ remains parametrizable (Gallot et al., 2004). We define the representative trajectory γ_0 by

$$\forall t \in \mathbb{R}, \quad \gamma_0(t) \;=\; \gamma_0^1(t)\,\mathbb{1}_{]-\infty,t_R^1]}(t) \;+\; \sum_{\ell=2}^{m-1} \gamma_0^\ell(t)\,\mathbb{1}_{]t_R^{\ell-1},t_R^\ell]}(t) \;+\; \gamma_0^m(t)\,\mathbb{1}_{]t_R^{m-1},+\infty[}(t)\,.$$

In other words, given a *manifold-template* of the geodesic components M_0, we build γ_0 so that the restriction of γ_0 to each sub-interval of $\boldsymbol{t_R}$ is the deformation of a geodesic curve $\bar{\gamma}_0^\ell$ living on M_0 by the corresponding isometry ϕ_0^ℓ. In practice, M_0 is chosen in order to catch the geometric nature of the observed data : if we are studying a score as in Section I, M_0 will be the standard finite segment $]0,1[$ for instance. The choice of

the isometries ϕ_0^ℓ and the geodesics $\bar{\gamma}_0^\ell$ have to be done with the aim of having an "as regular as possible" (at least continuous) curve γ_0 at the breaking-up time points t_R^ℓ. In the following section, we propose a way to meet this criterion in one dimension and in the shape framework. However, the freedoms in the choice of ϕ_0^ℓ and $\bar{\gamma}_0^\ell$ induce a wide panel of models.

b. Boundary Conditions

Because of the piecewise nature of our representative trajectory γ_0, constraints have to be formulated on each interval of the subdivision $\boldsymbol{t_R}$. Following the formulation of the *local existence and uniqueness theorem* (Gallot et al., 2004), constraints on geodesics are generally formulated by forcing a value and a tangent vector at a given time-point. However, as soon as there is more that one geodesic component, *i.e.* $m > 1$, such an approach cannot ensure the curve γ_0 to be at least continuous. That is why we re-formulate these constraints in our model as boundary conditions. Let $\bar{\boldsymbol{A}} = (\bar{A}^0, \dots, \bar{A}^m) \in (M_0)^{m+1}$. Let $t_0 \in \mathbb{R}$ be a real value representing an initial time and $t_1 \in \mathbb{R}$ representing a final one. We impose that $\bar{\gamma}_0^1(t_0) = \bar{A}^0$, $\bar{\gamma}_0^m(t_1) = \bar{A}^m$ and that

$$\forall \ell \in [\![1, m-1]\!], \quad \bar{\gamma}_0^\ell(t_R^\ell) = \bar{A}^\ell \qquad \text{and} \qquad \bar{\gamma}_0^{\ell+1}(t_R^\ell) = \bar{A}^\ell \,.$$

Note that we can apply the constraints on γ_0^ℓ instead of $\bar{\gamma}_0^\ell$ by defining $A^\ell = \phi_0^\ell(\bar{A}^\ell)$ for each ℓ. Notably, the $2m$ constraints are defined step by step. In the case where the geodesics could be written explicitly, such constraints do not complicate the model. In more complicated case, as the one shown for shapes in Section II of Chapter 5, we use shooting or matching methods to enforce this constraints.

From this representative curve, we derive a modeling of the individual trajectories that mimics the individual evolution of subjects and best fits the individual observations.

I – 2. Individual Trajectories: Space and Time Warping

We want the individual trajectories to represent a wide variety of behaviors and to derive from the group characteristic path by spatio-temporal transformations. To do that, we define for each component of the piecewise-geodesic curve γ_0 a couple of transformations: the *diffeomorphic component deformations* and the *time component reparametrizations* which characterize respectively the spatial and the temporal variability of propagation among the population. Moreover, we decree as few constraints as possible in the construction: at least continuity and control of the slopes at the (individual) breaking-up points.

a. Time Component Reparametrizations

For compactness, we denote t_0 by t_R^0 from now on.

To allow different paces in the progression and different rupture times for each individual, we introduce some temporal transformations $\psi_i^\ell \colon \mathbb{R} \to \mathbb{R}$, called *time-warp*, that are defined for the subject $i \in [\![1, n]\!]$ and for the geodesic component $\ell \in [\![1, m]\!]$ by

$$\psi_i^\ell(t) \;=\; \psi_{i\,(\alpha_i^\ell, \tau_i^\ell)}^\ell(t) \;=\; \alpha_i^\ell(t - t_R^{\ell-1} - \tau_i^\ell) + t_R^{\ell-1}\,, \quad \text{where} \quad (\alpha_i^\ell, \tau_i^\ell) \in \mathbb{R}^+ \times \mathbb{R}\,.$$

The parameters τ_i^ℓ correspond to the time-shifts between the representative and the individual progression onset ; the α_i^ℓ are the acceleration factors that describe the pace of individuals, being faster or slower than the population characteristic. For all individuals $i \in [\![1, n]\!]$, let $\boldsymbol{t_{R,i}} = (t_{R,i}^\ell)_{\ell \in [\![1, m-1]\!]}$ denote the individual sequence of rupture times which is the subdivision of $\mathbb{R}$ such that for all $\ell \in [\![1, m-1]\!]$, $\psi_i^\ell(t_{R,i}^\ell) = t_R^\ell$ *i.e.* such that

$$t_{R,i}^\ell \;=\; t_{R,i\,(\alpha_i^\ell, \tau_i^\ell)}^\ell \;=\; t_R^{\ell-1} + \tau_i^\ell + \frac{t_R^\ell - t_R^{\ell-1}}{\alpha_i^\ell}\,.$$

To ensure good adjunction at the rupture times, we demand that for all $\ell \in [\![1, m]\!]$, $\psi_i^\ell(t_{R,i}^{\ell-1}) = t_R^{\ell-1}$. Hence the time reparametrizations are constrained and only the acceleration factors α_i^ℓ and the first time shift τ_i^1 are free: all other time shift, $\ell \in [\![2, m]\!]$, are defined by $\tau_i^\ell = t_{R,i}^{\ell-1} - t_R^{\ell-1}$.

In the following, we will sometimes refer to the individual initial and final times which are defined, for all $i \in [\![1, n]\!]$, by $t_0^i = t_0 + \tau_1$ and $t_1^i = t_R^{m-1} + \tau_i^m + \frac{t_1 - t_R^{m-1}}{\alpha_i^m}$.

b. Diffeomorphic Component Deformations

Concerning the space variability, we introduce m diffeomorphisms $\phi_i^\ell \colon M_0^\ell \to \phi_i^\ell(M_0^\ell)$ to enable the different components of the individual trajectories to vary irrespectively of each other. We enforce the adjunctions to be at least continuous and therefore the mappings ϕ_i^ℓ to satisfy

$$\forall \ell \in [\![1, m-1]\!], \quad \phi_i^\ell \circ \gamma_0^\ell(t_R^\ell) \;=\; \phi_i^{\ell+1} \circ \gamma_0^{\ell+1}(t_R^\ell)\,.$$

Note that, as the individual paths are no longer required to be geodesic, the mappings ϕ_i^ℓ do not need to be isometric.

For all individuals $i \in [\![1, n]\!]$ and all component $\ell \in [\![1, m]\!]$, we set $\gamma_i^\ell = \phi_i^\ell \circ \gamma_0^\ell \circ \psi_i^\ell$ and define the corresponding individual curve γ_i by

$$\forall t \in \mathbb{R}, \quad \gamma_i(t) \;=\; \gamma_i^1(t)\, \mathbb{1}_{]-\infty, t_{R,i}^1]}(t) \;+\; \sum_{\ell=2}^{m-1} \gamma_i^\ell(t)\, \mathbb{1}_{]t_{R,i}^{\ell-1}, t_{R,i}^\ell]}(t) \;+\; \gamma_i^m(t)\, \mathbb{1}_{]t_{R,i}^{m-1}, +\infty[}(t)\,.$$

Finally, the observations $\boldsymbol{y}_i = (y_{i,j})_{j \in [\![1, k_i]\!]}$ are assumed to be distributed along the curve γ_i and perturbed by an additive Gaussian noise $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2 I_{k_i d})$, where $\sigma \in \mathbb{R}^+$:

$$\forall (i, j) \in [\![1, n]\!] \times [\![1, k_i]\!], \quad y_{i,j} \;=\; \gamma_i(t_{i,j}) \;+\; \varepsilon_{i,j}\,, \quad \text{where} \quad \varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2 I_d)\,.$$

By construction, for each $(i,j) \in [\![1,n]\!] \times [\![1,k_i]\!]$, there exist $\ell \in [\![1,m]\!]$ such that $\gamma_i(t_{i,j})$ lies on the submanifold $\phi_i^\ell(M_0^\ell)$ of $\mathbb{R}^d$. Thus, the previous sum is well-defined. In particular, we do not assume that the noisy-observation remain on the manifold.

The choice of the isometries ϕ_0^ℓ and the diffeomorphisms ϕ_i^ℓ induces a large range of piecewise-geodesic models. For example, if $m = 1$, $\phi_0^1 = Id$ and if ϕ_i^1 denotes the application that maps a curve onto its parallel curve for a given non-zero tangent vector $\boldsymbol{w}_i$, we feature the model proposed by Schiratti et al. (2015, 2017). In Chapter 5, we propose two other specific models which can be used for chemotherapy monitoring for instance.

I – 3. Toward a Coherent and Tractable Statistical Generative Model

Now that we have presented the geometrical objects we will deal with, we set up a comprehensive statistical framework in order to estimate the different parameters that control the trajectories.

We first introduce some notations in order to clearly state our statistical generative model. Let $\boldsymbol{z}_i^\psi = (\alpha_i^\ell, \tau_i^\ell)_{\ell \in [\![1,m]\!]}$ denote the individual temporal variables and similarly $\boldsymbol{z}_i^\phi$ denote the individual spatial variables, *i.e.* the variables associated to the variation of the m diffeomorphic deformations ϕ_i^ℓ. Likewise, let $\boldsymbol{z}_{\text{pop}}$ denote the population variable, *i.e.* the variable associated to the variation of the m isometric mappings ϕ_0^ℓ.

Let $p_{\text{ind}} \in \mathbb{N}$ be the dimension of each vector $\boldsymbol{z}_i = (\boldsymbol{z}_i^\psi, \boldsymbol{z}_i^\phi)$ such that $\forall i \in [\![1,n]\!]$, $\mathcal{Z}_i \subset \mathbb{R}^{p_{\text{ind}}}$ denotes the space of random effects. Similarly, let $p_{\text{pop}} \in \mathbb{N}$ be the dimension of $\boldsymbol{z}_{\text{pop}}$ and $\mathcal{Z}_{\text{pop}} \subset \mathbb{R}^{p_{\text{pop}}}$ denotes the space of fixed effects.

To cover many situations, we do not explicit here the individual spatial variables $\boldsymbol{z}_i^\phi$. However, for examples, we propose an instantiation of this generic model for one dimension manifolds and piecewise-logistically distributed data at Section I and for shape analysis at Section II of Chapter 5. Moreover, our generic approach encompass a large variety of models as such proposed by Schiratti et al. (2017), Bône et al. (2018) and Koval et al. (2018).

a. Modeling Constraints...

In a modeling perspective, we are interested in understanding the individual behaviors with respect to the characteristic one. Thus, we focus on the variance of the random effects $\boldsymbol{z}_i = (\boldsymbol{z}_i^\psi, \boldsymbol{z}_i^\phi)$ rather than their distributions. Moreover, as we want the representative path to characterize the pattern of behavior of the individual trajectories, we have to slightly modify the individual parameters $\boldsymbol{z}_i$ in such a way that for all i, $\mathbb{E}(\boldsymbol{z}_i) = 0$. In particular, if our model were linear, this would have ensure the representative trajectory to be the mean (in the statistical meaning) of the individual ones. Concerning the individual temporal variables for instance, the acceleration parameters $(\alpha_i^\ell)_{\ell \in [\![1,m]\!]}$ have to be positive and equal to one on average while the time shifts $(\tau_i^\ell)_{\ell \in [\![1,m]\!]}$ are of any signs

and must be zero on average. For these reasons, we set $\alpha_i^\ell = \mathrm{e}^{\xi_i^\ell}$ and consider the "new" temporal variable, still denoted $\boldsymbol{z}_i^\psi$ for compactness, $\boldsymbol{z}_i^\psi = (\xi_i^\ell, \tau_i^\ell)_{\ell \in [\![1,m]\!]}$. We proceed in the same way for the individual spatial variables z_i^ϕ, when required (for centered or positive variables).

To sum up, we assume that there exists a symmetric positive definite matrix $\Sigma \in \mathscr{S}^+_{p_{\text{ind}}}(\mathbb{R})$ such that $\boldsymbol{z}_i \sim \mathcal{N}(0, \Sigma)$, and now want to estimate Σ. Hence, the parameters we are interested in are $\theta = (\boldsymbol{z}_{\text{pop}}, \Sigma, \sigma) \in \mathcal{Z}_{\text{pop}} \times \mathscr{S}^+_{p_{\text{ind}}}(\mathbb{R}) \times \mathbb{R}^+$.

b. ...and Computational Feasibility

Given a n-sample, we target $\hat{\theta}_n$ an estimation of our parameters. Following the classical approach for maximum likelihood estimation in nonlinear mixed effects models, we use the MCMC-SAEM algorithm. However, the theoretical convergence of this algorithm is proved only if the model belongs to the curved exponential family (Allassonnière et al., 2010; Delyon et al., 1999). This framework is also important for numerical performances. Without further hypothesis, our model does not satisfy this constraint. Therefore, we proceed as in Kuhn and Lavielle (2005): we assume that $\boldsymbol{z}_{\text{pop}}$ is the realization of independent Gaussian random variables with fixed small variances and estimate the means of those variables. So, the parameters we want to estimate are $\theta = (\overline{\boldsymbol{z}_{\text{pop}}}, \Sigma, \sigma)$ and we define the set of admissible parameters by $\Theta = \mathbb{R}^{p_{\text{pop}}} \times \mathscr{S}^+_{p_{\text{ind}}}(\mathbb{R}) \times \mathbb{R}^+$.

The fixed and random effects $\boldsymbol{z} = (\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)_{i \in [\![1,n]\!]}$ are considered as latent variables, *i.e.* as hidden variables that are not directly observed but can be inferred by the observations. Our model writes in a hierarchical way as

$$\begin{cases} y \,|\, z, \theta \;\sim \displaystyle\bigotimes_{i=1}^{n} \bigotimes_{j=1}^{k_i} \mathcal{N}\big(\, \gamma_i(t_{i,j}) \,,\, \sigma^2 \,\big) \,, \\ z \,|\, \theta \;\sim \mathcal{N}\big(\, \overline{\boldsymbol{z}_{\text{pop}}} \,,\, D_{\text{pop}}^{-1} \,\big) \displaystyle\bigotimes_{i=1}^{n} \mathcal{N}(0, \Sigma) \,, \end{cases}$$

where $\boldsymbol{\sigma}_{\text{pop}} \in \mathbb{R}_+^{p_{\text{pop}}}$ is an hyperparameter of the model and $D_{\text{pop}} = \boldsymbol{\sigma}_{\text{pop}}^2 I_{p_{\text{pop}}} \in \mathscr{M}_{p_{\text{pop}}}(\mathbb{R})$ is the diagonal matrix of size p_{pop} whose diagonal entries are given by the vector $\boldsymbol{\sigma}_{\text{pop}}^2$. The products $\otimes$ mean that the corresponding entries are considered to be independent. In other words, we assume that each of the measurement noises is independent of all the others. Of course, it may not be the case in practice. But, as all the observations for a given subject come from a single curve, this assumption is reasonable in our context. Moreover, this assumption leads us to a more computationally tractable algorithm.

II. Parameters Estimation

As said just above, we want to estimate $\theta = (\overline{\boldsymbol{z}_{\text{pop}}}, \Sigma, \sigma) \in \mathbb{R}^{p_{\text{pop}}} \times \mathscr{S}^+_{p_{\text{ind}}}(\mathbb{R}) \times \mathbb{R}^+$. As we want our model to be consistent with low sample size high-dimensional data analysis, we consider a Bayesian framework, *i.e.* we assume the following priors

$$(\Sigma, \sigma) \sim \mathcal{W}^{-1}(V, m_\Sigma) \otimes \mathcal{W}^{-1}(v, m_\sigma) , \quad \text{where} \quad V \in \mathscr{S}^+_{p_{\text{ind}}}(\mathbb{R}), \quad v, m_\Sigma, m_\sigma \in \mathbb{R}$$

and $\mathcal{W}^{-1}(V, m_\Sigma)$ denotes the inverse Wishart distribution with scale matrix V and degrees of freedom m_Σ. Regularization has indeed proven its fruitfulness in this context (Giraud, 2014). In order for the inverse Wishart to be non-degenerate, the degrees m_Σ and m_σ must satisfy $m_\Sigma > 2p_{\text{ind}}$ and $m_\sigma > 2$. In practice, we yet use degenerate priors but with well-defined posteriors. In the spirit of the one-dimension inverse Wishart distribution, we define the density function distribution of higher dimension as

$$f_{\mathcal{W}^{-1}(V,m_\Sigma)}(\Sigma) = \frac{1}{\Gamma_{p_{\text{ind}}}\left(\frac{m_\Sigma}{2}\right)} \left(\frac{\sqrt{|V|}}{2^{\frac{p_{\text{ind}}}{2}} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} \operatorname{tr}\left(V\Sigma^{-1} \right) \right) \right)^{m_\Sigma} ,$$

where $\Gamma_{p_{\text{ind}}}$ is the multivariate gamma function and, for all matrices A, $|A|$ denotes the determinant of the matrix A.

The estimates are obtained by maximizing the posterior density on θ conditionally on the observations $\boldsymbol{y} = (y_{i,j})_{(i,j)\in[\![1,n]\!]\times[\![1,k_i]\!]}$.

In the following paragraphs, we first show that the model is well-posed *i.e.* that for any finite sample the *maximum* we are looking for exists. We then prove a consistency theorem which ensures that the set of parameters which well-explain the observations is non-empty and that the MAP estimator converges to this set. Last, we explain how to use the MCMC-SAEM algorithm to produce MAP estimates.

II – 1. Existence of the Maximum a Posteriori Estimator

The inverse Wishart priors on the variances not only regularize the log-likelihood of the model, they also ensure the existence of the MAP estimator.

Theorem 3.1 (*Existence of the MAP estimator*)

Given a piecewise-geodesic model and the choice of probability distributions for the parameters and latent variables of the model, for any data set $(t_{i,j}, y_{i,j})_{(i,j)\in[\![1,n]\!]\times[\![1,k_i]\!]}$, there exists

$$\widehat{\theta}_{MAP} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\ q(\theta|\boldsymbol{y}) .$$

The demonstration of the theorem uses the following lemma.

Lemma 3.2.1. Given a piecewise-geodesic model and the choice of a probability distribution for the parameters and latent variables of the model, the posterior $\theta \mapsto q(\theta|\boldsymbol{y})$ is continuous on the parameter space Θ.

Proof: Let $\mathcal{Z} = \mathcal{Z}_{\text{pop}} \times \prod_{i=1}^{n} \mathcal{Z}_i$ denote the space of latent variables. Using Bayes rule, for all $\theta \in \Theta$,

$$q(\,\theta\,|\,\boldsymbol{y}\,) \;=\; \frac{1}{q(\boldsymbol{y})} \left(\int_{\mathcal{Z}} q(\,\boldsymbol{y}\,|\,\boldsymbol{z},\theta\,)\; q(\,\boldsymbol{z}\,|\,\theta\,)\,\mathrm{d}\boldsymbol{z} \right) q_{prior}(\theta)\,.$$

The density functions $\theta \mapsto q_{prior}(\theta)$ and $\theta \mapsto q(\boldsymbol{y}|\boldsymbol{z},\theta)q(\boldsymbol{z}|\theta)$ are continuous on Θ for all $\boldsymbol{z} \in \mathcal{Z}$. Moreover, for all $\theta \in \Theta$ and all $\boldsymbol{z} \in \mathcal{Z}$,

$$q(\,\boldsymbol{y}\,|\,z,\theta\,) \;=\; \frac{1}{(\sigma\sqrt{2\pi})^k} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \big(y_{i,j} - \gamma_i(t_{i,j})\big)^2 \right)$$

and so, for all $\theta \in \Theta$ and $\boldsymbol{z} \in \mathcal{Z}$,

$$q(\,\boldsymbol{y}\,|\,\boldsymbol{z},\theta\,)\, q(\,\boldsymbol{z}\,|\,\theta\,) \;\leqslant\; \frac{1}{(\sigma\sqrt{2\pi})^k}\, q(\,\boldsymbol{z}\,|\,\theta\,)\,,$$

which is positive and integrable as a probability distribution. As a consequence, $\boldsymbol{z} \mapsto q(\boldsymbol{y}|\boldsymbol{z},\theta)\, q(\boldsymbol{z}|\theta)$ is integrable – and positive – on $\mathcal{Z}$ for all $\theta \in \Theta$ and $\theta \mapsto q(\boldsymbol{y}|\theta)$ is continuous. □

Proof of Theorem 3.1 – Existence of the MAP: We use the Alexandrov one-point compactification $\overline{\Theta} = \Theta \cup \{\infty\}$ of the parameters space Θ, where a sequence $(\theta_n)_{n\in\mathbb{N}}$ converges toward the point ∞ if and only if it eventually steps out of every compact subset of Θ. Thus, given the result of Lemma 3.2.1, it suffices to prove that $\lim_{\theta\to\infty} \log\, q(\theta|\boldsymbol{y}) = -\infty$. We keep the notation of the previous proof and proceed similarly. In particular, for all $\theta \in \Theta$,

$$\log\, q(\,\theta\,|\,\boldsymbol{y}\,) \;\leqslant\; -\log\, q(\boldsymbol{y}) - k\log(\sqrt{2\pi}) - k\log(\sigma) + \log\, q_{prior}(\theta)\,.$$

By computing the prior distribution q_{prior}, we remark that there exists λ which does not depend on the parameter θ such as

$$\begin{aligned}\log\, q(\,\theta\,|\,\boldsymbol{y}\,) \;\leqslant\; \lambda(\boldsymbol{y}) - (k + m_\sigma)\log(\sigma) \\ - \frac{m_\sigma}{2}\left(\frac{v}{\sigma}\right)^2 - \frac{m_\Sigma}{2}\left[\log(|\Sigma|) + \frac{m_\Sigma}{2}\,\mathrm{tr}\left(V\Sigma^{-1}\right)\right].\end{aligned}$$

Let $\mu(V)$ denote the smallest eigenvalue of V, $\rho(\Sigma^{-1})$ the largest eigenvalue of Σ^{-1}, which is also its operator norm, and $\langle \Sigma \,|\, V \rangle_F$ the Frobenius inner product of Σ with V. We know that $\langle \Sigma \,|\, V \rangle_F \geqslant \mu(V)\,\rho(\Sigma^{-1})$ and that $\log(|\Sigma^{-1}|) \leqslant p_{\text{ind}} \log\left(\|\Sigma^{-1}\|\right)$ so that

$$\log(|\Sigma^{-1}|) - \operatorname{tr}\left(V\Sigma^{-1}\right) \;\leqslant\; p_{\text{ind}} \log\Big(\|\Sigma^{-1}\|\Big) - \mu(V)\,\|\Sigma^{-1}\|\,.$$

As a consequence, it comes that

$$\lim_{\|\Sigma\|+\|\Sigma^{-1}\|\to+\infty} \left\{ \frac{m_\Sigma}{2} \left[\log(|\Sigma^{-1}|) - \operatorname{tr}\left(V\Sigma^{-1}\right)\right] \right\} = -\infty\,.$$

Likewise,

$$\lim_{\sigma+\sigma^{-1}\to+\infty} \left\{ -(k+m_\sigma)\log(\sigma) - \frac{m_\sigma}{2}\left(\frac{v}{\sigma}\right)^2 \right\} = -\infty$$

hence the result. □

We have detailed the previous proof in order to emphasize the necessity of prior distribution on the variances Σ and σ to ensure the existence of the *maximum a posteriori*.

II – 2. Consistency of the Maximum a Posteriori Estimator

We are now interested in the consistency of the MAP estimator without making strong assumptions on the distribution of the observations $\boldsymbol{y}$. In particular, we do not assume that the observations are generated by the model.

We denote $P(\mathrm{d}\boldsymbol{y})$ the distribution governing the observations and Θ_* the set of admissible parameters inducing a model distribution close to $P(\mathrm{d}\boldsymbol{y})$:

$$\Theta_* \;=\; \left\{ \theta_* \in \Theta \;\middle|\; \mathbb{E}_{P(\mathrm{d}\boldsymbol{y})}\left[\log q(\boldsymbol{y}|\theta_*)\right] \;=\; \sup_{\theta\in\Theta^\omega} \mathbb{E}_{P(\mathrm{d}\boldsymbol{y})}\left[\log q(\boldsymbol{y}|\theta)\right] \right\}.$$

The MAP estimator is said consistent if it converges to the set Θ_* (on every compact of Θ possibly). Classical results of consistency assume that the space Θ_* is non-empty (see the Wald's consistency theorem (van der Vaart, 2000)). However, such an hypothesis is not entirely satisfactory: we have no guarantee that Θ_* is actually non-empty. We propose below a reasonable framework in which the convergence of the MAP estimator toward the corresponding non-empty set Θ_* is guaranteed.

a. Two Kinds of Latent Variables

To this end and for any $\omega \in \mathbb{R}^+$, we define the space Θ^ω of admissible parameters such that on average, the fixed effects are bounded by ω:

$$\Theta^\omega = \{\theta = (\overline{\boldsymbol{z}_{\text{pop}}}, \Sigma, \sigma) \in \Theta \mid \|\overline{\boldsymbol{z}_{\text{pop}}}\|_2 \leqslant \omega\}\,, \quad \text{where} \quad \Theta = \mathbb{R}^{p_{\text{pop}}} \times \mathscr{S}^+_{p_{\text{ind}}}(\mathbb{R}) \times \mathbb{R}^+\,.$$

As the assumption only concern the average behavior of the population variable $\boldsymbol{z}_{\text{pop}}$, it is not restrictive. Moreover, fixed effects are most of the time bounded (but potentially with high bounds) in applications. In this new framework, for all $\omega \in \mathbb{R}^+$,

$$\Theta_*^\omega = \left\{ \theta \in \Theta^\omega \;\middle|\; \mathbb{E}_{P(\mathrm{d}\boldsymbol{y})} \left[\log q(\boldsymbol{y}|\theta) \right] = \mathbb{E}^*(\omega) \right\} ,$$

where

$$\mathbb{E}^*(\omega) = \sup_{\theta \in \Theta^\omega} \mathbb{E}_{P(\mathrm{d}\boldsymbol{y})} \left[\log q(\boldsymbol{y}|\theta) \right] .$$

To state the consistency of the MAP estimator, we first have to give some notations. For all $i \in [\![1, n]\!]$, we assume the existence of two subsets of $\mathcal{Z}_i$, $\mathcal{Z}_i^{\text{reg}}$ and $\mathcal{Z}_i^{\text{crit}}$, such that $\mathcal{Z}_i = \mathcal{Z}_i^{\text{reg}} \times \mathcal{Z}_i^{\text{crit}}$. In other words, we assume that each component of each individual latent variable $\boldsymbol{z}_i$ may be of two sorts: *regular* or *critical*. We will respectively denote $\boldsymbol{z}_i^{\text{reg}}$ and $\boldsymbol{z}_i^{\text{crit}}$ this sub-variables leading to write, up to permutations, $\boldsymbol{z}_i = (\boldsymbol{z}_i^{\text{reg}}, \boldsymbol{z}_i^{\text{crit}})$. Likewise, we assume that the components of the population latent variables can be regular or critical, *i.e.* that there exists $\mathcal{Z}_{\text{pop}}^{\text{reg}}, \mathcal{Z}_{\text{pop}}^{\text{crit}} \subset \mathcal{Z}_{\text{pop}}$ such that $\boldsymbol{z}_{\text{pop}} = (\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_{\text{pop}}^{\text{crit}}) \in \mathcal{Z}_{\text{pop}}^{\text{reg}} \times \mathcal{Z}_{\text{pop}}^{\text{crit}}$. To stay consistent with the previous notations, we denote $p_{\text{ind}}^{\text{reg}}$, $p_{\text{ind}}^{\text{crit}}$, $p_{\text{pop}}^{\text{reg}}$ and $p_{\text{pop}}^{\text{crit}}$ the dimension of the ambient space of the matching sets: $\mathcal{Z}_i^{\text{reg}} \subset \mathbb{R}^{p_{\text{ind}}^{\text{reg}}}$ and so on.

b. Consistency of the Maximum a Posteriori Estimator

In the following, we want to study the effect of the variables $(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)$ on the trajectories. To this end, we introduce for all i the notation $\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i) = (\gamma_i(t_{i,j}))_{j \in [\![1,k_i]\!]} \in \mathbb{R}^{k_i}$ and more generally the functions $\vec{\boldsymbol{\gamma}}_i \colon \mathcal{Z}_{\text{pop}} \times \mathcal{Z}_i \to \mathbb{R}^{k_i}$. Let $\ell \in [\![1, n]\!]$, consider a ℓ-tuple of individuals and denote by $k^\ell = \sum_{i=1}^{\ell} k_i$ the total number of measures for this ℓ-tuple. Let $\boldsymbol{y}^\ell = (y_i)_{i \in [\![1,\ell]\!]} \in \mathbb{R}^{k^\ell}$ and $\boldsymbol{z}^\ell = (\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)_{i \in [\![1,\ell]\!]} \in \mathbb{R}^{p_{\text{pop}} + \ell\, p_{\text{ind}}}$ be the vectors made up of the ℓ corresponding vectors. As in the one-by-one case, we define by $\vec{\boldsymbol{\gamma}}^\ell \colon \mathcal{Z}_{\text{pop}} \times \mathcal{Z}_i^\ell \to \mathbb{R}^{k^\ell}$ the function which maps the vector $\boldsymbol{z}^\ell$ to the one $(\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i))_{i \in [\![1,\ell]\!]}$.

For all vectors of the form $(\boldsymbol{a}, \boldsymbol{b}) \in \mathbb{R}^{p_a} \times \mathbb{R}^{p_b}$, where p_a and p_b are any integer number and for all indices $v \in [\![1, p_a + p_b]\!]$, $(\boldsymbol{a}, \boldsymbol{b})_v$ and $(\boldsymbol{a}, \boldsymbol{b})_{-v}$ refer respectively to

$$(\boldsymbol{a}, \boldsymbol{b})_v = \left((a_1, \ldots, a_{p_a}) , (b_1, \ldots, b_{p_b}) \right)_v = \begin{cases} a_v & \text{if } \; v \leqslant p_a \\ b_{v-p_a} & \text{else} \end{cases}$$

and

$$(\boldsymbol{a}, \boldsymbol{b})_{-v} = \begin{cases} \left((a_1, \ldots, a_{v-1}, a_{v+1}, \ldots, a_{p_a}), \boldsymbol{b} \right) & \text{if } \; v \leqslant p_a \\ \left(\boldsymbol{a} , (b_1, \ldots, b_{v-p_a-1}, b_{v-p_a+1}, \ldots, b_{p_b}) \right) & \text{else} \end{cases} .$$

Last, for all $k \in \mathbb{N}$, $\mathscr{L}_k$ refers to the Lebesgue measure on $\mathbb{R}^k$.

Theorem 3.2 (*Consistency of the MAP estimator*)

Assume that there exists an integer $\ell \in [\![1, n]\!]$ such that:

(H 1) The number of observations is bigger than the one of latent variables: $p^\ell < k^\ell$, where $k^\ell = \sum_{i=1}^{\ell} k_i$ and $p^\ell = p_{\text{pop}} + \ell\, p_{\text{ind}}$;

(H 2) The times of acquisitions $\boldsymbol{t}_i = (t_{i,j})_{j \in [\![1,k_i]\!]}$ are independent and identically distributed;

(H 3) The density $P(\mathrm{d}\boldsymbol{y}^\ell)$ is continuous with polynomial tail decay of degree bigger than the dimension of the truncated space of latent variables, *i.e.* bigger than $p^\ell + 1$, apart from a compact subset K of $\mathbb{R}^{k^\ell}$;

(H 4) The individual trajectories grow super-linearly with respect to the regular variables: for all individuals $i \in [\![1, n]\!]$ and for all $v \in [\![1, p_{\text{pop}}^{\text{reg}} + p_{\text{ind}}^{\text{reg}}]\!]$, there exists two functions $a_{i,v}, b_{i,v} \colon \mathbb{R}^{p_{\text{pop}}^{\text{reg}} + p_{\text{ind}}^{\text{reg}} - 1} \to \mathbb{R}$ which depend only of $(\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-v}$ and such that

$$\forall (\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i) \in \mathcal{Z}_{\text{pop}} \times \mathcal{Z}_i, \quad a_{i,v}\left((\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-v}\right) \geqslant 0\,,$$

where

$$a_{i,v}\left((\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-v}\right) = 0 \quad \text{iff} \quad (\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-v} = 0\,,$$

and

$$\begin{aligned}\left\|\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)\right\|_\infty \;\geqslant\; a_{i,v}\left((\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-v}\right) \left|(\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_v\right| \\ + \; b_{i,v}\left((\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-v}\right)\,;\end{aligned}$$

(H 5) Critical variables induce critical trajectories: for all individuals $i \in [\![1, n]\!]$ and for all $v \in [\![1, p_{\text{pop}}^{\text{crit}} + p_{\text{ind}}^{\text{crit}}]\!]$, there exists a critical trajectory $\gamma_{i,v}^{\text{crit}}$ which depends only of $\left(\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}}\right)$ and $\left(\boldsymbol{z}_{\text{pop}}^{\text{crit}}, \boldsymbol{z}_i^{\text{crit}}\right)_{-v}$ such that

$$\lim_{\left|(\boldsymbol{z}_{\text{pop}}^{\text{crit}}, \boldsymbol{z}_i^{\text{crit}})_v\right| \to +\infty} \vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i) \;=\; \gamma_{i,v}^{\text{crit}} \quad \text{and} \quad \mathscr{L}_{k_i}(\{y_i = \gamma_{i,v}^{\text{crit}}\}) = 0\,.$$

Let $(\hat{\theta}_n)_{n \in \mathbb{N}}$ denote any MAP estimator. Then $\Theta_*^\omega \neq \varnothing$ and for any $\varepsilon \in \mathbb{R}_+^*$,

$$\lim_{n \to \infty} \mathbb{P}\left[\delta(\hat{\theta}_n, \Theta_*^\omega) \geqslant \varepsilon\right] \;=\; 0\,,$$

where δ in any metric compatible with the topology on Θ^ω.

The proof is postponed in Section III.

If the times of observations $\boldsymbol{t}_i$ are identically distributed, the individual numbers of measurements k_i are in particular all equal. Thus, under **(H 2)**, Assumption **(H 1)** writes in a more concise manner as $p^\ell < \ell\, k_1$. However, as **(H 2)** is not required for all intermediate results (see the proof in Section III), we keep the more general statement for **(H 1)**. The condition **(H 2)** is for instance met if we assume that the times $\boldsymbol{t}_i$ are regularly spaced, that is to say that for all individuals $i \in [\![1, n]\!]$ and all measurements $j \in [\![1, k_1]\!]$, $t_{i,j}$ follows the uniform distribution $\mathcal{U}([T_{j-1}, T_j])$, where T is a maximum of the set $\{t_{i,j} | i \in [\![1, n]\!],\, j \in [\![1, k_1]\!]\}$ and $(T_0 = 0 < T_1 < \ldots < T_{k_1} = T)$ is a subdivision of $[0, T]$.

The condition $p^\ell < k^\ell$ means that without enough observations for at least some individuals, we cannot build a consistent model. Such an assumption is quite reasonable as we have no chance to catch the trajectories behavior with certitude with less observations than the constraints over them. The assumption on the distribution $P(\mathrm{d}\boldsymbol{y})$ is really weak and always fulfilled in practice. Moreover, as the theorem holds for all $\omega \in \mathbb{R}^+$, the boundary over the average of the population latent variable $\overline{\boldsymbol{z}_{\text{pop}}}$ is not really restrictive.

For compactness, we have stated the theorem by considering that a latent variable may be of only one kind: regular or critical. Actually, a single latent variable can be of two kinds: critical in the neighbourhood of $+\infty$ and regular around $-\infty$, and *vice-versa* (see the proof for details). This remark is all the more important in view of some applications and Chapter 5 but is treated by our proof.

II – 3. Estimation with the MCMC-SAEM Algorithm

As explain at the paragraph I.3.b, a stochastic version of the EM algorithm is adopted, namely the stochastic approximation expectation-maximization (SAEM) algorithm. As the conditional distribution $q(\boldsymbol{z}|\boldsymbol{y}, \theta)$ involves he renormalization constant which is our target function, the simulation step is replaced using a Monte-Carlo Markov Chain (MCMC) sampling algorithm, leading to consider the MCMC-SAEM algorithm (Allassonnière et al., 2010; Kuhn and Lavielle, 2005). It alternates between simulation, stochastic approximation and maximization steps until convergence. The simulation step is achieved using a symmetric random walk Hasting-Metropolis within Gibbs sampler (Robert and Casella, 1999).

Algorithm 1: Overview of the SAEM for the generic piecewise-geodesic model.

Input: $\theta^* = (\overline{z_{\text{pop}}}^*, \Sigma^*, \sigma^*)$, (V, m_Σ), (v, m_σ), `maxIter`, `Nburnin`.

Output: $\theta = (\overline{z_{\text{pop}}}, \Sigma, \sigma)$.

Initialization: $\theta \leftarrow \theta^*$; $S \leftarrow 0$; $(\varepsilon_{\texttt{iter}})_{\texttt{iter}>0}$; $z_{pop} \leftarrow \overline{z_{\text{pop}}}$; $(\boldsymbol{z}_i)_i \leftarrow 0$;

for *iter* $= 1$ **to** *maxIter* **do**

  Simulation: $(\boldsymbol{z}_{\text{pop}}, (\boldsymbol{z}_i)_i) \leftarrow \texttt{sampler}(\boldsymbol{z}_{\text{pop}}, (\boldsymbol{z}_i)_i)$;

  Stochastic Approximation: $\boldsymbol{S}_1 \leftarrow \boldsymbol{S}_1 + \varepsilon_{\texttt{iter}} (\boldsymbol{z}_{\text{pop}} - \boldsymbol{S}_1)$;

$$S_2 \leftarrow S_2 + \varepsilon_{\texttt{iter}} \left(\frac{1}{n} \sum_i {}^t\boldsymbol{z}_i \boldsymbol{z}_i - S_2 \right) ;$$

$$S_3 \leftarrow S_3 + \varepsilon_{\texttt{iter}} \left(\frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \big(y_{i,j} - \gamma_i(t_{i,j}) \big)^2 - S_3 \right) ;$$

  Maximization: $\overline{z_{\text{pop}}} \leftarrow \boldsymbol{S}_1$; $\Sigma \leftarrow \dfrac{nS_2 + m_\Sigma V}{n + m_\Sigma}$; $\sigma \leftarrow \sqrt{\dfrac{kS_3 + m_\sigma v^2}{k + m_\sigma}}$;

end

The complete log-likelihood of our model writes

$$\begin{aligned}
\log q(\boldsymbol{y}, \boldsymbol{z}, \theta) \;=\; & -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \big(y_{i,j} - \gamma_i(t_{i,j}) \big)^2 \;-\; k \log(\sigma) \\
& - \frac{1}{2} \sum_{i=1}^{n} \big({}^t\boldsymbol{z}_i \, \Sigma^{-1} \, \boldsymbol{z}_i \big) \;-\; \frac{n}{2} \log(|\Sigma|) \\
& - \frac{1}{2} {}^t\big(\boldsymbol{z}_{\text{pop}} - \overline{\boldsymbol{z}_{\text{pop}}} \big) D^{-1} \big(\boldsymbol{z}_{\text{pop}} - \overline{\boldsymbol{z}_{\text{pop}}} \big) \;-\; \frac{1}{2} \log(|D|) \;-\; \frac{1}{2} \operatorname{tr}\left(V \Sigma^{-1} \right) \\
& + \frac{m_\Sigma}{2} \big(\log(|V|) - \log(|\Sigma|) \big) \;+\; m_\sigma \log\left(\frac{v}{\sigma} \right) \;-\; \frac{m_\sigma}{2} \left(\frac{v}{\sigma} \right)^2 \;+\; \mathit{csts} \, .
\end{aligned}$$

It is clear to see that this model belongs to the curved exponential family: up to a multiple constant, the sufficient statistics are defined as

$$\boldsymbol{S}_1(\boldsymbol{y}, \boldsymbol{z}) \;=\; \boldsymbol{z}_{\text{pop}} \in \mathbb{R}^{p_{\text{pop}}} \;, \quad S_2(\boldsymbol{y}, \boldsymbol{z}) \;=\; \frac{1}{n} \sum_{i=1}^{n} {}^t\boldsymbol{z}_i \boldsymbol{z}_i \in \mathscr{M}_{p_{\text{ind}}}(\mathbb{R})$$

$$\text{and} \quad S_3(\boldsymbol{y}, \boldsymbol{z}) \;=\; \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \big(y_{i,j} - \gamma_i(t_{i,j}) \big)^2 \in \mathbb{R} \, .$$

By denoting `iter` the increment, $\boldsymbol{z}^{(\texttt{iter})}$ the current sample and $S_u^{(\texttt{iter})}$ the current approximation of the u$^{\text{th}}$ sufficient statistics, $u \in \{1, 2, 3\}$, the stochastic approximation step is defined as:

$$S_u^{(\texttt{iter})} = S_u^{(\texttt{iter})} + \varepsilon_{\texttt{iter}} \left(S_u(\boldsymbol{y}, \boldsymbol{z}^{(\texttt{iter})}) - S_u^{(\texttt{iter})} \right) ,$$

where $(\varepsilon_{\texttt{iter}})$ is a sequence positive step size (see below).

The maximization step is straightforward given the sufficient statistics of our exponential model: we update the parameters by taking a barycentre between the corresponding sufficient statistic and the prior (when there exists). In other words,

$$\overline{\boldsymbol{z}_{\text{pop}}}^{(\texttt{iter}+1)} = \boldsymbol{S}_1(\boldsymbol{y}, \boldsymbol{z}^{(\texttt{iter})}), \quad \Sigma^{(\texttt{iter}+1)} = \frac{nS_2(\boldsymbol{y}, \boldsymbol{z}^{(\texttt{iter})}) + m_\Sigma V}{n + m_\Sigma}$$

$$\text{and} \quad \sigma^{2(\texttt{iter}+1)} = \frac{kS_3(\boldsymbol{y}, \boldsymbol{z}^{(\texttt{iter})}) + m_\sigma v^2}{k + m_\sigma}\,.$$

Finally, given an adapted sampler (the symmetric random walk Hastings-Metropolis within Gibbs Sampler for instance) and the sequence $(\varepsilon_{\texttt{iter}})_{\texttt{iter}}$ defined by

$$\forall\, \texttt{iter} \geqslant 1, \quad \varepsilon_{\texttt{iter}} = \mathbb{1}_{\texttt{iter} \leqslant \texttt{Nburnin}} + (\texttt{iter} - \texttt{Nburnin})^{-0.65}\, \mathbb{1}_{\texttt{iter} > \texttt{Nburnin}}\,,$$

our algorithm writes as Algorithm 1 . Some experimental results are presented in Section III of Chapter 5.

III. Proof of the Consistency Theorem for Bounded Population Variable

The proof of the theorem relies on several lemmas. Lemma 3.3.3 is the heart of the proof: we control here the behavior of the log-likelihood at the boundary points of the parameters space Θ_*^ω and prove that this set is non-empty. It is based on Lemma 3.3.2 which states the integrability of the supremum over the parameter space of the positive part of the log-likelihood. Lemma 3.3.1 is derived from Allassonnière et al. (2007). We transpose the proof of the cited article here (with few more details) as this lemma is critical in the proof of Lemma 3.3.2 and not such classical.

In the following, we freely (and without reminder) use the notations introduced in Section II.2. Moreover, **(H 1)**, **(H 2)**, **(H 3)**, **(H 4)** and **(H 5)** refer to the hypothesis of the consistency theorem (Theorem 3.2, page 75).

III – 1. Lemmas

We first recall that the minimal number of balls of radius $r \in \mathbb{R}_+^*$ required to cover a compact set $K \in \mathbb{R}^p$ is bounded from above by $\left(\frac{\mathcal{D}\text{iam}(K)}{r}\right)^p$.

Lemma 3.3.1 (Preliminary of measure theory)**.** Let $p < q$ be two integers. Then, for any differentiable map $f\colon \mathbb{R}^p \to \mathbb{R}^q$ and any compact subset K of $\mathbb{R}^p$, there exists a constant λ which depends only on p and q such that

$$\int_{\mathbb{R}^q \setminus f(K)} \log^+ \frac{1}{d(\boldsymbol{y}, f(K))}\, \mathrm{d}\boldsymbol{y} \;<\; \lambda \left(\sup_K \|\mathcal{D}f\| + 2\right)^q \mathcal{D}\text{iam}(K)^p\,,$$

where d is the euclidean distance on $\mathbb{R}^q$, $\mathcal{D}f$ the differential of f and $\mathcal{D}\mathrm{iam}(K)$ the diameter of the compact K. Especially,

$$\int_{\mathbb{R}^q \setminus f(K)} \log^+ \frac{1}{d(\boldsymbol{y}, f(K))}\, \mathrm{d}\boldsymbol{y} \;<\; +\infty\,.$$

Proof: For all $\rho, \rho_1, \rho_2 \in \mathbb{R}_+^*$, $\rho_1 < \rho_2$, let

$$M_{\rho_1,\rho_2} = \big\{\boldsymbol{y} \in \mathbb{R}^q \mid \rho_1 \leqslant d(\boldsymbol{y}, f(K)) \leqslant \rho_2\big\} \quad \text{and} \quad M_\rho = M_{0,\rho}\,.$$

For all $\rho \in \mathbb{R}_+^*$, due to the compactness of K, there exists a finite set $\Lambda_\rho \subset K$ such that $K \subset \bigcup_{\boldsymbol{x}\in\Lambda_\rho} \mathcal{B}(\boldsymbol{x}, \rho)$ and $|\Lambda_\rho| \leqslant \left(\frac{\mathcal{D}\mathrm{iam}(K)}{\rho}\right)^p$. Let $\tau = \sup_K \|\mathcal{D}f\|$. According to the mean value theorem, $M_{0,\rho} \subset \mathcal{B}\big(f(\boldsymbol{x}), (\tau+2)\rho\big)$ and

$$\begin{aligned}
\mathscr{L}_q(M_\rho) \;&\leqslant\; \sum_{\boldsymbol{x}\in\Lambda_\rho} \mathscr{L}_q\Big(\mathcal{B}\big(f(\boldsymbol{x}), (\tau+2)\rho\big)\Big) \\
&\leqslant\; \frac{\sqrt{\pi}^p\,(\tau+2)^p}{\Gamma\left(\frac{p}{2}+1\right)} \times \big(\mathcal{D}\mathrm{iam}(K)\big)^p \times \rho^{q-p}\,.
\end{aligned}$$

Let $s \in \,]0,1[$. Then, from the Abel transformation,

$$\begin{aligned}
\int_{\mathbb{R}^q \setminus f(K)} \log^+ \frac{1}{d(\boldsymbol{y}, f(K))}\, \mathrm{d}\boldsymbol{y} \;&=\; \sum_{n=0}^{+\infty} \int_{M_{s^{n+1},s^n}} \log^+ \frac{1}{d(\boldsymbol{y}, f(K))}\, \mathrm{d}\boldsymbol{y} \\
&\leqslant\; \sum_{n=0}^{+\infty} \log \frac{1}{s^{n+1}}\; [\mathscr{L}_q(M_{s^n}) - \mathscr{L}_q(M_{s^{n+1}})] \\
&\leqslant\; -\log(s) \sum_{n=0}^{+\infty} \mathscr{L}_q(M_{s^n})\,.
\end{aligned}$$

Hence the result as $s \in \,]0,1[$. □

Lemma 3.3.2. Assume **(H 1)**, **(H 3)**, **(H 4)** and

(H′ 5) Bounded regular variables implies bounded trajectories: For all individuals $i \in [\![1, n]\!]$, if there exists $b \in \mathbb{R}$ such that $\|(\boldsymbol{z}_{\mathrm{pop}}^{\mathrm{reg}}, \boldsymbol{z}_i^{\mathrm{reg}})\|_\infty < b$ then there exists $R \in \mathbb{R}_+^*$ such that $\|\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\mathrm{pop}}, \boldsymbol{z}_i)\|_\infty < R$.

Then, for any such ℓ,

$$\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\sup_{\theta\in\Theta}\left(\sum_{i=1}^{\ell} \log q(\boldsymbol{y}_i|\theta)\right)^+\right] \;<\; +\infty\,.$$

Proof: Let $i \in [\![1, \ell]\!]$, $\Gamma_i = \mathcal{I}m(\vec{\boldsymbol{\gamma}}_i)$ and $\Gamma^\ell = \prod_{i=1}^{\ell} \Gamma_i$. For all $\theta \in \Theta^\omega$,

$$\begin{aligned} q(\boldsymbol{y}_i|\theta) = & \frac{1}{(\sigma\sqrt{2\pi})^{k_i}} \int_{\overline{\mathcal{Z}}_i} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}_i - \vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)\|_2^2\right) q(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i|\theta)\ \mathrm{d}(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i) \\ \leqslant & \frac{1}{(\sigma\sqrt{2\pi})^{k_i}} \exp\left(-\frac{1}{2\sigma^2} d(\boldsymbol{y}_i, \Gamma_i)^2\right), \end{aligned}$$

where d denotes the Euclidean distance on $\mathbb{R}^{k_i}$. Thus for all $\theta \in \Theta^\omega$,

$$\sum_{i=1}^{\ell} \log q(\boldsymbol{y}_i|\theta) \ \leqslant\ -\frac{k^\ell}{2}\log(2\pi\sigma^2) \ -\ \frac{1}{2\sigma^2} d(\boldsymbol{y}^\ell, \Gamma^\ell)^2,$$

where d denotes now the Euclidean distance on $\mathbb{R}^{k^\ell}$, $k^\ell = \sum_{i=1}^{\ell} k_i$. As the right hand side is maximized for $\sigma^2 = \frac{1}{k^\ell} d(\boldsymbol{y}^\ell, \Gamma^\ell)^2$, there exists a constant $\lambda \in \mathbb{R}_+^*$ such that

$$\sup_{\theta\in\Theta} \left(\sum_{i=1}^{\ell} \log q(\boldsymbol{y}_i|\theta)\right)^+ \ \leqslant\ \lambda \ +\ k^\ell \log^+ \frac{1}{d(\boldsymbol{y}^\ell, \Gamma^\ell)}.$$

1. Assume there exists $i_0 \in [\![1, n]\!]$ such that that $\|(\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_{i_0}^{\text{reg}})\|_\infty \geqslant b$ for all $b \in \mathbb{R}$.

   For all $r_1, r_2 \in \mathbb{R}$ we define a compact subset $\Gamma^\ell_{r_1,r_2}$ of Γ^ℓ by setting

   $$\begin{aligned} \bar{\mathcal{A}}(r_1, r_2) \ =\ \Big\{ \boldsymbol{z}^\ell \in \mathbb{R}^{p^\ell} \ \mid\ & r_1 \leqslant \|(\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{i\in[\![1,\ell]\!]}\|_\infty, \\ & \|(\boldsymbol{z}_{\text{pop}}^{\text{crit}}, \boldsymbol{z}_i^{\text{crit}})_{i\in[\![1,\ell]\!]}\|_\infty \leqslant r_2 \Big\} \end{aligned}$$

   and $\quad \Gamma^\ell_{r_1,r_2} \ =\ \{\, \vec{\boldsymbol{\gamma}}^\ell(\boldsymbol{z}^\ell) \mid \boldsymbol{z}^\ell \in \bar{\mathcal{A}}(r_1, r_2) \,\}.$

   Especially, $\lim_{r_2\to\infty} \Gamma^\ell_{0,r_2} = \Gamma^\ell$. Moreover, $\vec{\boldsymbol{\gamma}}^\ell$ is differentiable *a.e.*, at least one-side differentiable everywhere and there exists $\tau \in \mathbb{R}$ such that $\sup_{\mathbb{R}^{p\ell}} \|\mathcal{D}_{\boldsymbol{z}^\ell} \vec{\boldsymbol{\gamma}}^\ell\| < \tau$. So, according to Lemma 3.3.1, for all $r_1, r_2 \in \mathbb{R}$, there exists $\mu \in \mathbb{R}$ which depends only on p^ℓ and k^ℓ such that

   $$\mathbb{E}\left[\log^+ \frac{1}{d(\boldsymbol{y}^\ell, \Gamma^\ell_{r_1,r_2})}\right] \ <\ \mu\,(\tau+2)^{k^\ell}\ {r_2}^{p^\ell}.$$

   As in the proof of Lemma 3.3.1, we set $\overline{\Gamma^\ell_{r_1,r_2}} = \{\boldsymbol{y}^\ell \in \mathbb{R}^{k^\ell} \mid d(\boldsymbol{y}^\ell, \Gamma^\ell_{r_1,r_2}) \leqslant 1\}$ and we have for all $r_1, r_2 \in \mathbb{R}$,

   $$\begin{aligned} \int_{\mathbb{R}^{k^\ell}} \log^+ \frac{1}{d(\boldsymbol{y}^\ell, \Gamma^\ell_{r_1,r_2})} P(\mathrm{d}\boldsymbol{y}^\ell) \ &=\ \int_{\overline{\Gamma^\ell_{r_1,r_2}}} \log^+ \frac{1}{d(\boldsymbol{y}^\ell, \Gamma^\ell_{r_1,r_2})} P(\mathrm{d}\boldsymbol{y}^\ell) \\ &\leqslant\ \bar{\mu}\,{r_2}^{p^\ell} \sup_{\overline{\Gamma^\ell_{r_1,r_2}}} P(\boldsymbol{y}^\ell), \end{aligned}$$

where $\bar{\mu} = \mu\,(\tau+2)^{k^\ell} \in \mathbb{R}$. Let $R_1, R_2 \in \mathbb{N}$ such that $K \subset \bar{\mathcal{B}}(0,R_1)$ and $R_1 < R_2$. By definition of the distance to a subset, it comes that

$$\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\log^+ \frac{1}{d(\boldsymbol{y}^\ell, \Gamma^\ell_{0,R_2})}\right] \leqslant \bar{\mu}\, {R_1}^{p^\ell} \sup_{\overline{\Gamma^\ell_{0,R_1}}} P(\boldsymbol{y}^\ell) + \bar{\mu} \sum_{r=R_1}^{R_2-1} (r+1)^{p^\ell} \sup_{\overline{\Gamma^\ell_{r,r+1}}} P(\boldsymbol{y}^\ell)\,.$$

The first term is finite as $P(\mathrm{d}\boldsymbol{y})$ is continuous. Besides, if $\boldsymbol{y}^\ell \in \overline{\Gamma^\ell_{r,r+1}}$, there exists $\boldsymbol{z}^\ell \in \bar{\mathcal{A}}(r,r+1)$ such that $\|\vec{\boldsymbol{\gamma}}^\ell(\boldsymbol{z}^\ell) - \boldsymbol{y}^\ell\|_\infty \leqslant 1$. Let $i \in [\![1,n]\!]$ and $v \in [\![1, p^{\mathrm{reg}}_{\mathrm{pop}} + p^{\mathrm{reg}}_{\mathrm{ind}}]\!]$ so that $\|(\boldsymbol{z}^{\mathrm{reg}}_{\mathrm{pop}}, \boldsymbol{z}^{\mathrm{reg}}_i)_{i\in[\![1,n]\!]}\|_\infty = |(\boldsymbol{z}^{\mathrm{reg}}_{\mathrm{pop}}, \boldsymbol{z}^{\mathrm{reg}}_i)_v|$. Such a couple exists due to the existence of i_0. Moreover, there exists $a_{i,v}\big((\boldsymbol{z}^{\mathrm{reg}}_{\mathrm{pop}}, \boldsymbol{z}^{\mathrm{reg}}_i)_{-v}\big)$ and $b_{i,v}\big((\boldsymbol{z}^{\mathrm{reg}}_{\mathrm{pop}}, \boldsymbol{z}^{\mathrm{reg}}_i)_{-v}\big)$ as in **(H 4)** and by definition of $\boldsymbol{z}^\ell$ and the infinite norm,

$$\begin{aligned}\|\boldsymbol{y}^\ell\|_\infty \;\geqslant\; \|\vec{\boldsymbol{\gamma}}^\ell(\boldsymbol{z}^\ell)\|_\infty - 1 &\geqslant \|\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\mathrm{pop}}, \boldsymbol{z}_i)\|_\infty - 1\\ &\geqslant a_{i,v}\,|(\boldsymbol{z}^{\mathrm{reg}}_{\mathrm{pop}}, \boldsymbol{z}^{\mathrm{reg}}_i)_v| + b_{i,v} - 1\\ &\geqslant a_{i,v} \times r + b_{i,v} - 1\,.\end{aligned}$$

Consequently,

$$\sup_{\overline{\Gamma^\ell_{r,r+1}}} P(\boldsymbol{y}^\ell) \;\leqslant\; \sup\Big\{ P(\boldsymbol{y}^\ell) \;\mid\; \|\boldsymbol{y}^\ell\|_\infty \geqslant a_{i,v} \times r + b_{i,v} - 1 \Big\}$$

and the series $\sum (r+1)^{p^\ell} \sup_{\overline{\Gamma^\ell_{r,r+1}}} P(\boldsymbol{y}^\ell)$ converge since $P(\mathrm{d}\boldsymbol{y})$ has a polynomial decay tail of degree bigger than $p^\ell+1$ apart from K by assumption **(H 3)**.

2. Assume that there exists $b \in \mathbb{R}$ such that $\|(\boldsymbol{z}^{\mathrm{reg}}_{\mathrm{pop}}, \boldsymbol{z}^{\mathrm{reg}}_i)\|_\infty \leqslant b$ for all $i \in [\![1,n]\!]$. Then, by assumption **(H′ 5)**, there exists $R \in \mathbb{R}^*_+$ such that for all i, $\|\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\mathrm{pop}}, \boldsymbol{z}_i)\|_\infty < R$. In particular, $\|\vec{\boldsymbol{\gamma}}^\ell(\boldsymbol{z}^\ell)\|_\infty < R$ and $\Gamma^\ell \subset \bar{\mathcal{B}}(0,R)$. Thus,

$$\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\log^+ \frac{1}{d(\boldsymbol{y}^\ell, \Gamma^\ell)}\right] \;\leqslant\; \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\log^+ \frac{1}{d(\boldsymbol{y}^\ell, \bar{\mathcal{B}}(0,R))}\right].$$

Yet, by still denoting $\overline{\bar{\mathcal{B}}(0,R)} = \{\boldsymbol{y}^\ell \in \mathbb{R}^{k^\ell} \mid d(\boldsymbol{y}^\ell, \bar{\mathcal{B}}(0,R)) \leqslant 1\}$ and applying Lemma 3.3.1 to the compact $K = \bar{\mathcal{B}}(0,R)$ and $f = Id$, there exists $\mu \in \mathbb{R}$ such that

$$\begin{aligned}\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\log^+ \frac{1}{d(\boldsymbol{y}^\ell, \bar{\mathcal{B}}(0,R))}\right] &= \int_{\overline{\bar{\mathcal{B}}(0,R)}} \log^+ \frac{1}{d(\boldsymbol{y}^\ell, \bar{\mathcal{B}}(0,R))} P(\mathrm{d}\boldsymbol{y}^\ell)\\ &\leqslant \mu\, 3^{k^\ell}\, R^{p^\ell} \sup_{\overline{\bar{\mathcal{B}}(0,R)}} P(\boldsymbol{y}^\ell) \;< +\infty\,.\end{aligned}$$

Finally, in both cases, $\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\sup_{\theta\in\Theta}\left(\sum_{i=1}^{\ell}\log q(\boldsymbol{y}_i|\theta)\right)^+\right] < +\infty$. □

Lemma 3.3.3. Assume **(H 1)**, **(H 3)**, **(H 4)** and **(H 5)**. Let

$$\overline{\mathscr{S}^+_{p_{\mathrm{ind}}}(\mathbb{R})} = \mathscr{S}^+_{p_{\mathrm{ind}}}(\mathbb{R}) \cup \{\infty\}$$

be the one point Alexandrov compactification of $\mathscr{S}^+_{p_{\mathrm{ind}}}(\mathbb{R})$ and consider the compactification of the parameter space Θ^ω

$$\overline{\Theta^\omega} = \left\{\theta = (\overline{\boldsymbol{z}_{\mathrm{pop}}}, \Sigma, \sigma) \in \mathbb{R}^{p_{\mathrm{pop}}} \times \overline{\mathscr{S}^+_{p_{\mathrm{ind}}}(\mathbb{R})} \times \overline{\mathbb{R}^+} \mid \|\overline{\boldsymbol{z}_{\mathrm{pop}}}\| \leqslant \omega\right\},$$

where $\overline{R^+} = [0, +\infty[\cup \{+\infty\}$. Then, we have for all $\omega \in \mathbb{R}$,

(C 1) $P(\mathrm{d}\boldsymbol{y}^\ell)$ almost surely, for any sequence $\theta_\kappa = (\overline{\boldsymbol{z}_{pop}}_\kappa, \Sigma_\kappa, \sigma_\kappa)$ of elements from Θ^ω such that $\lim_{\kappa\to\infty}\theta_\kappa \in \overline{\Theta^\omega} \setminus \Theta^\omega$,

$$\lim_{\kappa\to\infty}\sum_{i=1}^{\ell}\log q\left(\boldsymbol{y}_i|\theta_\kappa\right) = -\infty\,;$$

(C 2) For any sequence $(\theta_\kappa) \in \Theta^{\omega\mathbb{N}}$ such that $\lim_{\kappa\to\infty}\theta_\kappa \in \overline{\Theta^\omega} \setminus \Theta^\omega$,

$$\lim_{\kappa\to\infty}\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\log q(y|\theta_\kappa)\right] = -\infty\,;$$

(C 3) The mapping $\theta \mapsto \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\log q(y|\theta)\right]$ is continuous on Θ^ω and $\Theta^\omega_* \neq \varnothing$.

Proof: We recall that a sequence $(\Sigma_\kappa)_{\kappa\in\mathbb{N}}$ of $\overline{\mathscr{S}^+_{p_{\mathrm{ind}}}(\mathbb{R})}$ converge toward the point ∞ if it eventually steps out of every compact subset of $\mathscr{S}^+_{p_{\mathrm{ind}}}(\mathbb{R})$. Let prove the three points in order.

1. As

$$\overline{\Theta^\omega} \setminus \Theta^\omega = \left\{ (\Sigma, \sigma) \in \overline{\mathscr{S}^+_{p_{\mathrm{ind}}}(\mathbb{R})} \times \overline{\mathbb{R}^+} \;\middle|\; \|\Sigma\| = +\infty \wedge \|\Sigma^{-1}\| = +\infty \wedge \sigma \in \{0, +\infty\} \right\}$$

we proceed by disjunction. Let

$$\forall\kappa \in \mathbb{N}, \quad \theta_\kappa = (\overline{\boldsymbol{z}_{pop}}_\kappa, \Sigma_\kappa, \sigma_\kappa) \in \Theta^\omega\,.$$

(i) Assume that, up to extraction of a subsequence, $\|\Sigma_\kappa\| \to \infty$ or $\|\Sigma_\kappa^{-1}\| \to \infty$.

Let $M = \|\boldsymbol{y}^{\ell}\|_{\infty}$. For all individuals $i \in [\![1, n]\!]$ and all $\kappa \in \mathbb{N}$, the marginal density of $\boldsymbol{y}_i$ given θ_κ is given by :

$$q(\boldsymbol{y}_i|\theta_\kappa) \;=\; \frac{1}{(\sigma_\kappa\sqrt{2\pi})^{k_i}} \int_{\mathcal{Z}_{\text{pop}}\times\mathcal{Z}_i} \exp\left(-\frac{1}{2\sigma_\kappa^2}\|\boldsymbol{y}_i - \vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)\|_2^2\right) q(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i|\theta_\kappa)\, \mathrm{d}(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)\,.$$

Let $x \geqslant 1$, $\mathcal{Z}_{i,-1}^{\text{reg}} = \left\{(\boldsymbol{z}_{i,2}^{\text{reg}}, \ldots, \boldsymbol{z}_{i,p_{\text{ind}}^{\text{reg}}}^{\text{reg}}) \mid \boldsymbol{z}_i^{\text{reg}} \in \mathcal{Z}_i^{\text{reg}}\right\}$ and likewise $\mathcal{Z}_{\text{pop},-1}^{\text{reg}}$. Let $\bar{\mathcal{B}}_{i,1}^x$ be the closed ball defined by

$$\bar{\mathcal{B}}_{i,1}^x \;=\; \bar{\mathcal{B}}_{i,1}^x\Big((\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-1}\Big) \;=\; \bar{\mathcal{B}}\left(0\,,\, \frac{xM - b_{i,1}\big((\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-1}\big)}{a_{i,1}\big((\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-1}\big)}\right),$$

where $a_{i,1}\big((\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-1}\big)$ and $b_{i,1}\big((\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{-1}\big)$ are defined as in **(H 4)**. Thus, by slicing the integral in half and bounding the exponential on $\bar{\mathcal{B}}_{i,1}^x$ by 1,

$$\begin{aligned} q(\boldsymbol{y}_i|\theta_\kappa) \;\leqslant\; & \frac{1}{(\sigma_\kappa\sqrt{2\pi})^{k_i}} \int_{\bar{\mathcal{B}}_{i,1}^x\times\mathcal{Z}_{i,-1}} q(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i|\theta_\kappa)\, \mathrm{d}(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i) \\ & + \frac{1}{(\sigma_\kappa\sqrt{2\pi})^{k_i}} \int_{\bar{\mathcal{B}}_{i,1}^{x\,\complement}\times\mathcal{Z}_{i,-1}} \exp\left(-\frac{1}{2\sigma_\kappa^2}\|\boldsymbol{y}_i - \vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)\|_2^2\right) \\ & \qquad q(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i|\theta_\kappa)\, \mathrm{d}(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)\,, \end{aligned}$$

where $\mathcal{Z}_{i,-1} = \mathcal{Z}_{\text{pop},-1}^{\text{reg}} \times \mathcal{Z}_{\text{pop}}^{\text{crit}} \times \mathcal{Z}_{i,-1}^{\text{reg}} \times \mathcal{Z}_i^{\text{crit}}$. Moreover, by conditioning,

$$\int_{\bar{\mathcal{B}}_{i,1}^x\times\mathcal{Z}_{i,-1}} q(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i|\theta_\kappa)\, \mathrm{d}(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i) \;=\; \int_{\bar{\mathcal{B}}_{i,1}^x} q(\boldsymbol{z}_{\text{pop},1}^{\text{reg}}, \boldsymbol{z}_{i,1}^{\text{reg}}|\theta_\kappa)\, \mathrm{d}(\boldsymbol{z}_{\text{pop},1}^{\text{reg}}, \boldsymbol{z}_{i,1}^{\text{reg}})\,.$$

By continuity of $(\boldsymbol{z}_{\text{pop},1}^{\text{reg}}, \boldsymbol{z}_{i,1}^{\text{reg}}) \mapsto q(\boldsymbol{z}_{\text{pop},1}^{\text{reg}}, \boldsymbol{z}_{i,1}^{\text{reg}}|\theta_\kappa)$ and compactness of $\bar{\mathcal{B}}_{i,1}^x$,

$$\int_{\bar{\mathcal{B}}_{i,1}^x\times\mathcal{Z}_{i,-1}} q(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i|\theta_\kappa)\, \mathrm{d}(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i) \;\leqslant\; \sup_{\bar{\mathcal{B}}_{i,1}^x} q(\boldsymbol{z}_{\text{pop},1}^{\text{reg}}, \boldsymbol{z}_{i,1}^{\text{reg}}|\theta_\kappa)\, \mathscr{L}_1(\bar{\mathcal{B}}_{i,1}^x)\,.$$

Since the marginal of a multivariate distribution is a multivariate distribution whose mean vector and covariance matrix are obtained by dropping the irrelevant variables, $\lim_{\|\Sigma_\kappa\|\to\infty} q(\boldsymbol{z}_{\text{pop},1}^{\text{reg}}, \boldsymbol{z}_{i,1}^{\text{reg}}|\theta_\kappa) = 0$ and the first integral goes to zero as $\|\Sigma_\kappa\| \to \infty$.

In the same way of the proof of Theorem 3.1, the marginal density $q(\boldsymbol{z}_{\text{pop},1}^{\text{reg}}, \boldsymbol{z}_{i,1}^{\text{reg}}|\theta_\kappa)$ is controlled by the operator norm of the covariance matrix Σ_κ^{-1} from which we have drop the irrelevant variables. Hence, as $\|\Sigma_\kappa^{-1}\| \to \infty$, the first integral converges toward zero as well.

The second integral is maximized at $\sigma_\kappa^2 = \frac{1}{k_i}\|\boldsymbol{y}_i - \vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)\|^2$. Thus, due to the Cauchy-Schwarz inequality, there exists a constant $c \in \mathbb{R}_+^*$ such that for all $(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i) \in \bar{\mathcal{B}}_{i,1}^{x\,\complement} \times \mathcal{Z}_{i,-1}$,

$$\begin{aligned}\|\boldsymbol{y}_i - \vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)\|_2^2 &\geqslant c\left(a_{i,1} \times \frac{xM - b_{i,1}}{a_{i,1}} + b_{i,1} - \|\boldsymbol{y}_i\|_\infty\right)^2 \\ &\geqslant c\left((x-1)M\right)^2\end{aligned}$$

and by bounding the marginal density $q(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i|\theta_\kappa)$ on $\bar{\mathcal{B}}_{i,1}^{x\,\complement} \times \mathcal{Z}_{i,-1}$ by 1, the second integral is bounded from above by

$$\left(\frac{k_i}{2\pi}\right)^{\frac{k_i}{2}} \mathrm{e}^{-\frac{k_i}{2}} \frac{1}{\left(\sqrt{c}\,(x-1)M\right)^{k_i}}\,.$$

Therefore,

$$\begin{aligned}\limsup_{\kappa\to\infty} \sum_{i=1}^{\ell} \log q(\boldsymbol{y}_i|\theta_\kappa) \;\leqslant\; &-\frac{k^\ell}{2}\Big[1 + \log(2\pi) + \log\left(\sqrt{c}\,(x-1)M\right)\Big] \\ &+ \frac{1}{2}\sum_{i=1}^{\ell} k_i \log k_i\,.\end{aligned}$$

Since x can be chosen arbitrarily large, we obtain the result for the case $\|\Sigma_\kappa\| \to +\infty$ as well as $\|\Sigma_\kappa^{-1}\| \to +\infty$.

(ii) Assume that, up to extraction of a subsequence, $\sigma_\kappa \to 0$ or $\sigma_\kappa \to \infty$. Let $M = \|\boldsymbol{y}^\ell\|_\infty$. With the same notations as in the proof of Lemma 3.3.2, for all $\kappa \in \mathbb{N}$,

$$\sum_{i=1}^{\ell} \log q(\boldsymbol{y}_i|\theta_\kappa) \;\leqslant\; -\frac{k^\ell}{2}\log(2\pi\sigma_\kappa^2) - \frac{1}{2\sigma^2} d(\boldsymbol{y}^\ell, \Gamma^\ell)^2\,,$$

where $\Gamma^\ell = \mathcal{I}m(\vec{\boldsymbol{\gamma}}^\ell)$ and d denotes the Euclidean distance on $\mathbb{R}^{k^\ell}$. Let us prove that $d(\boldsymbol{y}^\ell, \Gamma^\ell) > 0$ *a.s.* : the result will go along whatever $\sigma_\kappa \to 0$ or $\sigma_\kappa \to +\infty$ with the previous inequality. Let $\mathcal{Z}^\ell = \mathcal{Z}_{\text{pop}} \times \prod_{i=1}^{\ell} \mathcal{Z}_i$. Due to **(H 4)**, for all $i \in [\![1, n]\!]$,

$$\lim_{\|(\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})\|_\infty \to \infty} \|\gamma_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)\|_\infty \;=\; +\infty\,,$$

and so for all $\varepsilon \in \mathbb{R}_+^*$ non-negative, there exists $R \in \mathbb{R}$ such as for all $\boldsymbol{z}^\ell \in \mathcal{Z}^\ell$ satisfying $\|\boldsymbol{z}^\ell\| > R$, $\|\vec{\boldsymbol{\gamma}}^\ell(\boldsymbol{z}^\ell)\| > M + \varepsilon$. In particular, by definition of M, $\|\boldsymbol{y}^\ell - \vec{\boldsymbol{\gamma}}^\ell(\boldsymbol{z}^\ell)\|_\infty > 0$ for $\|(\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}})_{i\in[\![1,\ell]\!]}\|_\infty$ sufficiently large.

On the other hand, if at least a critical variable blows up, then by **(H 5)** there exists a critical trajectory γ_i^{crit} such that

$$\lim_{\|(\boldsymbol{z}_{\text{pop}}^{\text{crit}},\boldsymbol{z}_i^{\text{crit}})\|_\infty\to\infty}\|\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}},\boldsymbol{z}_i)\|_\infty \;=\; \gamma_i^{\text{crit}}$$

and as soon as this variable becomes sufficiently large, $\boldsymbol{y}_i \neq \gamma_i^{\text{crit}}$ *a.s.* Thus $\|\boldsymbol{y}^\ell - \vec{\boldsymbol{\gamma}}^\ell(\boldsymbol{z}^\ell)\|_\infty > 0$ *a.s.* for $\|(\boldsymbol{z}_{\text{pop}}^{\text{crit}}, \boldsymbol{z}_i^{\text{crit}})_{i\in[\![1,\ell]\!]}\|_\infty$ sufficiently large.
In other words, there exists $R \in \mathbb{R}_+^*$ such that for all $\boldsymbol{z}^\ell \in \mathcal{Z}^\ell$, if $\|\boldsymbol{z}^\ell\|_\infty > R$, then $\|\boldsymbol{y}^\ell - \vec{\boldsymbol{\gamma}}^\ell(\boldsymbol{z}^\ell)\|_\infty > 0$ *a.s.* So, by contraposition, if there exists $\boldsymbol{z}^\ell \in \mathcal{Z}^\ell$ such that $\|\boldsymbol{y}^\ell - \vec{\boldsymbol{\gamma}}^\ell(\boldsymbol{z}^\ell)\|_\infty = 0$ (at least *a.s.*) then $\|\boldsymbol{z}^\ell\|_\infty \leqslant R$. Especially,

$$\{\boldsymbol{z}^\ell \in \mathcal{Z}^\ell \mid \boldsymbol{y}^\ell = \vec{\boldsymbol{\gamma}}^\ell(\boldsymbol{z}^\ell) \text{ a.s.}\} \;\subset\; \bar{\mathcal{B}}(0,R)\,.$$

As **(H 3)** assumes that $P(\mathrm{d}\boldsymbol{y})$ has a continuous density and since $\vec{\boldsymbol{\gamma}}^\ell\left(\bar{\mathcal{B}}(0,R)\right)$ is a sub-manifold of dimension $p^\ell < k^\ell$, it comes that $P\left[\,\boldsymbol{z}^\ell \in \bar{\mathcal{B}}(0,R)\,\right] = 0$. Hence,

$$\mathscr{L}_{k^\ell}\left(\left\{\boldsymbol{y}^\ell \mid d(\boldsymbol{y}^\ell, \mathcal{I}m(\vec{\boldsymbol{\gamma}}^\ell)) = 0\right\}\right) \;=\; 0\,.$$

2. Let $f_\kappa(\boldsymbol{y}^\ell) = \sum_{i=1}^\ell \log q(\boldsymbol{y}_i|\theta_\kappa)$. From **(C 1)**, we deduce that, up to extraction, the negative part $\left(f_\kappa(\boldsymbol{y}^\ell)\right)^-$ is almost surely a non-decreasing and non-negative sequence converging to $+\infty$. From the monotone convergence theorem we then have

$$\liminf_{\kappa\to+\infty} \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\left(f_\kappa(\boldsymbol{y}^\ell)\right)^-\right] \;=\; +\infty$$

and so

$$\lim_{\kappa\to+\infty} \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\left(f_\kappa(\boldsymbol{y}^\ell)\right)^-\right] \;=\; +\infty\,.$$

Concerning the positive part $\left(f_\kappa(\boldsymbol{y}^\ell)\right)^+$, using the dominated convergence theorem, Lemma 3.3.2 and the point **(C 1)**, we get

$$\lim_{\kappa\to+\infty} \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\left(f_\kappa(\boldsymbol{y}^\ell)\right)^+\right] \;=\; 0\,.$$

Actually, for all $i \in [\![1,n]\!]$ the application $(\boldsymbol{z}_{\text{pop}}^{\text{reg}}, \boldsymbol{z}_i^{\text{reg}}) \mapsto \gamma_i^{\text{crit}}$ is continuous by continuity of the function $\vec{\boldsymbol{\gamma}}_i$ and so **(H′ 5)** holds.
Finally, we have proved that

$$\lim_{\kappa\to+\infty} \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\sum_{i=1}^\ell \log q(\boldsymbol{y}_i|\theta_\kappa)\right] \;=\; -\infty$$

and **(C 2)** follows immediately.

3. The continuity statement is straightforward. If Θ_*^ω is empty, any maximizing sequence θ_κ of $\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\log q(\boldsymbol{y}^\ell|\theta)\right]$ satisfies (up to extraction of a subsequence) $\theta_\kappa \in \Theta^\omega$, $\|\Sigma_\kappa\| \to +\infty$, $\|\Sigma_\kappa^{-1x}\| \to +\infty$, $\sigma_\kappa \to 0$ or $\sigma_\kappa \to +\infty$, which is on contradiction with conclusion **(C 2)**.

□

III – 2. Proof of the Consistency Theorem

We follow in the following proof the classical approach of van der Vaart (2000).

Proof of Theorem 3.2: As in Lemma 3.3.3, let $\overline{\Theta^\omega}$ denote the one point Alexandrov compactification of the parameter space Θ^ω. We have already proved at [Lemma 3.3.3 **(C 3)**] that $\Theta_*^\omega \neq \varnothing$. To achieve the proof, let us first demonstrate that for all $\theta_\infty \in \overline{\Theta^\omega}$ such that $\delta\left(\theta_\infty, \Theta_*^\omega\right) \geqslant \varepsilon$ there exists an open set $\mathcal{U} \subset \Theta^\omega$ such that

$$\frac{1}{\ell}\,\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\sup_{\theta\in\mathcal{U}\cap\Theta^\omega}\sum_{i=1}^{\ell}\log q(\boldsymbol{y}_i|\theta)\right] \;<\; \mathbb{E}^*(\omega)\,. \tag{0}$$

Let $\varepsilon \geqslant 0$, $(\mathcal{U}_h) \subset \Theta^{\omega\mathbb{N}}$ be a non-increasing sequence of open subsets of Θ^ω for which $\bigcap_{h\geqslant 0}\mathcal{U}_h = \{\theta_\infty\}$ and f_h be the function defined by

$$f_h\colon\; \boldsymbol{y}^\ell = (\boldsymbol{y}_i)_{i\in[\![1,\ell]\!]} \;\mapsto\; \frac{1}{\ell}\sup_{\theta\in\mathcal{U}_h}\sum_{i=1}^{\ell}\log q(\boldsymbol{y}_i|\theta)\,.$$

1. If $\theta_\infty \in \Theta^\omega$, through the continuity of the map $\theta \mapsto \sum_{i=1}^{\ell}\log q(\boldsymbol{y}_i|\theta)$ and the definition of the sequence $(\mathcal{U}_h)$,

$$\lim_{h\to+\infty} f_h(\boldsymbol{y}^\ell) \;=\; \frac{1}{\ell}\sum_{i=1}^{\ell}\log q(\boldsymbol{y}_i|\theta_\infty)\,.$$

So, according to the monotone convergence theorem, Lemma 3.3.2 and since $\theta_\infty \notin \Theta_*^\omega$,

$$\lim_{h\to+\infty}\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[f_h(\boldsymbol{y}^\ell)\right] \;=\; \frac{1}{\ell}\sum_{i=1}^{\ell}\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\log q(\boldsymbol{y}_i|\theta_\infty)\right] \;<\; \mathbb{E}^*(\omega)\,.$$

2. If $\theta_\infty \notin \Theta^\omega$, *i.e.* if $\theta_\infty \in \overline{\Theta^\omega}\setminus\Theta^\omega$, we can prove that for all observations $\boldsymbol{y}^\ell \in \mathbb{R}^{k^\ell}$ $\lim_{h\to\infty} f_h(\boldsymbol{y}^\ell) = -\infty$ $P(\mathrm{d}\boldsymbol{y}^\ell)$ *a.s.* We proceed by contradiction : assume that there exists a measurable set $A \in \mathscr{B}(\mathbb{R}^{k^\ell})$ such that

$\mathbb{P}(\boldsymbol{y}^\ell \in A) > 0$ and for all $\boldsymbol{y}^\ell \in A$, $\inf_{h\in\mathbb{N}} f_h(\boldsymbol{y}^\ell) > -\infty$. Then, by definition of the *infimum*, for all $\boldsymbol{y}^\ell \in A$ there exists a sequence $(h_n) \in \mathbb{R}^{\mathbb{N}}$ such as $\liminf_{n\to+\infty} f_{h_n}(y) > -\infty$. However for all $\boldsymbol{y}^\ell \in A$, $h \mapsto f_h(\boldsymbol{y}^\ell)$ is non-increasing and reaches its *infimum* limit for $h = +\infty$ and thus $\lim_{n\to+\infty} \mathcal{U}_{h_n} = \mathcal{U}_\infty = \{\theta_\infty\}$. Finally, up to considering a sequence $(\theta_{n,n'}) \in \mathcal{U}_{h_n}^{\mathbb{N}}$ for all subsets $\mathcal{U}_{h_n} \subset \Theta^\omega$ such that for all $n \in \mathbb{N}$,

$$\lim_{n'\to+\infty} \sum_{i=1}^{\ell} \log q(\boldsymbol{y}_i|\theta_{n,n'}) \;=\; \sup_{\theta\in\mathcal{U}_n} \sum_{i=1}^{\ell} \log q(\boldsymbol{y}_i \mid \theta)\,,$$

concatenating, reindexing those sequences and using the continuity of the map $\theta \mapsto \sum_{i=1}^{\ell} \log q(\boldsymbol{y}_i|\theta)$ we know that there exists a sequence $(\theta_n) \in \Theta^{\omega\mathbb{N}}$ such that

$$\lim_{n\to\infty} \theta_n = \theta_\infty \qquad \text{and} \qquad \liminf_{n\to+\infty} \sum_{i=1}^{\ell} \log q(\boldsymbol{y}_i|\theta_n) > -\infty\,.$$

Moreover, $\theta_\infty = (\overline{\boldsymbol{z}_{pop}}_\infty, \Sigma_\infty, \sigma_\infty) \in \overline{\Theta^\omega} \setminus \Theta^\omega$ and thus $\|\Sigma_\infty\| = +\infty$, $\|\Sigma_\infty^{-1}\| = +\infty$ or $\sigma_\infty \in \{0, +\infty\}$ in contradiction to [Lemma 3.3.3 **(C 1)**]. So for all observations $\boldsymbol{y}$, $\lim_{h\to\infty} f_h(\boldsymbol{y}^\ell) = -\infty$ $P(\mathrm{d}\boldsymbol{y})$ *a.s.* As in the proof of Lemma 3.3.3, Hypothesis **(H 5)** implies **(H′ 5)** and according to Lemma 3.3.2 and the monotone convergence theorem,

$$\lim_{h\to+\infty} \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)} \left[f_h(\boldsymbol{y}^\ell) \right] \;=\; -\infty \;<\; \mathbb{E}^*(\omega)\,.$$

That is, in both cases $\lim_{h\to+\infty} \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)} \left[f_h(\boldsymbol{y}^\ell) \right] \;<\; \mathbb{E}^*(\omega)$ and there exists an open set $\mathcal{U} \subset \Theta^\omega$ such that

$$\frac{1}{\ell}\, \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)} \left[\sup_{\theta\in\mathcal{U}\cap\Theta^\omega} \sum_{i=1}^{\ell} \log q(\boldsymbol{y}_i|\theta) \right] \;<\; \mathbb{E}^*(\omega)$$

as announced.

Let $K_\varepsilon = \{\theta \in \overline{\Theta^\omega} \mid \delta(\theta, \Theta_*^\omega) \geqslant \varepsilon\}$. Through the compactness of K_ε, there exists an open finite cover $(\mathcal{U}_\alpha)_{\alpha\in[\![1,A]\!]}$ of K_ε satisfying (0). Thus, denoting $q_n = \lfloor \frac{n}{\ell} \rfloor$ and $r_n = n - q_n\ell$ the quotient and the rest of the euclidean division of n by ℓ , we get for all $\theta \in K_\varepsilon$,

$$\sup_{\theta\in K_\varepsilon\cap\Theta^\omega} \sum_{i=1}^{n} \log q(\boldsymbol{y}_i|\theta) \;\leqslant\; \sup_{\alpha\in[\![1,A]\!]} \left(\sum_{q=0}^{q_n} \sup_{\theta\in\mathcal{U}_\alpha\cap\Theta^\omega} \sum_{r=1}^{\ell} \log q(\boldsymbol{y}_{q\ell+r}|\theta) \right.$$
$$\left. + \sup_{\theta\in\mathcal{U}_\alpha\cap\Theta^\omega} \sum_{r=\ell+1}^{r_n} \log q(\boldsymbol{y}_{q_n\ell+r}|\theta) \right).$$

However, according to the strong law of large numbers, Assumption **(H 2)** and (0),

$$\lim_{q_n\to\infty}\frac{1}{q_n}\sum_{q=0}^{q_n}\sup_{\theta\in\mathcal{U}_\alpha\cap\Theta^\omega}\sum_{r=1}^{\ell}\log q(\boldsymbol{y}_{q\ell+r}|\theta)\ \leqslant\ \ell\,\mathbb{E}^*(\omega)$$

hence, since $\lim_{n\to+\infty} q_n = +\infty$ and $r_n < \ell$ for all $n \in \mathbb{N}$,

$$\begin{aligned}\limsup_{n\to\infty}\left[\frac{q_n}{n}\sup_{\alpha\in[\![1,A]\!]}\left(\frac{1}{q_n}\sum_{q=0}^{q_n}\sup_{\theta\in\mathcal{U}_\alpha\cap\Theta^\omega}\sum_{r=1}^{\ell}\log q(\boldsymbol{y}_{q\ell+r}|\theta)\right)\right]\\ =\ \frac{1}{\ell}\times\sup_{\alpha\in[\![1,A]\!]}\left(\mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\left[\sup_{\theta\in\mathcal{U}_\alpha\cap\Theta^\omega}\sum_{r=1}^{\ell}\log q(\boldsymbol{y}_{q_n\ell+r}|\theta)\right]\right)\ <\ \mathbb{E}^*(\omega)\,.\end{aligned}$$

Otherwise, for all $r \in [\![\ell+1, \ell_n]\!]$, $\log q(\boldsymbol{y}_{q_n\ell+r}|\theta) \leqslant -k^\ell \log q(\sigma\sqrt{2\pi})$ so

$$\frac{1}{n}\sup_{\alpha\in[\![1,A]\!]}\left(\sup_{\theta\in\mathcal{U}_\alpha\cap\Theta^\omega}\sum_{r=\ell+1}^{r_n}\log q(\,\boldsymbol{y}_{q_n\ell+r}|\theta\,)\right)\ \leqslant\ \frac{k^\ell(r_n-1)}{n}\log(\sigma\sqrt{2\pi})\,.$$

Thereafter

$$\limsup_{n\to\infty}\left[\frac{1}{n}\sup_{\alpha\in[\![1,A]\!]}\left(\sup_{\theta\in\mathcal{U}_\alpha\cap\Theta^\omega}\sum_{r=\ell+1}^{r_n}\log q(\,\boldsymbol{y}_{q_n\ell+r}|\theta\,)\right)\right]\ \leqslant\ 0$$

and

$$\limsup_{n\to\infty}\frac{1}{n}\sup_{\theta\in K_\varepsilon\cap\Theta^\omega}\sum_{i=1}^{n}\log q(\boldsymbol{y}_i|\theta)\ <\ \mathbb{E}^*(\omega)\,. \tag{1}$$

By definition of Θ_*^ω and according to the strong law of large numbers and **(H 2)**, for all $\theta^* \in \Theta_*^\omega$ $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \log q(\boldsymbol{y}_i|\theta^*) = \mathbb{E}^*(\omega)$ *a.s.* Moreover for all $i \in [\![1, n]\!]$,

$$q(\boldsymbol{y}_i|\hat\theta_n)\ =\ \frac{q(\hat\theta_n\mid\boldsymbol{y}_i)\,q(\boldsymbol{y}_i)}{q_{prior}(\hat\theta_n)}\ \geqslant\ \frac{q(\theta_*\mid\boldsymbol{y}_i)\,q(\boldsymbol{y}_i)}{q_{prior}(\hat\theta_n)}\ =\ \frac{q(\boldsymbol{y}_i|\theta_*)\,q_{prior}(\theta_*)}{q_{prior}(\hat\theta_n)}$$

and so

$$\sum_{i=1}^{n}\log q(\boldsymbol{y}_i|\hat\theta_n)\ \geqslant\ \sum_{i=1}^{n}\log q(\boldsymbol{y}_i|\theta_*)\ +\ \Big(\log q_{prior}(\theta_*)-\log q_{prior}(\hat\theta_n)\Big)\,.$$

Since q_{prior} is upper-bounded on Θ^ω, there exists $M \in \mathbb{R}^+$ such that

$$\frac{1}{n}\Big(\log q_{prior}(\theta_*)-\log q_{prior}(\hat\theta_n)\Big)\ \geqslant\ \frac{1}{n}\log\left(\frac{q_{prior}(\theta_*)}{M}\right)$$

i.e. $\liminf_{n\to+\infty}\frac{1}{n}\Big(\log q_{prior}(\theta_*)-\log q_{prior}(\hat\theta_n)\Big)\geqslant 0$ and

$$\liminf_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}\log q(\boldsymbol{y}_i|\hat\theta_n)\ \geqslant\ \mathbb{E}^*(\omega)\,. \tag{2}$$

The result follows from Equations 1 and 2 by contradiction : Assume that for all $n \in \mathbb{N}$, $\hat{\theta}_n \in K_\varepsilon$ *i.e.* that $\delta(\hat{\theta}_n, \Theta_*^\omega) \geqslant \varepsilon$. Then

$$\sum_{i=1}^{n} \log q(\boldsymbol{y}_i|\hat{\theta}_n) \leqslant \sup_{\theta \in K_\varepsilon \cap \Theta^\omega} \sum_{i=1}^{n} \log q(\boldsymbol{y}_i|\theta)$$

and by taking the limit superior, we get

$$\mathbb{E}^*(\omega) \overset{(2)}{\leqslant} \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log q(\boldsymbol{y}_i|\hat{\theta}_n) \overset{(1)}{<} \mathbb{E}^*(\omega)$$

i.e. $\mathbb{E}^*(\omega) < \mathbb{E}^*(\omega)$. Hence $\lim_{n\to\infty} \mathbb{P}\left[\delta(\hat{\theta}_n, \Theta_*^\omega) \geqslant \varepsilon\right] = 0$. □

IV. Discussion and Perspective

We have proposed a coherent statistical framework for the spatio-temporal analysis of piecewise geodesic manifold valued measurements. This model allows each individual to have his own intrinsic geometry and his own time parametrization. Unlike previous similar works (Schiratti et al., 2015, 2017), it allows for piecewise geodesic trajectories. Relaxing the classical geodesic assumption widen the application scope of the model in biology and medicine. The model is built in a hierarchical way as a non-linear mixed effects model whose fixed effects define a representative trajectory of the global evolution in the space of measurements and random effects account for the spatio-temporal variability of the trajectories at the individual level.

Estimation was formulated as a well-defined MAP problem and numerically performed through the MCMC-SAEM algorithm. Experimentations performed at Chapter 5 highlight the robustness of our model to noise and its performance in catching individual behaviors. We believe that the complexity of our model ensures its practical identifiability, even if it is not structurally identifiable (Lavielle and Aarons, 2016). Besides, as the posterior likelihood is not convex, the MAP could be difficult to determine numerically.

Part III. focuses on exploring some possible improvement of the numerical scheme.

Our model can be applied to a wide variety of situations and data sets. In particular, we can address medical follow-up such as neurodegenerative diseases or chemotherapy monitoring. The example of chemotherapy monitoring developed in the following chapter is especially interesting in a modeling perspective as the patients are treated and tumors may respond, stabilize or progress during the treatment, with different conducts for each phase. At the age of personalized medicine, to give physicians decision support systems is really important. Therefore learning correlations between phases is crucial. This has

been taken into account in our experimentations. More generally, the inter-individual variability allows us to personalize the model to new patient and thus perform predictive medicine.

– Chapter IV –

Encode 3D Anatomical Shapes: Currents and Varifolds Shape Spaces

Afin de permettre l'étude de modèles pour les formes anatomiques en 3 dimension, on redonne dans le présent chapitre et sans démonstration les fondements mathématiques de deux espaces de forme très utilisés dans ce contexte : les courants (Vaillant et Glaunès, 2005) et les varifolds (Charon et Trouvé, 2013).

Contents

WITHOUT pretending to make a complete or self-contained presentation of the wide theory of currents and varifolds, we try to outline in this chapter the minimal theoretical background needed for the following chapter.

I. The Representation by Currents

Currents were historically introduced by De Rham (1973) as a way to generalize the Schwartz distributions to the space of differential forms. Later, Federer (1969) developed widely the theory and connected currents to geometric measure theory. However, the use of currents for computational anatomy is more recent and goes back to Glaunès (2005)' and Durrleman (2010)'s works during their respective PhD.

The essential features of currents is that they provide an embedding of the set of all shapes of a given dimension into a common Banach space. Moreover, currents provide an intrinsic representation of shapes, in the sense that it is independent of shape parametrization.

I – 1. Differential Forms and Currents

We start by briefly covering the mathematical foundations of differential forms and providing some notations. We take consistent notations with Durrleman (2010).

Let E be a euclidean space of dimension n and $d \in [\![0, n]\!]$. We denote $\Lambda^d E$ the d-times exterior power of E, which is a vector space of dimension $\binom{n}{d}$ spanned by the set of simple d-vectors $\xi_1 \wedge \ldots \wedge \xi_d$. We remind that the wedge-product is bilinear and antisymetric by construction. Moreover, whether $d = 1$ or $d = n - 1$, $\Lambda^d E = E$. $\Lambda^d E$ is equipped with the euclidean metric given for two simple d-vectors $\xi = \xi_1 \wedge \ldots \wedge \xi_d$ and $\zeta = \zeta_1 \wedge \ldots \wedge \zeta_d$ by the determinant of the Gram matrix $\langle \xi \,|\, \zeta \rangle = \mathcal{D}et\left(\langle \xi_i \,|\, \zeta_j \rangle\right)_{i,j}$. In particular, $|\xi|$ gives the volume of the corresponding parallelotope. We denote by $\Lambda^d E^*$ the dual vector space of $\Lambda^d E$, which can be identified to the space of alternate d-forms on E (Durrleman, 2010).

We note $\Omega_0^d(E) = \mathscr{C}_0^0\left(E, \Lambda^d E^*\right)$ the set of continuous d-dimensional forms on E vanishing at infinity equipped with the infinite norm

$$\forall \omega \in \Omega_0^d(E), \quad \|\omega\|_\infty \;=\; \sup_{x \in E} \sup_{|\xi|=1} \omega_x(\xi)\,,$$

which makes it a Banach space. Note that if $\mathrm{d}x^1, \ldots, \mathrm{d}x^n$ is a basis of E^*, any $\omega \in \Omega_0^d(E)$

can be written as

$$\forall x \in E, \quad \omega_x \;=\; \sum_{1 \leqslant i_1 < \ldots < i_d \leqslant n} a_{i_1,\ldots,i_d}\, \mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^d\,,$$

where $a_{i_1,\ldots,i_d}$ are continuous functions, vanishing at infinity. This definition can obviously be extended to k-times continuously differentiable forms, leading to the space $\Omega^d_{0,k}(E)$ of the k-times continuously differentiable forms ω such that ω and all its differentials up to order k vanish at infinity.

Definition 4.1. *The space of d-currents on E is the space $\Omega^d_0(E)'$ of all continuous linear forms on the space of differential forms.*

Note that the above definition is quite more restrictive that the classical one (Federer, 1969; Morgan, 1995). in the particular case $d = 0$, the previous definition is nothing more than the definition of usual distribution on E, *i.e.* the dual of the space of real-valued functions. As well as for distributions, the simplest example of currents is given by *Dirac currents* δ^ξ_x defined as

$$\forall x \in E, \quad \forall \xi \in \Lambda^d E, \quad \forall \omega \in \Omega^d_0(E), \quad \delta^\xi_x(\omega) = \omega_x(\xi)\,.$$

I – 2. Oriented Shapes as Currents

As an extension of distributions, currents were meant to carry information of local geometry: this property is crucial for computational anatomy application and we will now detail the relationship between currents and shapes, namely how every d-dimensional oriented submanifolf X of E of finite volume can be represented by an element of $\Omega^d_0(E)'$.

A well-known result of integration (Boothby, 1986) ensures us that any d-dimensional differential form can be integrated along X. We thus define an element of $\Omega^d_0(E)'$ by setting

$$T_X\colon\, \omega \;\mapsto\; \int_X \omega\,.$$

Moreover, this representation is *geometric* in the sense that it only depends of the shape X and not of its parametrization. Indeed, assume that X is given through a parametrization $\gamma\colon \mathcal{U} \to E$, where $\mathcal{U}$ is an open subset of $\mathbb{R}^d$. The previous expression thus rewrites

$$\forall \omega \in \Omega^d_0(E), \quad T_X(\omega) \;=\; \int_{\mathcal{U}} \omega_{\gamma(u)} \left(\frac{\partial \gamma}{\partial u_1} \wedge \ldots \wedge \frac{\partial \gamma}{\partial u_d} \right) \mathrm{d}u_1 \ldots \mathrm{d}u_d\,,$$

which is independent of any positive parametrization according to the change of variable formula.

Obviously, the set of d-currents is much larger. For instance, we can generalize submanifolds from the point of view of measure theroy, leading to consider rectifiable subsets of E (*cf* Section II.2 for a definition of rectifiable subsets), which are still currents and can be treated exactly in the same way as submanifolds: by integrating differential

forms. In the literature, we refer to this type of currents as rectifiable currents (Federer, 1969).

As explained in Section II.1 (or II.1 in french), the definition of shape spaces is completed if we can define the action of diffeomorphims over the current space $\Omega_0^d(E)'$. We define the *transport* of currents by diffeomorphisms of E through the pull-back and push-forward operations. More precisely, for all currents $T \in \Omega_0^d(E)'$ and all diffeomorphisms $\phi \in \mathscr{C}^1(E)$ whose differential is bounded, the *push-forward* of T is the current defined by

$$\forall \omega \in \Omega_0^d(E), \quad (\phi_\# T)(\omega) \;=\; T(\phi^\# \omega)\,,$$

where the differential form $\phi^\# \omega$ refers to the *pull-back* of ω and is defined by

$$\forall x \in E, \quad \forall \xi = (\xi_1 \wedge \ldots \wedge \xi_d) \in \Lambda^d E, \quad (\phi_\# \omega)_x(\xi) \;=\; \omega_{\phi(x)}(\mathrm{d}_x\phi(\xi_1) \wedge \ldots \wedge \mathrm{d}_x\phi(\xi_d))\,,$$

and where $\mathrm{d}_x\phi$ denotes the differential of the morphism ϕ at the point x.

Last, the change of variable for the integration of differential forms (Federer, 1969) gives that for all shape X of E, *i.e.* a submanifold of E or more generally a rectifiable subset of E, and for all diffeomorphism ϕ whose differential is bounded,

$$\phi_\# T_X = T_{\phi(X)}$$

and we thus transport the shape X through the diffeomorphism ϕ, in the common sense of transport.

To sum up, the currents framework gives us the capability to represent any d-dimensional oriented submanifolf X of E of finite volume as a geometric object, *i.e.* independently of the parametrization of X, and to transport it.

As the dual space of differential forms, the space of currents naturally inherits its operator norm, which is called the *mass norm* (Federer, 1969; Morgan, 1995) and defined as

$$\forall T \in \Omega_0^d(E)', \quad M(T) \;=\; \sup_{\|\omega\|_\infty \leqslant 1} |T(\omega)|\,.$$

But, as noticed by Durrleman (2010), this norm is not adapted for shape comparison. To overcome this issue and based on the very good results obtained within the LDDMM framework (*cf.* page 39 or 11 in french), Glaunès (2005) proposed to introduce again a reproducing kernel Hilbert space (RKHS).

This approach consists in defining a vector kernel $K\colon E \times E \to \mathcal{L}(\Lambda^p E)$ on E and its associated RKHS W. It is then shown that, under some regularity assumptions on the kernel, the space of d-currents is continuously embedded in the dual W' which is also a Hilbert space. As a result, currents and therefore shapes can be compared using a Hilbert norm.

I – 3. Currents for Meshed Surfaces

All previous definitions deal with properly defined submanifolds of E. In practice, we have only access to meshed surfaces. Fortunately, the formalism introduced above adapts with no change to the case of discrete geometry: any meshed shape can be transcribed into a current by approximating each cell of the mesh with one Dirac current located at the center of the cell and a simple d-vector encoding the local volume element (Durrleman, 2010). As a consequence, the meshed d-dimensional shape can be approximated by a finite sum of Dirac currents of the form

$$T_X \;\simeq\; \sum_{i=1}^{m} \delta_{x_i}^{\xi_i} \,.$$

We detail thereafter the case of 3 dimension.

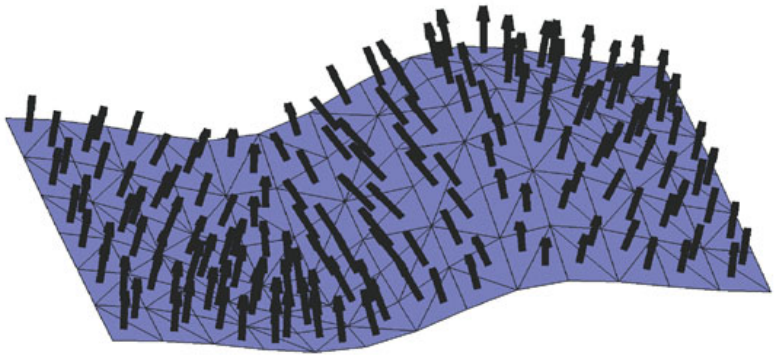


Figure 4.1 *– Representation of a meshed surface in Dirac currents.*
Illustration taken from Charon and Trouvé (2014).

Definition 4.2. *Let $d \in [\![1,3]\!]$. An oriented d-mesh in $E = \mathbb{R}^3$ is a finite collection of oriented d-dimensional simplexes. Each simplexe is called face or cell of the mesh.*

For each cell $i \in [\![1,m]\!]$, we denote $x_1^i, \ldots, x_{d+1}^i$ its vertices and $\bar{x}_i = \frac{1}{d+1}\sum_{j=1}^{d+1} x_j^i$ its center of mass. For each cell i and all edge $j \in [\![1,d]\!]$, we define d vectors parallel to the edges of the cell i by setting

$$\xi_j^i \;=\; \frac{1}{(d!)^{1/d}} \left(x_{j+1}^i - x_1^i\right) \,.$$

In the previous sum, we chose the order of the vertices so that the basis $(\xi_j^i)_{j\in[\![1,d]\!]}$ has the same orientation as the mesh cell, in such a way that $|\xi^i| = |\xi_1^i \wedge \ldots \wedge \xi_d^i|$ is the d-volume of the cell i.

In definition 4.2, we require the faces of the mesh to be simplexes, so that the surfaces mesh cells must be triangles ($d = 2$) or tetrahedrons ($d = 3$). Obviously, we can extend

the previous definition to whatever type of faces but the vector thus have to be defined carefully to keep the interpretation of their norm as being equal to the d-volume of the cell. Moreover, the definition can be extended to any dimension.

Note that for any scalar v and any d-vector α, β and any point $x \in \mathbb{R}^d$ we have

$$v\,\delta_x^\alpha + \delta_x^\beta \;=\; \delta_x^{v\alpha+\beta}\,.$$

Proposition 4.3 (*Durrleman (2010)*)

Let X a d-mesh. Assume that the differential forms $\omega \in \Omega_0^d(\mathbb{R}^3)$ are $\mathscr{C}^1$ and verify $\|\nabla_x\omega\|_\infty \leqslant C_\infty\|\omega\|_\infty$ for a fixed constant C_∞. Let the current

$$\widetilde{T}_{\widetilde{X}} \;=\; \sum_{i=1}^{m} \delta_{\bar{x}_i}^{\xi^i}\,, \qquad \text{where} \quad \bar{x}_i = \frac{1}{d+1}\sum_{j=1}^{d+1} x_j^i$$

and $\xi^i = \xi_1^i \wedge \ldots \wedge \xi_d^i$. Then,

$$M(T_X - \widetilde{T}_{\widetilde{X}}) \;\leqslant\; C_\infty\, M(T_X) \max_{i\in[\![1,m]\!]} \mathcal{D}iam(c_i)\,,$$

where $\mathcal{D}iam(c_i)$ denotes the diameter of the cell c_i, $i \in [\![1,m]\!]$.

In particular, if the sampling of the mesh becomes finer, that is to say if $\max_{i\in[\![1,m]\!]} \mathcal{D}iam(c_i) \to 0$, the current $\widetilde{T}_{\widetilde{X}}$ converges toward T_X.

Last but not least, thanks to the RKHS nature of W, the inner product between two Dirac currents writes:

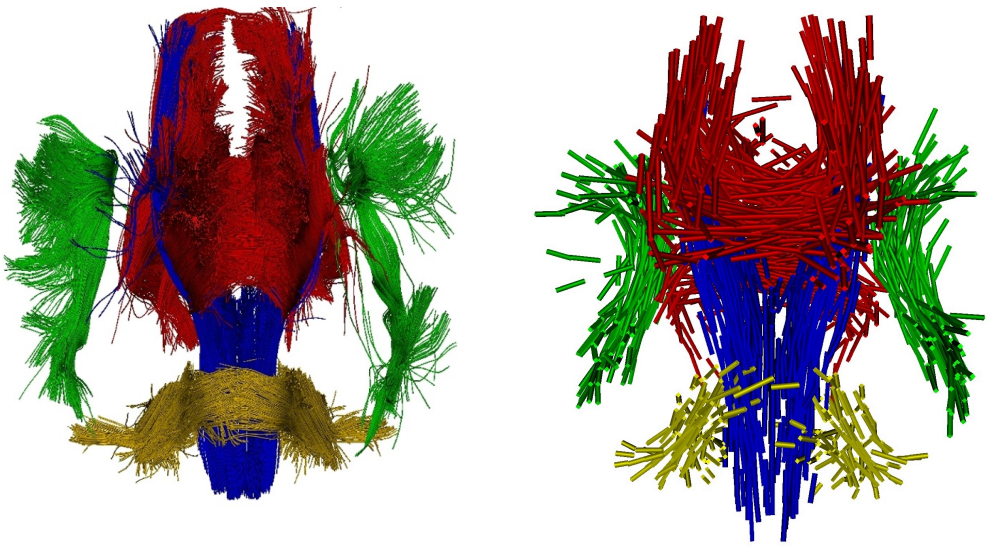
$$\forall x_1, x_2 \in E \quad ,\forall \xi_1, \xi_2 \in \Lambda^d E, \quad \left\langle \delta_{x_1}^{\xi_1} \,\middle|\, \delta_{x_2}^{\xi_2} \right\rangle_{W'} \;=\; \left\langle \xi_1 \,\middle|\, K(x_1,x_2)\xi_2 \right\rangle .$$

In this discrete setting, since shapes are represented as finite sums of Dirac, the RKHS distance allows for close form computations as double sums of kernel equation. We thus obtain a corresponding attachment term between any shapes X and Y by setting

$$A(X,Y) \;=\; \|T_X - T_Y\|_{W'}^2\,.$$

I – 4. The Issue of Shape Orientation

As mentioned previously, the current associated to a given d-dimensional submanifold or more generally d-dimensional rectifiable subset X of E depends on its orientation. For instance, if we denote $\tilde{X}$ the same set than X but with opposite orientation, then $T_{\tilde{X}} = -T_X$. As a consequence, to compare two shape in the currents' setting there, the orientation of both shapes have to be consistent with each other.



(a) White matter fibber bundle *(b) Mathematical representation*

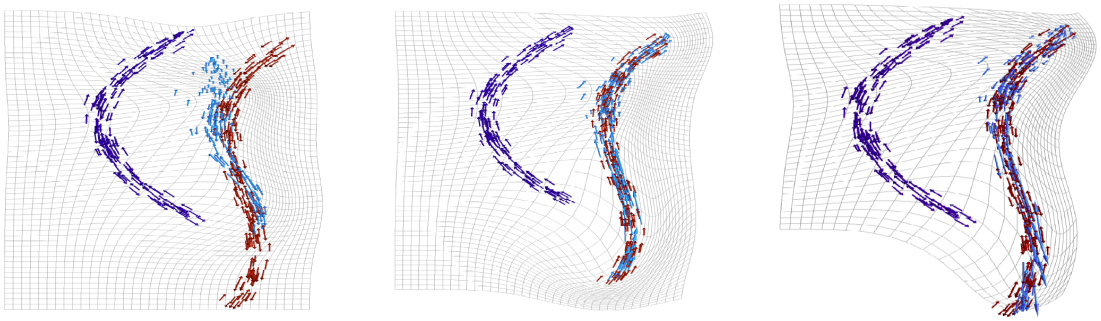
***Figure 4.2** – White matter fibber bundle.*
White matter fiber bundle estimated from Diffusion Tensor Imaging (DTI) illustrating the potential difficulty of consistent orientation of all different fibers.
Credit: *D. Ducreux, Hôpital Kremlin-Bicêtre et P. Fillard, CEA, Neurospin.*

A second issue, and more damaging, is that currents with opposite orientation within a small space domain may cancel each other with respect to the kernel metric. Let consider a Gaussian kernel with scale σ as a typical example, then

$$\forall x, y \in E, \quad \forall \xi \in \Lambda^d E, \quad \|\delta_x^\xi - \delta_y^{-\xi}\| \;=\; 2|\xi|^2 \left(1 - \mathrm{e}^{-\frac{|x-y|^2}{\sigma^2}}\right),$$

which vanish whenever $|x - y|$ is small compared to σ. As claimed by Charon and Trouvé (2013), this trivial fact has important consequences in practice. Indeed, even if orientable, orientating shapes consistently can be either a difficult or even ill-posed problem in certain datasets. However, this type of dataset may arrive naturally in the context of computational anatomy. This is for example the case of white matter fiber bundle which is composed of many different and possibly disconnected pieces of curves as illustrates Figure 4.2. Indeed, given a shape composed of N connected component, there are 2^N different possible orientations.

We report here an experiment conducted by Charon and Trouvé (2013) to illustrate the compulsory need for consistent orientations as we find that it illustrates well this drawback in an easy setting. They consider the matching between two fiber bundles consisting of many different pieces of curves with three settings: currents being oriented randomly, currents with a consistent orientation and varifolds (*cf* below). The results are displayed at Figure 4.3.



(a) Random orientation *(b) Consistent orientation* *(c) Varifolds*

***Figure 4.3** – Synthetic fiber bundles.*
Matching between two synthetic fiber bundles consisting of many different pieces of curves, each of them being oriented randomly in the currents' framework (Figure 4.3a) versus a consistent orientation of all pieces still in the currents' framework (Figure 4.3b) versus the varifolds's framework. The source shape is in dark blue, the target in red and the deformed source is in light blue.
Illustration taken from Charon and Trouvé (2013).

II. The Representation by Varifolds

Varifolds have been first introduced in the context of geometric measure theory (Allard, 1972; Almgren, 1966). Here, based on the PhD work of Charon (2013), we explain the fundamental relationship between varifolds. We also explain how varifolds can encode numerically non-oriented objects and so be usefull to computational anatomy.

II – 1. Grassmannian and Varifolds

Let E be a euclidean space of dimension n and $d \in [\![0, n]\!]$. The basic idea behind the varifolds is to represent any submanifold, and more generally any rectifiable set, oriented or not, as a distribution of unoriented tangent spaces spread in the embedding space E. For that purpose, we need a proper way to represent tangent spaces of dimension d in E, leading to consider the Grassmann manifold.

Definition 4.4. *The Grassmann manifold of dimension d in E, denoted $G_d(E)$, is the set of all d-dimensional subspaces of E.*

Note that the Grassmann $G_d(E)$ manifold can be identified to the quotient space of all families of d independent vectors of E by the equivalence relation obtained by identifying families that generate the same subspace. As a quotient space, one can show that $G_d(E)$ inherits a structure of Riemannian manifold of dimension $d(n-d)$ (Wong, 1967). Moreover, in the particular case where $d = 1$ or $d = n - 1$, the Grassmann manifold is nothing else that the real projective space of E.

Definition 4.5. *A d-dimensional varifold on E is a distribution on the product space $E \times G_d(E)$, i.e. an element of $\mathscr{C}_0\left(E \times G_d(E)\right)'$.*

Likewise than in the currents' framework, one can define Dirac varifolds:

$$\forall x \in E, \quad \forall V \in G_d(E), \quad \forall \omega \in \mathscr{C}_0\left(E \times G_d(E)\right), \quad \delta_{(c,V)}(\omega) \;=\; \omega(x, V)\,.$$

In other words, a Dirac varifold is the data of a position $x \in E$ and a d-dimensional (non-oriented) subspace V that will play the role of tangent space. In particular and contrary to currents' approach, the varifold encompass a notion of direction through the tangent space V.

II – 2. Rectifiable Sets and Hausdorff Measure

In the following sections, we will explain in what respect the varifolds offer a natural framework to embed non-oriented shapes in the context of computational anatomy both from the theoretical and numerical point of view. However, the variflods' framework relies on the notion of Rectifiable sets and Hausdorff measure, which we thus recall here. Rectifiable subsets can be thought as the proper generalization of submanifolds in the context of geometric measure theory. Let first recall the notion of Hausdorff measure.

Basically, the Hausdorff measure $\mathcal{H}^d$ is an outer measure that measures the d-dimensional volume of a subset in E. Properly, given a real number $r > 0$, we define

$$\forall X \subseteq E, \quad \mathcal{H}^d_r(X) \;=\; \inf\left\{ \sum_{i=1}^{+\infty} (\mathcal{D}iam\, \mathcal{B}_i)^d \,\middle|\, X \subseteq \bigcup_{i=1}^{+\infty} \mathcal{B}_i\,,\; \mathcal{D}iam\, \mathcal{B}_i < r \right\},$$

where the infimum is over all countable covers of X by sets $\mathcal{B}_i \subset X$ satisfying $\mathcal{D}iam\, \mathcal{B}_i < r$. As $r \mapsto \mathcal{H}^d_r(X)$ is monotone non-increasing, we can define the limit

$$\forall X \subseteq E, \quad \mathcal{H}^d(X) \;=\; \lim_{r \to 0} \mathcal{H}^d_r(X)\,,$$

which is possibly infinite. It is called the d-dimensional Hausdorff measure of X.

In particular, $\mathcal{H}^n = \mathscr{L}_n$ the usual Lebesgue measure. Moreover, if X is a p-dimensional submanifold of E, then $\mathcal{H}^d(X)$ is rigorously the d-volume of X if $d = p$, vanish if $p < d$ and equals $+\infty$ whenever $d < p$.

Definition 4.6. *A subset $X \subseteq E$ is said to be a d-dimensional rectifiable subset of E if:*

(i) $\mathcal{H}^d(X) < +\infty$ *;*

(ii) There exists a countable family $(\psi_i)_{i \in \mathbb{N}}$ of Lipschitz maps $\psi_i \colon \mathbb{R}^d \to E$ such that

$$\mathcal{H}^d\left(X \setminus \bigcup_{i \in \mathbb{N}} \phi_i(\mathbb{R}^d) \right) \;=\; 0\,.$$

In other words, a rectifiable subset is a subset of E which can be almost everywhere covered by a countable union of images of Lipschitzt maps form $\mathbb{R}^d$ to E and it is easy to convince oneself that this notion generalizes submanifolds. We take the definition of Charon (2013) but note that this definition actually referred to countable $\mathcal{H}^d$-rectifiable subset in Federer (1969).

II – 3. Non-Oriented Shapes as Varifolds

As well as the currents which were an efficient way to encode oriented shapes, the varifolds are the right notion to encore non-oriented ones. Indeed, let X be a non-oriented submanifold or rectifiable set of E of dimension d. One can associate to the set X a varifold μ_X which, in the measure point of view, is given by

$$\forall A \in \mathscr{B}\left(E \times G_d(E)\right), \quad \mu_X(A) \;=\; \mathcal{H}^d\left(\{x \in X | (x, T_x X) \in A\}\right)$$

or alternatively, as a continuous linear form on $\mathscr{C}_0\left(E \times G_d(E)\right)$,

$$\begin{aligned} \forall \omega \in \mathscr{C}_0\left(E \times G_d(E)\right), \quad \mu_X(\omega) \;&=\; \int_{E \times G_d(E)} \omega(x; V)\, \mathrm{d}\mu_X(x, V) \\ &=\; \int_X \omega(x, T_x X)\, \mathrm{d}\mathcal{H}^d(x)\,. \end{aligned}$$

Last, if X is more simply a smooth compact submanifold of E given by a parametrization $\gamma : \mathcal{U} \to E$, where $\mathcal{U}$ is an open subset of $\mathbb{R}^d$ then,

$$\forall \omega \in \mathscr{C}_0\left(E \times G_d(E)\right), \quad \mu_X(\omega) \;=\; \int_{\mathcal{U}} \omega(\gamma(u), T_{\gamma(u)} X)\, |\gamma'(u)|\, \mathrm{d}u\,,$$

where $\gamma'(u) = \bigwedge_{i=1}^d \frac{\partial \gamma}{\partial u_i} \in \Lambda^d E$ and $|\gamma'(u)|$ represents the local d-volume element. As for the currents' framework, the above formula is independent of the parametrization and we define a geometric representation again, *i.e.* a representation that depends uniquely on the shape X and not on the parametrization of this last one. Moreover and unlike in the currents' setting, here, the formula is also independent of the sign of the orientation.

In the same spirit than currents' setting, we now focus on the definition of a notion of transport of a varifold by a diffeomorphism in a way which is compatible with the intuitive notion of transport of shapes. Likewise, we define the transport of a varifold through the pull-back and the push-forward operations. Let $\mu \in \mathscr{C}_0\left(E \times G_d(E)\right)'$ be a varifold and $\phi \in \mathscr{C}^1(E)$ a diffeomorphism whose differential is bounded. Then, we define the transport $\phi_* \mu$ of μ by ϕ as the varifold defined by

$$\forall \omega \in \mathscr{C}_0\left(E \times G_d(E)\right), \quad (\phi_* \mu)(\omega) \;=\; \mu(\phi^* \omega)\,,$$

where $\phi^* \omega$ is the pull-back of ω by ϕ defined as

$$\forall x \in E, \quad \forall V \in G_d(E), \quad (\phi^* \omega)(x, V) \;=\; |\mathrm{d}_x \phi(u_1) \wedge \ldots \wedge \mathrm{d}_x \phi(u_d)|\; \omega(\phi(x), \mathrm{d}_x \phi \cdot V)\,,$$

where $(u_1, \ldots, u_d)$ is an orthogonal basis of V, , $\mathrm{d}_x\phi \cdot V$ denotes the element of $G_d(E)$ that is the image of V by $\mathrm{d}_x\phi$, and where term $|\mathrm{d}_x\phi(u_1) \wedge \ldots \wedge \mathrm{d}_x\phi(u_d)|$ is the d-dimensional Jacobian of ϕ on the subspace V and represents, as in the standard real case, the local change of d-dimensional volume of the transformation.

Last, for any d-dimensional shape and diffeomorphism ϕ (Charon and Trouvé, 2013),

$$\phi_* \mu_X \;=\; \mu_{\phi(X)}$$

and then the transported varifold transports the shape as desired.

Finally, the varifold setting turn out to be an efficient framework to represent non-oriented d-dimensional shapes, at least from the continuous setting. The next section is dedicated to discrete geometry and more specifically to meshed surfaces. Before, we focus on the definition of an adapted metric for varifolds comparison.

Reproducing kernels provided regularized metrics for the comparison of currents. In particular, there were particularly well-fitted to compute distances between discretized shapes because of the simple expression of dot product between two current Diracs. Charon and Trouvé (2013) proposed to use a similar approach to endow the varifolds with an adapted metric. However, the complex nature of varifolds makes the construction more tricky. In particular, kernel metric does not provide necessarily real distances on varifolds. However, a wide class of kernels (Proposition 4.8) overcome this issue. A brief introduction is provided here and we can refer to Charon (2013) or Charon and Trouvé (2013) for a more complete construction.

Proposition 4.7 (*Charon and Trouvé (2013)*)

Assume that we are given a positive real kernel k_e on the space E such that k_e is continuous, bounded and for all $x \in E$, the function $k_e(x; \cdot)$ vanishes at infinity. Assume that a second kernel k_t is defined on the manifold $G_d(E)$ and is also continuous. Then the RKHS W associated to the kernel $k_e \otimes k_t$ is continuously embedded into the space $\mathscr{C}_0(E \times Gd(E))$.

Consequently, there exists a continuous mapping of the space of varifolds $\mathscr{C}_0(E \times Gd(E))'$ into the dual of W. Just as for currents, we can compare varifolds and unoriented subsets through the Hilbert norm of W'.

Proposition 4.8

Let $k = k_e \otimes k_t$ be a kernel as in Proposition 4.7. Assume that kernel k_e is $\mathscr{C}_0$-universal and that the kernel k_t is such that $k_t(V; V) > 0$ for all $V \in G_d(E)$. Let $X = \bigcup_{i=1}^{N} X_i$ and $Y = \bigcup_{j=1}^{M} X_j$ be two finite unions of compact d-dimensional submanifolds of E. If $\|\mu_X - \mu_Y\| = 0$ then $X = Y$.

Finally, the main difficulty is to endow the Grassmann manifold $G_d(E)$ with a reproducing kernel. More approaches are allowed for this purpose. Charon and Trouvé

(2013) dedicate a whole section of their paper to this issue.

II – 4. Varifolds for Meshed Surfaces

Exactly in the same spirit as with currents, any meshed surfaces X can be coded by a finite sum of Dirac varifolds of the form

$$\mu_X \simeq \sum_{i=1}^{m} v_i \delta_{x_i,V_i} ,$$

where m denotes the number of cells in the mesh and, for each cell $i \in [\![1,m]\!]$, $x_i \in E$ is the center of the cell, $V_i \in G_d(E)$ the tangent space to the shape at the point x_i and $v_i \in \mathbb{R}_+^*$ the d-volume of the cell.

More precisely, with the same notations than the currents' ones, for each cell $i \in [\![1,m]\!]$ we denote $x_1^i, \ldots, x_{d+1}^i$ its vertices. Then, for each cell i, we define its center of mass $\bar{x}_i$ and d vectors parallel to its edges by setting

$$\bar{x}_i = \frac{1}{d+1} \sum_{j=1}^{d+1} x_j^i \qquad \text{and} \qquad \forall j \in [\![1,d]\!], \quad \xi_j^i = \frac{1}{(d!)^{1/d}} \left(x_{j+1}^i - x_d^i \right) .$$

We also define the vector space V_i to be tangent to the shape at the point x_i as

$$V_i = span(\xi_1^i, \ldots, \xi_d^i) .$$

Last, as previously, the d-volume of the cell i is given by $v_i = |\xi^i| = |\xi_1^i \wedge \ldots \wedge \xi_d^i|$.

Likewise, we define here simplexes mesh but the simplexe assumption can be relaxed.

The transport of a Dirac varifold can be computed straightforwardly: Let ϕ be a diffeomorphism whose differential is bounded. Then,

$$\forall x \in E, \quad \forall V \in G_d(E), \quad \phi_* \delta_{x,V} = |\mathrm{d}_x\phi(u_1) \wedge \ldots \wedge \mathrm{d}_x\phi(u_d)| \; \delta_{\phi(x),\mathrm{d}_x\phi \cdot V}$$

if $(u_1, \ldots, u_d)$ is an orthonormal basis of V. Moreover, the inner product between two Dirac varifolds writes (Charon and Trouvé, 2013)

$$\forall x_1, x_2 \in E, \quad \forall V_1, V_2 \in G_d(E), \quad \left\langle \delta_{x_1,V_1} \middle| \delta_{x_2,V_2} \right\rangle_{W'} = k_e(x_1,x_2) k_t(V_1,V_2)$$

with the notations of Proposition 4.7.

Finally, assume that we are matching two shapes X and Y represented by two unoriented d-dimensional meshed shapes. Then, the varifolds X and Y write

$$X = \sum_{i=1}^{m_X} v_i \delta_{x_i,V_i} \qquad \text{and} \qquad Y = \sum_{i=1}^{m_Y} w_i \delta_{y_i,W_i}$$

and the attachment distance between X and Y writes

$$\begin{aligned} A(X,Y) &= \|\mu_X - \mu_Y\|^2_{W'} = \|\mu_X\|^2_{W'} - 2\left\langle \mu_X \,\middle|\, \mu_Y \right\rangle_{W'} + \|\mu_Y\|^2_{W'} \\ &= \sum_{i,j=1}^{m_X} v_i v_j k_e(x_i,x_j) k_t(V_i,V_j) \; - \; 2\sum_{i=1}^{m_X}\sum_{j=1}^{m_Y} v_i w_j k_e(x_i,y_j) k_t(V_i,W_j) \\ &\qquad\qquad + \sum_{i,j=1}^{m_Y} w_i w_j k_e(y_i,y_j) k_t(W_i,W_j)\,. \end{aligned}$$

Charon and Trouvé (2013) provide a practical and low-computational-cost way to compute $k_t(V,W)$ through the notion of principal angles, which we do not detail in this dissertation.

At the very last point, Figure 4.3 illustrates the very-well behavior of varifolds regarding the orientation. Unfortunately, as a result, varifolds are sensitive to noise while currents were not.

– Chapter V –

Application to Chemotherapy Monitoring: Piecewise-Logistic Curve and Piecewise-Geodesic Shape Models

Ainsi le modèle proposé au chapitre 3 est-il à même de quantifier le rythme de progression d'un processus continu. Nous appliquons ce modèle au suivi de chimiothérapie et plus particulièrement au suivi du cancer métastatique du rein. En effet, dans ce contexte, la compréhension du rythme de progression du cancer est au cœur de la prise en charge médicale.

La première application concerne le suivi de scores RECIST pour response evaluation criteria in solid tumors en anglais. Ces scores étant des données scalaires, on réalise une instantiation du modèle générique pour des données réelles bornées en se plaçant sur le segment $[0, 1]$ munit de la métrique logistique. Ce modèle a été élaboré en collaboration avec des oncologues et radiologues de l'Hôpital Européen Georges Pompidou (HEGP). Des expériences numériques sur données réelles et synthétiques en valident la pertinence.

La seconde application porte sur le suivi de formes anatomiques 3D, toujours pour l'évaluation de la réponse tumorale. Ce modèle repose sur la notion de grandes déformations que nous avons discutée en introduction et s'applique aussi bien aux courants (Vaillant et Glaunès, 2005) qu'aux varifolds (Charon et Trouvé, 2013), qui sont des espaces de forme standards pour l'analyse de formes anatomiques et dont on a rappelé les fondements mathématiques au Chapitre 4. On propose également des expériences numériques sur données synthétiques.

Contents

UNNDERSTANDING the global disease progression is the key of chemotherapy monitoring. Indeed, physicians have to choose the best possible treatment and sequence of molecules for each of their patients, in the shortest possible time. Here, we propose two instantiations of the generic piecewise geodesic model, both in view of chemotherapy monitoring: the piecewise logistic curve model and the piecewise geodesic shape model.

We recall that patients are treated and so the evolution of the tumoral growth will fluctuate. Therefore, reaction-diffusion based tumor growth models do not apply in this context. Moreover, the two proposed models allow to bring a representative of the whole population comportment out for any kind of input data: scores, images, shapes, etc.

I. The Piecewise-Logistic Curve Model: Chemotherapy Monitoring through RECIST Score

In this section, we explicit the generic model with logistic geodesics and $M =]0, 1[$. This is motivated by the study of the RECIST score monitoring, which leads to consider one-dimension manifold, with one rupture point. As this explicit model is designed in view of our target application, we first give a short description of RECIST score.

I – 1. The RECIST Score

Patients suffering from the metastatic kidney cancer, take a drug each day and regularly have to check their tumor evolution. Indeed, during the past few years, the way renal metastatic cancer are monitoring was profoundly changed: a new class of anti-angiogenic therapies targeting the tumor vessels instead of the tumor cells has emerged and drastically improved survival by a factor of three (Escudier et al., 2016). These new drugs, however, do not cure the cancer, and only succeed in delaying the tumor growth, requiring the use of successive therapies which have to be continued or interrupted at the appropriate moment according to the patient's response. So, the new medicine process has also created a new scientific challenge: how to choose the most efficient drug therapy given he first monitoring times of the response profile of a given patient.

The RECIST (Response Evaluation Criteria In Solid Tumors) score (Therasse et al., 2000) is a set of published rules that measures the tumoral growth. Physicians select at most five lesions, with a sufficient diameter, and sum the longest diameter for all target lesions. This leads them to determine if the tumors in cancer patients respond (completely or partially), stabilize or progress during treatment.

The response to a given treatment has generally two distinct phases: first, tumor's size reduces; then, the tumor grows again. So, we have to build a model which allow to us to catch this behaviors. Moreover, a practical question is to quantify the correlation between both phases and to determine as accurately as possible the individual rupture times t_R^i which are related to an escape of the patient's response to treatment.

I – 2. The Piecewise-Logistic Curve Model

Our observations consist of patient's RECIST score over time, *i.e.* of sequences of bounded one-dimension measures. As explained above, we could make out two phases in the evolution of RECIST scores: a decreasing and a growing one. So, we set $m = 2$ and $d = 1$, which leads us to propose a way to build models for chemotherapy monitoring. This model has been designed after discussions with oncologists of the HEGP.

a. The Group-Representative Trajectory.

Let M_0 be the open interval $]0, 1[$, equipped with the logistic metric: $\forall x \in M_0$, $\forall \xi, \zeta \in T_x M_0 \simeq \mathbb{R}$,

$$g_x(\xi, \zeta) \;=\; \xi\, \mathcal{G}(x)\, \zeta \qquad \text{with} \qquad \mathcal{G}(x) \;=\; \frac{1}{x^2(1-x^2)}\,.$$

Given three real numbers γ_0^{init}, γ_0^{escap} and γ_0^{fin} we define two affine functions by setting down

$$\phi_0^1 \colon\, x \;\mapsto\; \left(\gamma_0^{\text{init}} - \gamma_0^{\text{escap}}\right) x + \gamma_0^{\text{escap}} \qquad \text{and} \qquad \phi_0^2 \colon\, x \;\mapsto\; \left(\gamma_0^{\text{fin}} - \gamma_0^{\text{escap}}\right) x + \gamma_0^{\text{escap}}\,.$$

This allows us to map M_0 onto the intervals $]\gamma_0^{\text{escap}}, \gamma_0^{\text{init}}[$ and $]\gamma_0^{\text{escap}}, \gamma_0^{\text{fin}}[$ respectively: if $\bar{\gamma}_0$ refers to the sigmoid function, $\phi_0^1 \circ \bar{\gamma}_0$ will be a logistic curve, growing from γ_0^{escap} to γ_0^{init}. For compactness, we note t_R the single breaking-up time at the population level and t_R^i at the individual one. Moreover, due to our target application, we force the first logistic to be decreasing and the second one increasing (this condition may be easily relaxed for other framework).

Logistics are defined on open intervals, with asymptotic constraints. We want to formulate our constraints on some non-infinite time-points, as explained in paragraph I.1.b of Chapter 3. So, we set a positive threshold ν, close to zero, and demand the logistics γ_0^1 and γ_0^2 to be ν-near from their corresponding asymptotes. More precisely, we impose the trajectory γ_0 to be of the form $\gamma_0 = \gamma_0^1\, \mathbb{1}_{]-\infty, t_R]} + \gamma_0^2\, \mathbb{1}_{]t_R, +\infty[}$, where, for all times $t \in \mathbb{R}$,

$$\gamma_0^1 : t \;\mapsto\; \frac{\gamma_0^{\text{init}} + \gamma_0^{\text{escap}}\, \mathrm{e}^{(at+b)}}{1 + \mathrm{e}^{(at+b)}} \qquad ; \qquad \gamma_0^2 : t \;\mapsto\; \frac{\gamma_0^{\text{fin}} + \gamma_0^{\text{escap}}\, \mathrm{e}^{-(ct+d)}}{1 + \mathrm{e}^{-(ct+d)}}$$

and a, b, c and d are some positive numbers given by the following constraints

$$\gamma_0^1(t_0) \;=\; \gamma_0^{\text{init}} - \nu\,, \quad \gamma_0^1(t_R) \;=\; \gamma_0^2(t_R) \;=\; \gamma_0^{\text{escap}} + \nu \quad \text{and} \quad \gamma_0^2(t_1) \;=\; \gamma_0^{\text{fin}} - \nu\,.$$

In order the previous logistics to be well-defined, we also have to enforce $\gamma_0^{\text{escap}} + 2\nu \leqslant \gamma_0^{\text{init}}$ and $\gamma_0^{\text{escap}} + 2\nu \leqslant \gamma_0^{\text{fin}}$. Thus, $p_{\text{pop}} = 5$ and

$$\mathcal{Z}_{\text{pop}} \;=\; \left\{ \left(\gamma_0^{\text{init}}, \gamma_0^{\text{escap}}, \gamma_0^{\text{fin}}, t_R, t_1\right) \in \mathbb{R}^5 \;\middle|\; \gamma_0^{\text{escap}} + 2\nu \leqslant \gamma_0^{\text{init}} \;\wedge\; \gamma_0^{\text{escap}} + 2\nu \leqslant \gamma_0^{\text{fin}} \right\}.$$

In our context, the initial time of the process is known: it is the beginning of the treatment. So, we assume that the representative initial time t_0 is equal to zero.

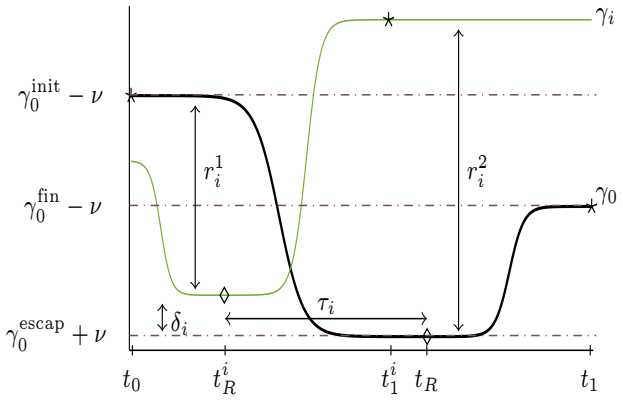


Figure 5.1 *– From representative to individual trajectory.*
Illustration of the non-standard constraints for γ_0 and the transition from the representative trajectory to an individual one: the trajectory γ_i is subject to a temporal and a spatial warp. In other "words", $\gamma_i \;=\; \phi_i^1 \circ \gamma_0^1 \circ \psi_i^1 \, \mathbb{1}_{]-\infty, t_R^i]} + \phi_i^2 \circ \gamma_0^2 \circ \psi_i^2 \, \mathbb{1}_{]t_R^i, +\infty[}$.

b. Individual Trajectories.

For each $i \in [\![1, n]\!]$, given $(\alpha_i^1, \alpha_i^2, \tau_i) \in \mathbb{R}_+^2 \times \mathbb{R}$, the time-warps (*cf.* Chap. 3, I.2.a) write

$$\psi_i^1(t) \;=\; \alpha_i^1(t - t_0 - \tau_i) + t_0 \qquad \text{and} \qquad \psi_i^2(t) \;=\; \alpha_i^2(t - t_R - \tau_i^2) + t_R \,,$$

where $\tau_i^2 \;=\; \tau_i + \left(\frac{1-\alpha_i^1}{\alpha_i^1}\right)(t_R - t_0)$.

In the same way as the time-warp, the diffeomorphisms ϕ_i^1 and ϕ_i^2 (*cf.* Chap. 3, I.2.b) are chosen to allow different amplitudes and rupture values: for each $i \in [\![1, n]\!]$, given the two scaling factors r_i^1 and r_i^2 and the space-shift δ_i, we define

$$\forall \ell \in \{1, 2\}, \qquad \phi_i^\ell(x) \;=\; r_i^\ell \left(x - \gamma_0(t_R)\right) + \gamma_0(t_R) + \delta_i \,.$$

Other choices are conceivable but in the context of our target applications, this one is the most appropriate: as we want to study the correlation between growth and decrease phase, none of the portions of the curves have to be favoured and affine functions allow us

to put the same weight on the whole curves. Mathematically, any regular and injective function defined on $]\gamma_0^{\text{escap}}, \gamma_0^{\text{init}}[$ (respectively $]\gamma_0^{\text{escap}}, \gamma_0^{\text{fin}}[$) works.

To sum up, each individual trajectory γ_i depends on the representative curve γ_0 through $\boldsymbol{z}_{\text{pop}} = \left(\gamma_0^{\text{init}}, \gamma_0^{\text{escap}}, \gamma_0^{\text{fin}}, t_R, t_1\right)$ fixed and $\boldsymbol{z}_i = (\alpha_i^1, \alpha_i^2, \tau_i, r_i^1, r_i^2, \delta_i)$ random effects. This leads to a non-linear mixed effects model. More precisely, we set for all individuals $i \in [\![1, n]\!]$

$$\forall \ell \in \{1,2\}, \quad \gamma_i^\ell \;=\; \phi_i^\ell \circ \gamma_0^\ell \circ \psi_i^\ell \qquad \text{and} \qquad t_R^i \;=\; t_0 + \tau_i + \frac{t_R - t_0}{\alpha_i^1}\,,$$

which leads us to write for all measurements $j \in [\![1, k_i]\!]$,

$$y_{i,j} \;=\; \gamma_i^1(t_{i,j})\,\mathbb{1}_{]-\infty, t_R^i]}(t_{i,j}) \;+\; \gamma_i^2(t_{i,j})\,\mathbb{1}_{]t_R^i, +\infty[}(t_{i,j}) \;+\; \varepsilon_{i,j}\,.$$

Figures 5.1 and 5.2 provide illustrations of the model. On each figure, the bold black curve represents the characteristic trajectory γ_0 and the color curves several individual trajectories.

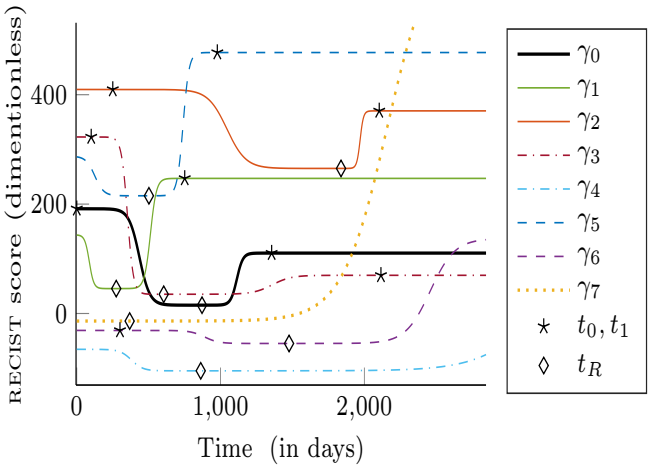


Figure 5.2 – *Diversity of individual trajectories.*
A typical representative trajectory in bold and several individual ones, for different vectors $\boldsymbol{z}_i$. The rupture times are represented by diamonds and the initial/final times by stars.

We proceed as in the paragraph I.3.a and set $\alpha_i^\ell = \mathrm{e}^{\xi_i^\ell}$ for $\ell \in \{1,2\}$. Likewise, the scaling parameters r_i^ℓ have to be positive and equal to one on average while the space shifts δ_i can be of any signs and must be zero on average. So, we set $r_i^\ell = \mathrm{e}^{\rho_i^\ell}$ for $\ell \in \{1,2\}$ leading to $\boldsymbol{z}_i = (\xi_i^1, \xi_i^2, \tau_i, \rho_i^1, \rho_i^2, \delta_i)$. In particular, $p_{\text{ind}} = 6$ and we assume that there exists $\Sigma \in \mathscr{S}^+_{p_{\text{ind}}}(\mathbb{R})$ such that $\boldsymbol{z}_i \sim \mathcal{N}(0, \Sigma)$ for all $i \in [\![1, n]\!]$. In view of our target application, this assumption is really important: usually, the random effects are studied independently. Here, we are interested in correlations between the two phases of patient's response to treatment in order to answer question like: does a fast response

induce a fast reprogression after the rupture time, which would mean that a fast response would decrease the susceptibility to this drug?

I – 3. Theoretical Analysis of the Piecewise-Logistic Curve Model

Theorem 3.1 applies as is leading to a well-defined MAP estimator for the piecewise logistic model. Moreover, at the risk of assuming some restriction concerning the distribution of our observations, the piecewise logistic model is consistent.

More precisely, let Θ^{PL} be the space of the admissible parameters for the Piecewise-Logistic model, *i.e.*

$$\Theta^{\text{PL}} \;=\; \{(\overline{\gamma_0^{\text{init}}}, \overline{\gamma_0^{\text{escap}}}, \overline{\gamma_0^{\text{fin}}}, \overline{t_R}, \overline{t_1}, \Sigma, \sigma) \in \mathbb{R}^{p_{\text{pop}}} \times \mathscr{S}^+_{p_{\text{ind}}}(\mathbb{R}) \times \mathbb{R}^+ \}\,.$$

We define

$$\Theta^{\omega,\text{PL}} \;=\; \{\, \theta \in \Theta^{\text{PL}} \mid \|(\overline{\gamma_0^{\text{init}}}, \overline{\gamma_0^{\text{escap}}}, \overline{\gamma_0^{\text{fin}}}, \overline{t_R}, \overline{t_1})\| \leqslant \omega \,\}$$

the space of the parameters associated to bounded on average fixed effects, for the Piecewise-Logistic model and, as in the generic framework, the space

$$\Theta^{\omega,\text{PL}}_* \;=\; \{\theta \in \Theta^{\omega,\text{PL}} \mid \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)} \Big[\log q(\boldsymbol{y}^\ell|\theta)\Big] = \mathbb{E}^*(\omega)\}\,,$$

where $\mathbb{E}^*(\omega) \;=\; \sup_{\theta\in\Theta^{\omega,\text{PL}}} \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)} \Big[\log q(\boldsymbol{y}^\ell|\theta)\Big]$.

Theorem 5.1 (*Consistency of the MAP, Piecewise-Logistic Curve Model*)

Assume that

(H 1) The number of acquisition is bigger than the one of latent variables: There exists $\ell \in [\![1, n]\!]$ such that $p^\ell < k^\ell$, where $k^\ell = \sum_{i=1}^{\ell} k_i$ and $p^\ell = p_{\text{pop}} + \ell\, p_{\text{ind}}$;

(H 2) The times of registration $t_i = (t_{i,j})_{j\in[\![1,k_i]\!]}$ are independent and identically distributed;

(H 3) The density $P(\mathrm{d}\boldsymbol{y}^\ell)$ is continuous with polynomial tail decay of degree bigger $p^\ell + 1$ apart from a compact subset K of $\mathbb{R}^{k^\ell}$;

Then, the Piecewise-Logistic model satisfies the hypothesis of Theorem 3.2. In particular, if $(\hat{\theta}_n)_{n\in\mathbb{N}}$ denote any MAP estimator, $\Theta^{\omega,\text{PL}}_* \neq \varnothing$ and for any $\varepsilon \in \mathbb{R}^*_+$,

$$\lim_{n\to\infty} \mathbb{P}\left[\delta(\hat{\theta}_n, \Theta^{\omega,\text{PL}}_*) \geqslant \varepsilon\right] \;=\; 0\,,$$

where δ in any metric compatible with the topology on $\Theta^{\omega,\text{PL}}$.

Proof: We demonstrate that, for all $i \in [\![1, n]\!]$, the variables $\left(\gamma_0^{\text{init}},\, \gamma_0^{\text{escap}},\, \gamma_0^{\text{fin}},\, \rho_i^1,\, \rho_i^2,\, \delta_i\right)$ are regular, that the variables $(t_R,\, t_1,\, \xi_i^1,\, \xi_i^2,\, \tau_i)$ are critical, and that $(\rho_i^1,\, \rho_i^2)$ are regular in the neighbourhood of $+\infty$ and critical near $-\infty$. See the remark after Theorem 3.2.

(H 4) Let $i \in [\![1, n]\!]$. By definition of $\vec{\boldsymbol{\gamma}}_i$,

$$\|\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)\|_\infty \;=\; \max \left\{ \begin{array}{l} |\, \gamma_0^{\text{escap}} + \nu + \delta_i \,| \\ |\, \gamma_0^{\text{escap}} + \nu + \delta_i + \mathrm{e}^{\rho_i^1}(\gamma_0^{\text{init}} - \gamma_0^{\text{escap}} - 2\nu) \,| \\ |\, \gamma_0^{\text{escap}} + \nu + \delta_i + \mathrm{e}^{\rho_i^2}(\gamma_0^{\text{fin}} - \gamma_0^{\text{escap}} - 2\nu) \,| \end{array} \right\} .$$

And we can check that for γ_0^{init}, γ_0^{escap}, γ_0^{fin}, ρ_i^1, ρ_i^2 and δ_i and that for ρ_i^1 and ρ_i^2 as soon as $|\rho_i^1|, |\rho_i^2| \geqslant 1$ there exists two functions a_i and b_i as in [Theorem 3.2 **(H 4)**].

(H 5) Let $i \in [\![1, n]\!]$ and $j \in [\![1, k_i]\!]$. By definition of $\vec{\boldsymbol{\gamma}}_i$,

$$\lim_{t_R \to +\infty} \vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)_j \;=\; \Big[\, \mathrm{e}^{\rho_i^1} \left(\gamma_0^{\text{init}} - \gamma_0^{\text{escap}} - 2\nu\right) + \gamma_0^{\text{escap}} + \nu + \delta_i \,\Big]\, \mathbb{1}_{[t_0, +\infty[}(t_{i,j})\,,$$

where $\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i)_j$ denotes the j^{th} coordinate of the vector $\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\text{pop}}, \boldsymbol{z}_i) \in \mathbb{R}^{k_i}$. However, by construction, $\gamma_0^{\text{init}} - \gamma_0^{\text{escap}}$ and γ_0^{escap} follow a normal distribution so

$$\mathscr{L}_{k_i}\Big(\Big\{\, y_{i,j} = \mathrm{e}^{\rho_i^1} \left(\gamma_0^{\text{init}} - \gamma_0^{\text{escap}} - 2\nu\right) + \gamma_0^{\text{escap}} + \nu + \delta_i \,\Big\}\Big) = 0\,.$$

Likewise for $t_R \to -\infty$. The same argument holds when t_1, ξ_i^1, ξ_i^2 or τ_i become infinite and when ρ_i^1 or ρ_i^2 go to $-\infty$.

□

II. The Piecewise-Geodesic Shape Model:
Chemotherapy Monitoring through Anatomical Shapes

A more precise way to follow-up cancer is to focus on the evolution of the tumors as anatomical shape. For this purpose, the tumors is segmented and transformed into a surface mesh or a curvilinear paths depending on the type of data we consider, typically depending on the number of layers in scanner we have access to. Let $d \in \{2, 3\}$ be the dimension of the ambient space.

Shape trajectories within the geodesic-framework developed by Schiratti et al. has already been addressed (Bône et al., 2018). We explain here quickly how to adapt this model to the piecewise-geodesic framework. In particular, we use consistent notations and admit the notion of exp-parallelism.

II – 1. The Piecewise-Geodesic Shape Model

In the same way as the piecewise-logistic curve model, we are targeting chemotherapy monitoring with two distinct phases in the evolution of the tumoral growth and thus we set $m = 2$. We keep the notation t_R for the single breaking-up time at the population level and t_R^i at the individual one. We consider only two geodesic components, so that an easy way to enforce the representative path to be continuous is to define the first component in the past and the second one in the future, from the rupture time t_R. Thus, provided that we follow forward the first component, we can use exactly the same construction that the one introduced by Bône et al. (2018). As a consequence, the following is applicable either for currents (Vaillant and Glaunès, 2005) or varifolds (Charon and Trouvé, 2013), allowing to consider shapes without any point correspondence.

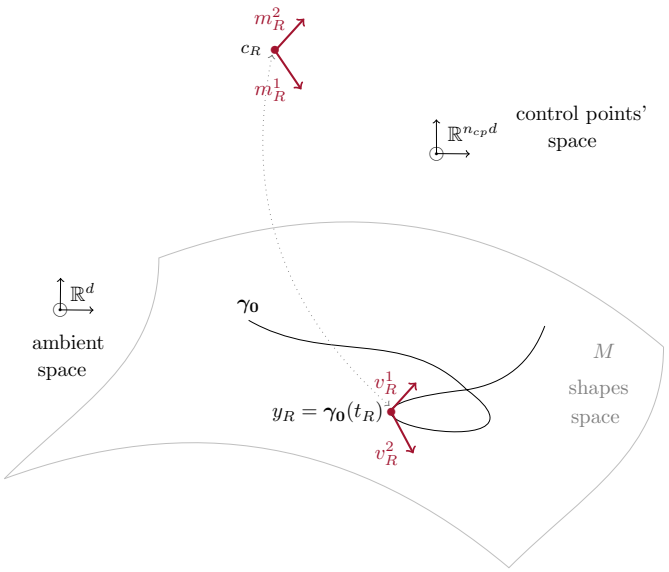


***Figure 5.3** – Construction of the group-representative trajectory.*
Let t_R be the rupture time and $y_R \in M$ the rupture shape, *i.e.* the shape of the representative path at the rupture time. We define the path γ_0 as the concatenation of the two geodesics starting at the rupture time t_R and the point y_R, in the directions associated to m_R^1 the backward and m_R^2 the forward momenta respectively and where the first one is followed backward.

a. The Group-Representative Trajectory.

Let $y_R \in M \subset \mathbb{R}^d$ be the rupture shape, *i.e.* the representative shape of the population at the rupture time t_R, and likewise $c_R \in \mathbb{R}^{n_{cp}d}$ be a set of n_{cp} rupture control points. Let $m_R^1 \in \mathbb{R}^{n_{cp}d}$ and $m_R^2 \in \mathbb{R}^{n_{cp}d}$ be respectively the backward and the forward momenta at the rupture time. We define the representative path by:

$$\gamma_0 \colon t \;\mapsto\; \mathcal{E}xp_{c_R,t_R,-t}(m_R^1) \circ y_R \; \mathbb{1}_{]-\infty,t_R]}(t) \;+\; \mathcal{E}xp_{c_R,t_R,t}(m_R^2) \circ y_R \; \mathbb{1}_{[t_R,+\infty[}(t)\,,$$

where $t \mapsto \mathcal{E}xp_{c_R,t_R,t}(m_R)$ denotes the exponential operator associated to the manifold of diffeomorphisms underlying the shape space (See Beg et al. (2005); Bône et al. (2018); Miller et al. (2006) for details about the construction of the shape space). In particular, the resulting backward and forward velocity vectors at the rupture time are respectively define by $v_R^1 = c_R \cdot m_R^1 = \sum_{q=1}^{n_{cp}} c_{R,q}\, m_{R,q}^1$ and $v_R^2 = c_R \cdot m_R^2$. Figure 5.3 sums up this construction.

b. Individual Trajectories.

At the individual level, as the initial representative time is not explicitly defined in this framework, we slightly modify the first time component reparametrization leading to

$$\psi_i^1(t) \;=\; \mathrm{e}^{\xi_i^1}(t - t_R - \tau_i) + t_R \qquad \text{and} \qquad \psi_i^2(t) \;=\; \mathrm{e}^{\xi_i^2}(t - t_R - \tau_i) + t_R\,.$$

The individual rupture times are then given by $t_R^i = t_R + \tau_i$ and we check that $\psi_i^1(t_R^i) = \psi_i^2(t_R^i) = t_R$.

The diffeomorphic component deformations consist of exp-parallelism of the representative path (Schiratti et al., 2015). Given a vector w, to define the exp-parallel of a curve γ in the direction of w, we first transport the vector w along the curve γ and then compute the flow given by the transported vector. We note $P_t \colon \mathbb{R}^{n_{cp}d} \to \mathbb{R}^{n_{cp}d}$ the parallel transport operator, which transport any vector $w \in \mathbb{R}^{n_{cp}d}$ along the curve γ_0 from $\gamma_0(t_R)$ to $\gamma_0(t)$ and we set: $\eta_w \colon t \mapsto \mathcal{E}xp_{c(t),0,1}\left(P_t(w)\right)$, where $c(t)$ is the set of control points for γ_0, at the time t:

$$c(t) \;=\; \mathcal{E}xp_{c_R,t_R,-t}(m_R^1) \circ c_R \; \mathbb{1}_{]-\infty,t_R]}(t) \;+\; \mathcal{E}xp_{c_R,t_R,t}(m_R^2) \circ c_R \; \mathbb{1}_{[t_R,+\infty[}(t)\,.$$

Thus, given a space shift momenta w_i for all individuals, the space deformation of the curve γ_0 is given by $t \mapsto \eta_{w_i}(t) \circ y_R$.

Last, for all individuals $i \in [\![1, n]\!]$, we define a subject-specific trajectory by setting:

$$\gamma_i \colon t \;\mapsto\; \eta_{w_i}(\psi_i^1(t)) \circ y_R \; \mathbb{1}_{]-\infty,t_R^i]}(t) \;+\; \eta_{w_i}(\psi_i^2(t)) \circ y_R \; \mathbb{1}_{[t_R^i,+\infty[}(t)\,.$$

c. Space-shift momenta and identifiability

Following Bône et al. (2018) and in the spirit of Independent Component Analysis (Hyvärinen et al., 2004), we assume that each space-shift momenta w_i is a linear combination of n_s sources $s_i \in \mathbb{R}^{n_s}$, *i.e.* that $w_i = A_{m_R^\perp} s_i$, where $A_{m_R^\perp} \in \mathcal{M}_{n_{cp}d,n_s}(\mathbb{R})$ calls

modulation matrix. As argued in Bône et al. (2018) and Schiratti et al. (2015), we have to ensure the orthogonality between m_R and w_i in order to ensure the identifiability of the model. This orthogonality condition prevents a confusion between the space shifts and the acceleration factors and can be achieved through projection techniques that we do not detail here.

To sum up, the population random effects are given by $z_{\text{pop}} = (y_R, c_R, m_R^1, m_R^2, t_R, A_{m_R^\perp})$ and the individual ones by $z_i = (\xi_i^1, \xi_i^2, \tau_i, s_i)$. To place ourselves in the hierachical framework detailed at the paragraph I.3.b, we assume that there exists $\Sigma \in \mathscr{S}_3^+(\mathbb{R})$ such that $z_i \sim \mathcal{N}(0, \Sigma) \otimes \mathcal{N}(0, 1)$ and that there exists small fixed variances such that the population latent variable follow a tight Gaussian distribution.

II – 2. Theoretical Analysis of the Piecewise-Geodesic Shape Model

As for the piecewise-logistic curve model, Theorem 3.1 applies and the MAP estimator for the piecewise-geodesic shape model is well-defined. We therefore focus on the consistency of this model.

Like previously, we define the space of admissible parameters associated to bounded on average fixed effects:

$$\Theta^{\omega,\text{PS}} \;=\; \{\, \theta \in \Theta^{\text{PS}} \;\mid\; \|(\overline{\gamma_0^{\text{init}}}, \overline{\gamma_0^{\text{escap}}}, \overline{\gamma_0^{\text{fin}}}, \overline{t_R}, \overline{t_1})\| \leqslant \omega \,\}$$

for all $\omega \in \mathbb{R}$, where

$$\Theta^{\text{PS}} \;=\; \{\, (\overline{y_R}, \overline{c_R}, \overline{m_R^1}, \overline{m_R^2}, \overline{t_R}, \overline{A_{m_R^\perp}}\Sigma, \sigma) \;\in\; \mathbb{R}^{p_{\text{pop}}} \times \mathscr{S}^+_{p_{\text{ind}}}(\mathbb{R}) \times \mathbb{R}^+ \,\}\,.$$

As in the generic and piecewise-logistic framework, we also define the space

$$\Theta_*^{\omega,\text{PS}} \;=\; \{\, \theta \in \Theta^{\omega,\text{PS}} \;\mid\; \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\Big[\log q(\boldsymbol{y}^\ell|\theta)\Big] = \mathbb{E}^*(\omega) \,\}\,,$$

where $\mathbb{E}^*(\omega) \;=\; \sup_{\theta \in \Theta^{\omega,\text{PL}}} \mathbb{E}_{P(\mathrm{d}\boldsymbol{y}^\ell)}\Big[\log q(\boldsymbol{y}^\ell|\theta)\Big]$.

Theorem 5.2 (*Consistency of the MAP, Piecewise-Geodesic Shape Model*)

Assume that

(H 1) The number of observations is bigger than the one of latent variables: There exists $\ell \in [\![1, n]\!]$ such that $p^\ell < k^\ell$, where $k^\ell = \sum_{i=1}^{\ell} k_i$ and $p^\ell = p_{\text{pop}} + \ell\, p_{\text{ind}}$;

(H 2) The times of acquisition $t_i = (t_{i,j})_{j \in [\![1,k_i]\!]}$ are independent and identically distributed;

(H 3) The density $P(\mathrm{d}\boldsymbol{y}^\ell)$ is continuous with polynomial tail decay of degree bigger $p^\ell + 1$ apart from a compact subset K of $\mathbb{R}^{k^\ell}$;

(H 4) For all individuals $i \in [\![1, n]\!]$, the function $t \mapsto \|\gamma_i(t)\|$ grows super-linearly or $t \mapsto \gamma_i(t)$ converges uniformly w.r.t the variable y_R toward a function $t \mapsto \gamma_i^*(t)$;

(H 5) For all individuals $i \in [\![1, n]\!]$, the variables c_R, m_R^1, m_R^2, $A_{m_R^\perp}$ and s_i are either regular or critical, in the sens of Theorem 3.2.

Then, the piecewise-geodesic shape model satisfies the hypothesis of Theorem 3.2. In particular, if $(\hat{\theta}_n)_{n\in\mathbb{N}}$ denote any MAP estimator, $\Theta_*^{\omega,\mathrm{PS}} \neq \varnothing$ and for any $\varepsilon \in \mathbb{R}_+^*$,

$$\lim_{n\to\infty} \mathbb{P}\left[\delta(\hat{\theta}_n, \Theta_*^{\omega,\mathrm{PS}}) \geqslant \varepsilon\right] \;=\; 0\,,$$

where δ in any metric compatible with the topology on $\Theta^{\omega,\mathrm{PS}}$.

Proof: Let's demonstrate that the variables y_R, t_R, ξ_i^1 , ξ_i^2 and τ_i are either regular of critical for all individuals $i \in [\![1, n]\!]$.

1. Let $i \in [\![1, n]\!]$ and $(\boldsymbol{z}_{\mathrm{pop}}, \boldsymbol{z}_i) \in \mathcal{Z}_{\mathrm{pop}} \times \mathcal{Z}_i$. By continuity of the parallel transport and the trajectory γ_0, there exists $a \in \mathbb{R}^+$ such that

$$\|\vec{\boldsymbol{\gamma}}_i(\boldsymbol{z}_{\mathrm{pop}}, \boldsymbol{z}_i)\|_\infty \;\geqslant\; a\,\|y_R = \gamma_0(t_R)\|$$

   and so, y_R is regular.

2. By continuity of the geodesic flow, for all times $t \in \mathbb{R}$, $\gamma_i(t)$ depends continuously of $\gamma_i(t_R)$. Thus, for all $t \in \mathbb{R}$, by continuity of the parallel transport, $\gamma_i(t)$ depends continuously of y_R. If γ_i^* exists, since the convergence of γ_i toward γ_i^* is uniform w.r.t the variable y_R, γ_i^* is also continuous w.r.t $y_R = \gamma_0(t_R) \sim \mathcal{N}\left(\overline{y_R}, *\right)$ and so is a continuous distribution.
   For all $i \in [\![1, n]\!]$, all $\ell \in \{1, 2\}$ and all $t \in \mathbb{R}$, we have $\lim_{\xi_i^\ell \to -\infty} \psi_i^\ell(t) = t_R$. Therefore, $\lim_{\xi_i^\ell \to -\infty} \gamma_i^\ell(t) = \gamma_i(t_R)$ which is a continuous distribution. So, for all $j \in [\![1, k_i]\!]$, $\mathscr{L}_{k_i}\left(\{y_{i,j} = \gamma_i(t_R)\}\right) = 0$.
   Last, since $\lim \psi_i^\ell(t) = \pm\infty$ when $|t_R|$, $|\tau_i|$ or ξ_i^ℓ converge toward $+\infty$, we get the result with Assumption **(H 4)**.

□

Proving **(H 5)** is a very interesting issue but outside of the scope of this section. Our conjecture is that a positive and restricted curvature for the shape space M will guarantee that, for all individuals $i \in [\![1, n]\!]$, c_R, m_R^1, m_R^2, $A_{m_R^\perp}$ and s_i are regular. Indeed, we guess that in geodesic shooting, sufficient initial momenta will enforce the trajectory to "go away", provided that the underlying manifold is "kind" enough. Note that the computation of the curvature for both currents and varifolds is still an open problem.

III. Experimental Results

Experimentations are performed for both models introduced above: the piecewise logistic curve model and the piecewise geodesic shape model. In order to validate our model and numerical scheme, we first run experiments on synthetic data for the piecewise logistic model. We then test our estimation algorithm on real data from the Hôpital Européen Georges Pompidou (HEGP – Georges Pompidou European Hospital). A medical paper is under progress to provide a more accurate interpretation of this results. Then, we run experiments on synthetic data for the piecewise geodesic shape model to confirm the performance of our model on more complicated data. Real data for this framework are being collected and preprocessed.

III – 1. Univariate Synthetic Data

We generate four types of data set, to put our algorithm in different situations. More precisely, we want to quantify its sensitivity to initialisation, sample size and noise.

a. Influence of the Initialization

The estimation is performed through the SAEM algorithm (Algorithm 1). This iterative algorithm is proven to converge toward a critical point of the observed likelihood. Therefore, as our model does no imply a convex likelihood, one may end up with a local *maximum* depending on the initialization point and the dynamic of our iterations. This choice of initialization appears crucial. In particular the choice of the initial mean population parameters $\overline{\boldsymbol{z}_{\text{pop}}}^{\,\text{init}}$ as illustrated bellow.

If our model were linear, the representative curve γ_0 would exactly be the one induced by the mean of the individual trajectories γ_i, *i.e.* the one where $\boldsymbol{z}_{\text{pop}} = \text{mean}_{i\in[\![1,n]\!]}\ \boldsymbol{z}_i$. Following this idea, we set in our experiments

$$\overline{\gamma_0^{\text{init}}}^{\,\text{init}} = \underset{i\in[\![1,n]\!]}{\text{mean}}\ y_{i,1} \quad ; \quad \overline{\gamma_0^{\text{escap}}}^{\,\text{init}} = \underset{i\in[\![1,n]\!]}{\text{mean}}\ \underset{j\in[\![1,k_i]\!]}{\min}\ y_{i,j} \quad ; \quad \overline{\gamma_0^{\text{fin}}}^{\,\text{init}} = \underset{i\in[\![1,n]\!]}{\text{mean}}\ y_{i,k_i}$$

$$\overline{t_R}^{\,\text{init}} = \frac{1}{2}\ \underset{i\in[\![1,n]\!]}{\text{mean}}\ t_{k_i} \qquad \text{and} \qquad \overline{t_1}^{\,\text{init}} = \underset{i\in[\![1,n]\!]}{\text{mean}}\ t_{k_i}\ .$$

Note that the choice of the initial covariance matrix $\Sigma^{\,\text{init}}$ and the residual noise $\sigma^{\,\text{init}}$ does not seem to be very influential. We just demand $\Sigma^{\,\text{init}}$ to be definite positive.

b. Influence of the Proposal Variances

The SAEM algorithm is very sensitive to the choice of the proposal variances in the sampling step. Thus, we have to carefully tune these variances in order the mean acceptance ratio to stay around the optimal rate – 24% as we are using a symmetric random

walk sampler. To decrease the influence of a bad calibration, we adapt the proposal variances over the iterations in the way of Roberts and Rosenthal (2007, 2009): every s^{th} batch of 50 iterations, we increase or decrease the logarithm of the proposal proposal variances by $\delta(s) = \min\left(0.001, \frac{1}{\sqrt{s}}\right)$ depending on whether the mean associated variable acceptance rate is bigger or smaller than the optimal one. Note that we have also tried to adapt the proposal variances as in Atchadé (2006) but the results we obtained were not satisfactory. Actually, it appears numerically that if we want the adaptive procedure to increase the efficiency of our algorithm, we must modify the proposal variance neither too often nor with a too big amplitude of change.

c. Construction of the Data Sets

For each type of data set, given the corresponding ground truth parameters θ^{true}, we generate three data sets of respective size 50, 100 and 250. Last, to put our algorithm on a more realistic situation, the synthetic individual times are non-periodically spaced and individual sizes vary between 12 and 18.

The first type – A – is said quasilinear in the sense that, for these data sets, the representative trajectory γ_0 is "close" to the mean trajectory described above. Hence, we put our algorithm in a favorable situation where the optimal representative trajectory is close to the initial one. The second type –A* – is a noisy version of A. The noise level is approximately 20% (against 2% for the non-noisy data set A).

On the contrary, the thirds type – B – is built in order to be "truly non-linear": the representative trajectory γ_0 is "far" from the curve built by $\overline{\boldsymbol{z}_{\text{pop}}}^{\text{init}}$. Likewise, the fourth type – B* – is a noisy version of B, with a 20 % noise level.

To measure this degree of non-linearity, we introduce the ratio $\Delta_{\mathscr{L}}(\overline{\boldsymbol{z}_{\text{pop}}})$ which is the relative error of $\overline{\boldsymbol{z}_{\text{pop}}}^{\text{init}}$:

$$\Delta_{\mathscr{L}}(\overline{\boldsymbol{z}_{\text{pop}}}) = \frac{\|\overline{\boldsymbol{z}_{\text{pop}}}^{\text{init}} - \overline{\boldsymbol{z}_{\text{pop}}}^{\text{true}}\|}{\|\overline{\boldsymbol{z}_{\text{pop}}}^{\text{true}}\|} .$$

Table 5.1 compiles this ratio for every data set, and for every parameter in $\overline{\boldsymbol{z}_{\text{pop}}}$. In particular, the initialization of $\overline{\gamma_0^{\text{escap}}}$ is in itself a challenge and very sensitive to the noise in the data set: even in the quasilinear case, $\overline{\gamma_0^{\text{escap}}}^{\text{init}}$ is quite far from $\overline{\gamma_0^{\text{escap}}}^{\text{true}}$.

d. Estimation of the Fixed Effects

Table 5.2 displays the relative errors for the estimated population parameters. In most case, these errors decrease with the size of the data set. More specific to our model, we observe that these errors are correlated to the subjective linearity of the model. With the exception of $\overline{\gamma_0^{\text{escap}}}$, the errors for estimating population parameters grow linearly with the non-linearity of the model. We suppose that the difference of scale between $\overline{\gamma_0^{\text{escap}}}$ and the others can, at least partly, explain this phenomena: $\overline{\gamma_0^{\text{escap}}}$ is about a few

Table 5.1 *– Degree of non-linearity.*
Relative errors (expressed as a percentage) for the initial population parameters $\overline{z_{\text{pop}}}^{\text{init}}$, according to the type of data set and the sample size n.

	n	$\Delta_{\mathscr{L}}(\overline{\gamma_0^{\text{init}}})$	$\Delta_{\mathscr{L}}(\overline{\gamma_0^{\text{escap}}})$	$\Delta_{\mathscr{L}}(\overline{\gamma_0^{\text{fin}}})$	$\Delta_{\mathscr{L}}(\overline{t_R})$	$\Delta_{\mathscr{L}}(\overline{t_1})$
A	50	7.08	17.01	5.94	1.97	1.98
	100	2.93	22.33	3.66	2.40	2.42
	250	2.16	24.06	2.12	3.52	3.54
A*	50	5.63	283.14	1.51	1.03	1.01
	100	3.38	259.25	0.07	4.75	4.76
	250	3.67	269.42	0.41	3.94	3.95
B	50	80.47	2.77	39.78	35.04	35.09
	100	88.17	4.39	51.83	36.14	36.19
	250	83.52	12.91	47.90	33.23	33.27
B*	50	59.25	201.98	33.46	28.85	28.89
	100	74.94	213.96	43.50	30.74	30.78
	250	79.14	229.40	47.30	34.39	34.44

tens of units ; $\overline{\gamma_0^{\text{init}}}$, $\overline{\gamma_0^{\text{fin}}}$ and $\overline{t_R}$ about a few hundreds and $\overline{t_1}$ about one thousand. Thus, a same absolute error will lead to markedly different relative error.

As Table 5.1 displays the relative error for for the initial population parameters $\overline{z_{\text{pop}}}^{\text{init}}$ and Table 5.2 the relative errors for the estimated population parameters $\overline{z_{\text{pop}}}^{\text{estim}}$, by comparing this two tables, we are able to quantify the contribution of the estimation-procedure in the knowledge of the population parameters. The first point to note is that this relative error generally decrease. Specifically, the population parameters are well-learned in quasilinear cases (data sets A and A*) and in particular in large data set ($n = 250$). Then, the algorithm we propose is not noise-sensitive: errors for non-noisy and noisy versions of a same type of data set are notably the same. And even better, for the non-linear data sets, the estimation is better performed in the noisy case than in the non-noisy one. It seems that the presence of noise helps the algorithm not to get stuck in potential well.

Hence, the degree of non-linearity in the data set seems to play a significant role in the estimation of the population parameters. To be certain that the poor estimation of $\overline{z_{\text{pop}}}$ when the ratio $\Delta_{\mathscr{L}}(\overline{z_{\text{pop}}})$ is too big is due to the non-linearity of the data set and not to a bad initialization, we have also performed estimations by assigning $\theta^{\text{init}} = \theta^{\text{true}}$. The results were better but not so significantly.

Last, note that the representative rupture time $\overline{t_R}$ is well-estimated, no matter the subjective linearity of the data set. In the view of chemotherapy monitoring, well-estimating the rupture time, which correspond to an escapement from the treatment, is

***Table 5.2** – Estimation of the fixed effects.*
Mean (standard deviation) relative errors (expressed as a percentage) over 50 runs, for the estimated parameters $\overline{z_{\text{pop}}}^{\text{estim}}$, according to the data set and the sample size n.

	n	$\overline{\gamma_0^{\text{init}}}$	$\overline{\gamma_0^{\text{escap}}}$	$\overline{\gamma_0^{\text{fin}}}$	$\overline{t_R}$	$\overline{t_1}$
A	50	6.03 (0.32)	10.25 (0.50)	3.69 (0.25)	**1.95 (0.13)**	2.43 (0.18)
	100	2.19 (0.17)	3.28 (0.22)	2.07 (0.18)	**1.69 (0.11)**	1.86 (0.17)
	250	1.30 (0.10)	1.96 (0.13)	1.53 (0.08)	**0.78 (0.06)**	1.67 (0.09)
A*	50	3.74 (0.26)	25.73 (1.64)	6.84 (0.40)	**3.32 (0.26)**	3.73 (0.26)
	100	2.35 (0.15)	12.20 (0.64)	1.35 (0.09)	**2.98 (0.22)**	2.29 (0.18)
	250	1.70 (0.12)	3.94 (0.29)	1.33 (0.09)	**1.36 (0.10)**	1.51 (0.10)
B	50	71.13 (1.33)	100.24 (8.09)	90.73 (2.54)	**7.78 (0.56)**	46.39 (1.32)
	100	58.73 (0.98)	58.88 (3.00)	84.99 (1.42)	**8.13 (0.57)**	42.06 (1.04)
	250	67.49 (0.47)	23.12 (1.54)	57.82 (0.74)	**6.01 (0.33)**	38.09 (0.36)
B*	50	41.61 (1.26)	29.86 (2.53)	46.38 (1.60)	**9.04 (0.58)**	29.90 (0.58)
	100	60.39 (0.81)	28.43 (2.06)	58.35 (1.07)	**8.11 (0.54)**	29.75 (0.50)
	250	55.89 (0.74)	15.56 (0.98)	59.90 (0.58)	**3.26 (0.25)**	39.28 (0.43)

very important.

e. Estimation of the Inter-Individual Variability

In the target of our application, the covariance matrix Σ gives a lot of information on the health status of a patient: pace and amplitude of tumor progression, individual rupture times, *etc.* Therefore, we have to pay special attention to the estimation of Σ.

Much as the representative trajectory is not always good-estimated, our algorithm always allows a well-understanding of the inter-individual variability. We present at Table 5.3 the Kullback-Leibler divergence from $\mathcal{N}(0, \Sigma^{\text{estim}})$ to $\mathcal{N}(0, \Sigma^{\text{true}})$, the relative error of the individual rupture times and the estimated residual noise. As for the estimation of the population parameters, errors decrease with the sample size n and are not significantly different between noisy and non-noisy versions of a same type of data set. Moreover, in that case, the errors seem to not rely on the subjective linearity of the data set. In the following, for the sake of brevity, the Kullback-Leibler divergence from $\mathcal{N}(0, \Sigma^{\text{estim}})$ to $\mathcal{N}(0, \Sigma^{\text{true}})$ will be referred to as the one from Σ^{estim} to Σ^{true}.

Moreover, the individual rupture times $\overline{t_R}$, the residual noise σ is always well-estimated.

***Table 5.3** – Variability and residual noise.*
Mean (standard deviation) of Kullback–Leibler divergences from Σ^{estim} to Σ^{true}, mean (standard deviation) relative errors (expressed as a percentage) for the individual rupture times ${t_R^i}^{\text{estim}}$ and mean estimated residual noise σ^{estim} according to the data set and the sample size n. All over 50 runs.

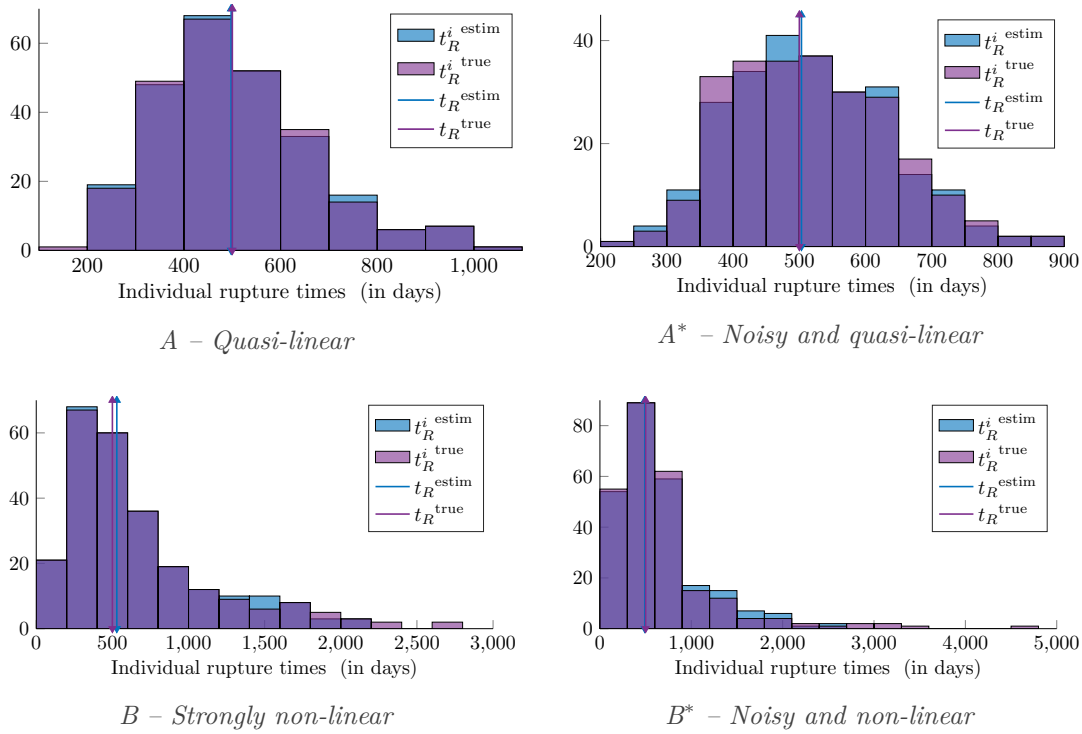
	n	Σ	t_R^i	σ
A	50	15.54 (5.17)	**0.49 (0.04)**	**2.03**
	100	8.45 (2.26)	**0.63 (0.06)**	**1.97**
	250	9.29 (3.13)	**0.57 (0.60)**	**2.06**
A*	50	16.52 (19.45)	**4.66 (0.45)**	**19.81**
	100	12.86 (4.26)	**3.85 (0.32)**	**19.03**
	250	6.72 (2.44)	**3.98 (0.32)**	**20.07**
B	50	16.53 (7.72)	**5.89 (3.45)**	**3.07**
	100	13.59 (5.42)	**4.44 (1.93)**	**2.14**
	250	22.24 (9.77)	**4.96 (1.93)**	**2.49**
B*	50	27.62 (17.71)	**14.32 (4.06)**	**19.93**
	100	23.98 (18.07)	**13.97 (3.71)**	**20.56**
	250	17.70 (5.35)	**11.57 (2.42)**	**21.38**

f. Reconstruction of the Individual Trajectories

Figure 5.4 illustrates the well-understanding of the variance within the population, including for the non-linear data set. Determining accurate individual rupture time t_R^i is all the most important as, in the aim of chemotherapy monitoring, these times are related to an escape of the patient's response to treatment.

An important point was to allow a lot of different individual behaviors. In our synthetic example, Figure 5.2 illustrates this variability. From a single representative trajectory (γ_0 in bold plain line), we can generate individuals who are cured at the end (dot-dashed lines: γ_3 and γ_4), some whose response to the treatment is bad (dashed lines: γ_5 and γ_6), some who only escape (no positive response to the treatments – dotted lines: γ_7). Likewise, we can generate "patients" with only positive responses or no response at all. The case of individual 4 is interesting in practice: the tumor still grows but so slowly that the growth is negligible, at least in the short-run.

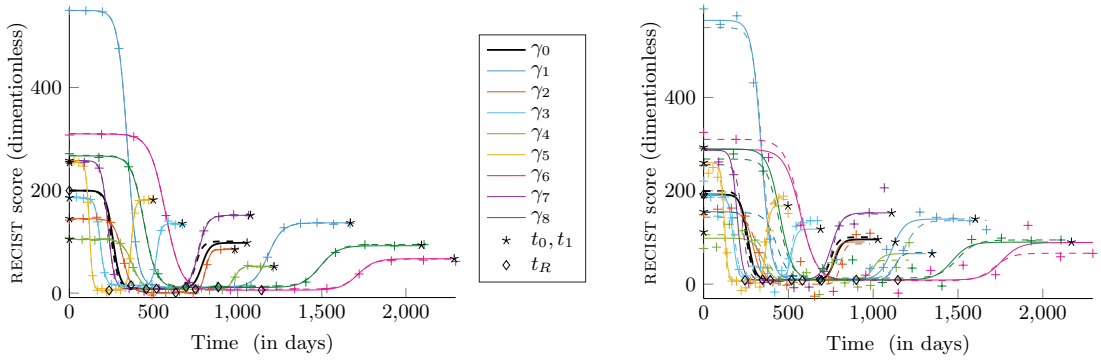
Figure 5.5 illustrates the qualitative performance of the estimation. We are notably able to understand various behaviors and fit subjects which are far from the characteristic path. Moreover, the noise seems to not reduce the quality of the estimation. We represent only five individuals but 250 subjects have been used to perform the estimation.



A – Quasi-linear *A* – Noisy and quasi-linear*

B – Strongly non-linear *B* – Noisy and non-linear*

Figure 5.4 *– Distribution of the individual rupture times t_R^i.*
Each subfigure compares the distribution of the (mean of the) estimated individual rupture times $t_R^{i\ \text{estim}}$ (in blue) and the distribution of the true individual rupture times $t_R^{i\ \text{true}}$ (in violet). In bold line, the estimated average rupture time $t_R{}^{\text{estim}}$ and the true average rupture time $t_R{}^{\text{true}}$ are relatively close to each other. $n = 250$.



A – Quasi-linear *A* – Noisy and quasi-linear*

Figure 5.5 *– Qualitative performance of the estimation and robustness to noise.*
On both figures, the estimated trajectories are in plain lines and the target curves in dashed lines. The (noisy) observations are represented by crosses. The representative path is in bold black line, the individuals in colour. $n = 250$.

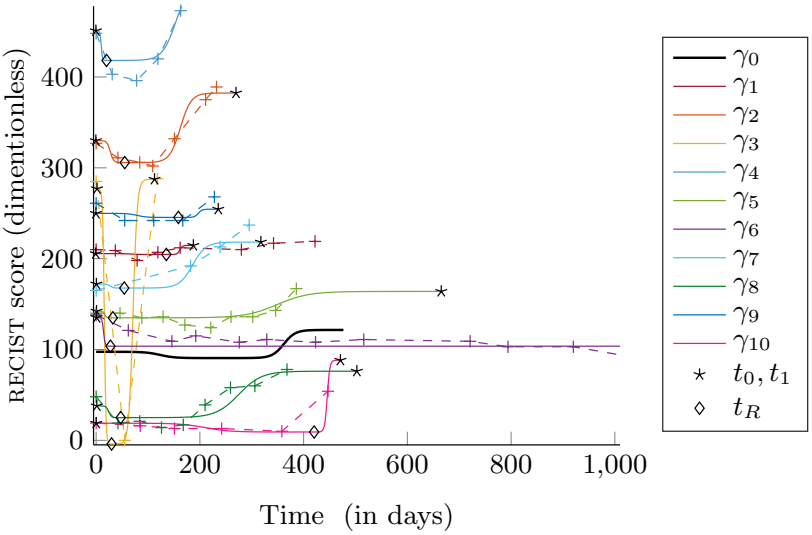


Figure 5.6 – *RECIST score of patients from the HEGP.*
We keep conventions of the previous figures: the representative path is in bold black line and the individuals in color. We represent only 10 patients among the 176.

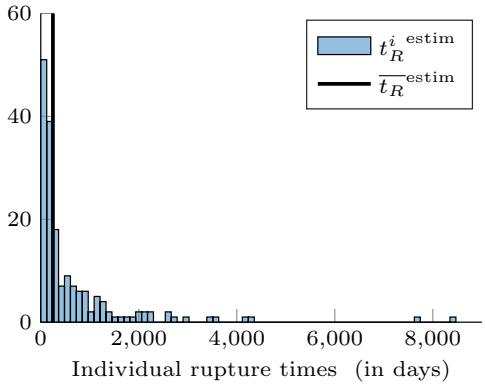
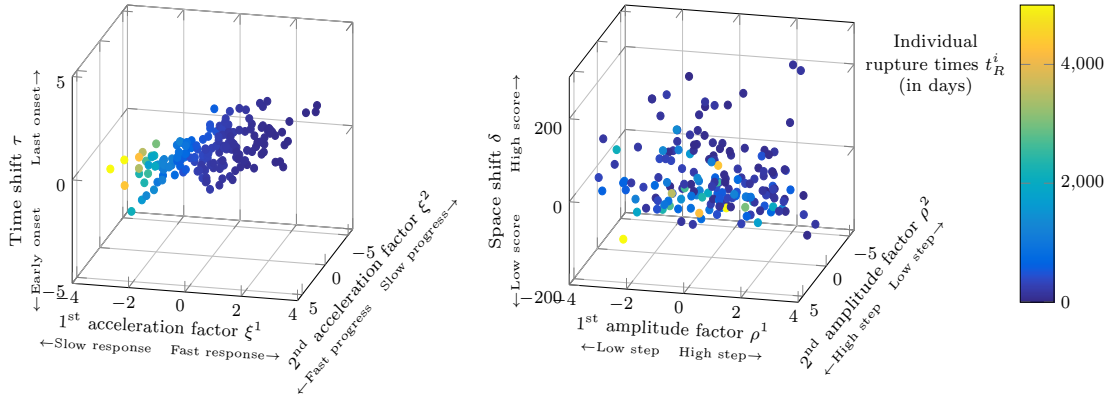


Figure 5.7 – *Distribution of the individual rupture times t_R^i.*
Histogram of the rupture times t_R^i for this run. In black bold line, the estimated average rupture time $\overline{t_R}$ is a good estimate of the average of the individual rupture times although there exists a large range of escape.

III – 2. Metastatic Kidney Cancer Monitoring

The algorithm is now run on RECIST score of real patients suffering from kidney cancer. The estimation is performed over a cohort of 176 patients of the HEGP. There is an average of 7 visits per subjects (min: 3, max: 22), with an average duration of 90 days between consecutive visits. We present here a run with a low residual standard variation with respect to the amplitude of the trajectories and complexity of the data



(a) The time warp. *(b) The space warp.*

***Figure 5.8** – Individual random effects.*
Fig. 5.8a: log-acceleration factors ξ_i^1 and ξ_i^2 against times shifts τ_i. Fig. 5.8b: log-amplitude factors ρ_i^1 and ρ_i^2 against space shifts δ_i. In both figures, the colour corresponds to the individual rupture time t_R^i. These estimates hold for the same run as Fig. 5.6.

set: $\sigma = 9.10$.

Figure 5.6 illustrates the qualitative performance of the model on ten patients. Although we cannot explain all the paths of progression, the algorithm succeeds in fitting various types of curves: from the curve γ_6 which is flat to the curve γ_3 which is spiky. From Figure 5.7, it seems that the rupture times occur early in the progression in average.

In Figure 5.8, we plot the individual estimates of the random effects (obtained from the last iteration) in comparison to the individual rupture times. Even though the parameters which lead the space warp, *i.e.* ρ_i^1, ρ_i^2 and δ_i are correlated, the correlation with the rupture time is not clear. In other words, the volume of the tumors seems to not be relevant to evaluate the escapement of a patient. On the contrary, which is logical, the time warp strongly impacts the rupture time.

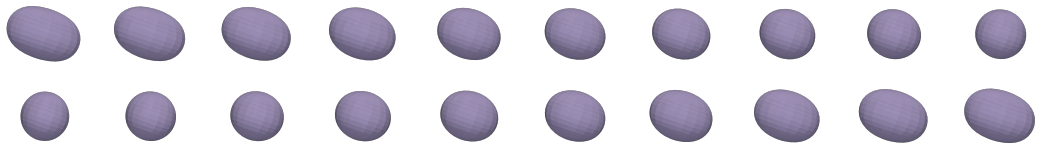
III – 3. Shape Synthetic Data

The dataset consists of 20 synthetic sequences of 3D-shapes built in accordance to the piecewise geodesic shape model described at the paragraph II. The estimation is still performed through the MCMC-SAEM algorithm (Algorithm 1). Real data are not yet available as the segmentation of the tumor has to be done manually, which is complex and time-consuming. This study will motivate new segmentations for future works.

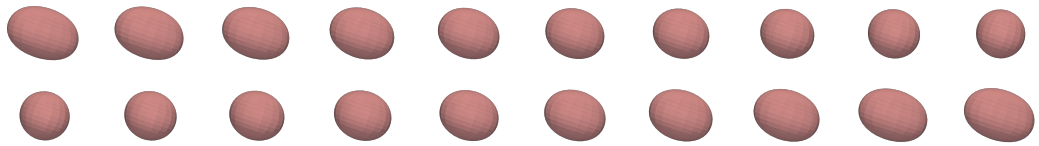
The control points used to construct the data are chosen to be regularly distributed. Thus, the algorithm has no reason to return the same control points: on the contrary, it will return more relevant control points. As momenta and control points share a single dynamic, we rather evaluate the performances on the reconstruction relative error which

***Table 5.4** – The piecewise geodesic shape model.*
Relative errors for the representative rupture shape $\overline{y_R}$ and the representative rupture times $\overline{t_R}$. Mean (standard deviation) relative errors for the individual rupture time $t_R^{i\ \text{estim}}$, within the population. Error of reconstruction for the template and mean (standard deviation) of the error of reconstruction for the individuals. All expressed as a percentage and for a type run.

$\overline{y_R}$	$\overline{t_R}$	Template reconstruction	t_R^i	Individuals reconstruction
1.30	0.01	9.72	0.31 (0.41)	7.94 (5.91)



(a) Template used to generate the dataset.

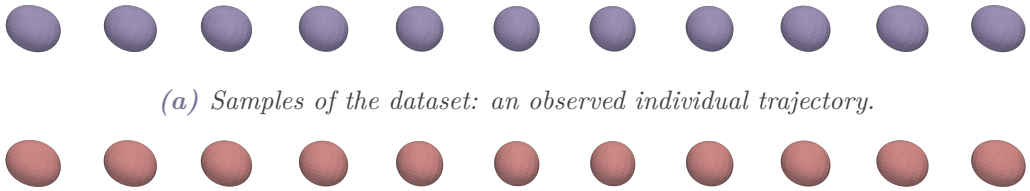


(b) Template estimated after 600 iterations.

***Figure 5.9** – Reconstruction of the template.*
Evolution of the template over time. In purple (Fig. 5.9a), the template used for the generation of the dataset ; In red (Fig. 5.9b), the one estimated by the algorithm.

summaries the goodness of fit of our algorithm.

Table 5.4 displays the relative errors for the estimated representative rupture shape, representative rupture time and individual rupture times. We emphasis the well-estimation of the rupture times t_R and $(t_R^i)_{i\in[\![1,n]\!]}$, which is critical in the target to our application to chemotherapy monitoring. We also provide the relative errors of reconstruction, *i.e.* the relative residual distances between the estimated trajectories and and their corresponding paths in the data set for both the representative path and the individuals ones. Figure 5.9 illustrates the qualitative performance of the reconstruction of the template and Figure 5.10 the qualitative performance of the reconstruction of subject type. Mainly, the reconstruction turns out to be very efficient.



(a) Samples of the dataset: an observed individual trajectory.

(b) Estimation of the same individual trajectory, 600 iterations.

***Figure 5.10** – Reconstruction of the individual trajectories.*
Evolution of a standard subject over time. We keep the same convention than at Figure 5.9: Fig. 5.10a (in purple) shows the shooting of an individual evolution path and Fig. 5.10b (in red)) the corresponding reconstructed one.

– Troisième partie –

A New Class of EM Algorithms

Motivations and State of the Art

Although the expectation-maximization (EM) algorithm introduced by Dempster et al. (1977) is a very popular and often efficient approach to *maximum* likelihood (or *maximum a posteriori*) estimation in incomplete data models, as it a simple use algorithm, it has one major issue : the computation of the expectation with respect to the conditional distribution. Indeed, in certain situations, the EM is not applicable because the expectation step cannot be performed in closed form. To overcome this restriction, many different options have been proposed. The first one is to replace the expectation step by a sampling of the unobserved data step. We refer to this EM version as the Stochastic EM (SEM) algorithm (Celeux and Diebolt, 1985). In particular, in the SEM, only one sample of the latent variable is drawn. A possible generalization of the SEM is the Monte-Carlo EM (MCEM) (Wei and Tanner, 1990), in which a Monte-Carlo implementation of the expectation in the E-step is carried out. In an alternative way, Delyon et al. (1999) proposed to replace the expectation step of the EM algorithm by one iteration of a stochastic approximation procedure, referred to as SAEM, standing for stochastic approximation EM. In addition to avoiding the computation of the expectation, introducing randomness may enable to escape local maxima. However, this is not theoretically proved nor numerically illustrated in the literature.

The convergence of the SAEM toward local *maxima* has been proved in Delyon et al. (1999) and its numerical efficiency has been demonstrated in several situations such as in inference in hidden Markov models (Cappé et al., 2005). However, despite appealing features, the limit point of this algorithm can strongly depend on its initialization. In order to avoid convergence toward local *maxima*, Lavielle and Moulines (1997) have proposed a simulated annealing version of the SAEM. The main idea was to allow the procedure to better explore the state-space by considering a tempered version of the model likelihood. More precisely, assuming that the data are corrupted by an additive Gaussian noise with variance σ^2, at each iteration k of the SAEM algorithm, they consider the "false" model in which the noise variance is equal to $((1 + T_k)\sigma)^2$, where (T_k) is a positive sequence ot temperatures that decreases slowly toward 0. Therefore, the bigger T_k, the more the likelihood of the model is flattened and the optimizing sequence can escape easily from local *maxima*. The simulations gave good results but there were no theoretical guarantee for this procedure. Based on the same idea, Lavielle (2014) has proposed to use the simulated-annealing process as a "trick" to better initialize the

SAEM algorithm. This initialization scheme is implemented in the MONOLIX software and gives impressive results on real data (Chan et al., 2011; Lavielle and Mentré, 2007; Samson et al., 2006).

All theoretical results regarding the convergence of the SAEM algorithm assume that we are able to sample from the posterior distribution, but in practice it may be intractable or have a high computational cost. To overcome this issue, Picchini and Samson (2018) have proposed to couple the SAEM algorithm to an approximate Bayesian computation step (ABC, see Marin et al. (2012) for a review), leading to the ABC-SAEM method in which ABC is used to sample from an approximation to the posterior distribution. Simulations show that this algorithm can be calibrated to return accurate inference, and in some situations it can outperform a version of the SAEM incorporating the bootstrap filter. However, Picchini and Samson (2018) do not provide any theoretical guarantee of its convergence. More broadly, when sampling from the posterior distribution is prohibitive, one may want to shift to variational inference (Blei et al., 2017; Jordan et al., 1999; Wainwright and Jordan, 2008).

Behind variational inference, the main idea is to replace the objective function by a minorant function which is a trade off between a likelihood and a Kullback-Leibler (KL) divergence between the conditional distribution and a parametric probability density function. This minimizing function is then optimized by a stochastic gradient descent. These methods are known to converge toward local minima for bounded parameters or positive objective functions.

We propose here a new stochastic approximation version of the EM algorithm where we do not sample from the exact distribution but rather from a distribution which converges to the conditional one along the algorithm iterations. This new procedure allows us to derive a wide class of SAEM-like algorithms, including the "trick" initialized SAEM of Lavielle (2014) and the ABC-SAEM algorithms, to cope with intractable or difficult sampling. We refer to this new algorithm as the approximated-SAEM.

This general framework allows us to build a procedure, with the thought of the simulated annealing version of the SAEM (Lavielle and Moulines, 1997), to prevent convergence toward local *maxima*. We introduce a sequence of temperatures and sample from a tempered version of the posterior distribution. Therefore, the posterior-likelihood of the model is "flattened" and the optimizing sequence can escape more easily from local *maxima*. We refer to this particular instantiation as the tempering-SAEM. Note that our tempering-SAEM differs to the ones of Lavielle and Moulines (1997) as we do not modify the model but only the sampling-step.

In Chapter 6, we review briefly the EM algorithm and some of its variants. Notably, we recall some theoretical results from Delyon et al. (1999), regarding the convergence of the SAEM algorithm, that are useful to Chapter 7.

In Chapter 7, we introduce our new stochastic approximation version of the EM

algorithm, namely the approximated-SAEM, and prove the convergence of this algorithm toward local *maxima* under usual assumptions. The demonstration of the convergence, by its similarity with the proof of the convergence of the SAEM, highlights the unconstraining nature of the different assumptions and therefore the great applicability of our algorithm. Thus, we provide a theoretical study of the convergence of the tempering-SAEM toward local *maxima*. We also give an heuristic to the convergence of the tempering-SAEM toward "less local" *maxima*.

Section II of Chapter 7 is dedicated to experiments. The first application we take into account is the *maximum* likelihood estimation of the parameters of a multivariate Gaussian mixture models. This example supports the previous heuristic discussion and gives intuitions into the behavior of the tempering-SAEM algorithm. The second application consists in independent factor analysis (Attias, 1999). In both applications, we focus on the contribution of the tempering-SAEM in comparison to the SAEM.

Chapter 7 is based on Allassonnière and Chevallier (2019).

– Chapter VI –

A Brief Review of the EM-like Algorithms

Numériquement l'estimation des paramètres du modèle introduit au Chaptire 3 est réalisée *via* une approximation stochastique de l'algorithme EM, à savoir l'algorithme SAEM. Avant d'étudier plus en profondeur cet algorithme, on présente dans ce chapitre la littérature classique concernant l'algorithme EM et ses variantes usuelles. Nous portons une attention particulière au papier séminal de Delyon et al. (1999).

Contents

The EM algorithm is a powerful computational technique for locating maxima of functions. It is widely used in statistics for *maximum* likelihood or *maximum a posteriori* estimation in incomplete data models. In this chapter, we briefly review the EM algorithm and some of its variants. Notably, we recall at Appendix A some theoretical results regarding the convergence of the SAEM algorithm, that are useful to the demonstration of the convergence of the approximated-SAEM algorithm introduced at Chapter 7.

I. The Expectation-Maximization Algorithm

We use in this chapter and in the following one the classical terminology of the missing data problem, even though EM algorithm applies to a more general context.

Let $\mathcal{Y} \subset \mathbb{R}^{n_y}$ denote the set of observations, $\mathcal{Z} \subset \mathbb{R}^{n_z}$ the set of latent variables and $\Theta \subset \mathbb{R}^{n_\theta}$ the set of admissible parameters. Let μ be a σ-finite positive Borel measure on $\mathcal{Z}$. For sake of simplicity, we will use the notation q for different likelihoods, specifying their variables in brackets. In particular, for all $(y;\theta) \in \mathcal{Y} \times \Theta$, $z \mapsto q(y,z;\theta)$ is the complete likelihood given the observation y and parameter θ and we assume it is integrable with respect to the measure μ. As for, we note $q(y;\theta) = \int_{\mathcal{Z}} q(y,z;\theta)\,\mathrm{d}\mu(z)$ the observed likelihood and $q(z|y;\theta) = \frac{q(y,z;\theta)}{q(y;\theta)}$ the conditional distribution of the missing data z given the observed data y. Our goal is to estimate the parameters that maximize the likelihood of the observations of n independent samples of a random variable Y, *i.e.* that maximize the observed data likelihood:

$$\text{Given } \; y_1^n = (y_1, \dots, y_n) \in \mathcal{Y}^n, \quad \widehat{\theta}_n^{MLE} \in \operatorname*{argmax}_{\theta\in\Theta} q(y_1^n;\theta)\,.$$

Unfortunately, MLE problem has generally no closed-form solution. The expectation-maximization (EM) algorithm and its variants are powerful algorithms which have demonstrated their efficiency in practice, for instance for clustering with the help of Gaussian Mixture Model (GMM. See Section II.1 of Chapter 7 for details).

For the sake of readability, y_1^n is noted as y throughout the rest of the chapter.

Given a current parameter estimate θ_k, the EM algorithm seeks to find the MLE by iteratively applying these two steps:

E-step: Compute the conditional expected log-likelihood

$$Q(\theta|\theta_k) \;=\; \int_{\mathcal{Z}} \log q(y,z;\theta) q(z|y;\theta_k)\;\mathrm{d}z \;=\; \mathbb{E}\left[\log q(Z|y,\theta_k)\right]\;;$$

M-step: Maximize $Q(\,\cdot\,|\theta_k)$ in the feasible set Θ:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta\in\Theta} Q(\theta|\theta_k)\,.$$

I – 1. Convergence of the EM Algorithm for Curved Exponential Families

This algorithm has been first introduced by Dempster et al. (1977) and convergence under very general conditions has been established by Wu (1983). Delyon et al. (1999) have reframed the convergence of the EM algorithm under easier hypothesis. We present here their results. In particular, they assume that the complete data likelihood belongs to the curved exponential family (M1). In practice, this assumption is not so restrictive and even complex models, such as the model we have introduced in Part II., satisfy this assumption.

(M1) The parameter space Θ is an open subset of $\mathbb{R}^{n_\theta}$. For all $y \in \mathcal{Y}$, $z \in \mathcal{Z}$ and $\theta \in \Theta$, the complete data likelihood function can be expressed as

$$q(y,z;\theta) \;=\; \exp\Big(-\psi(\theta) + \Big\langle S(y,z)\,\Big|\,\phi(\theta)\Big\rangle\Big)$$

where $S : \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_s}$ is a Borel function and $\mathcal{S}$ is an open subset of $\mathbb{R}^{n_s}$. The convex hull of $S(\mathbb{R}^{n_z})$ is included in $\mathcal{S}$. For all $\theta \in \Theta$, all $y \in \mathcal{Y}$, we have

$$\int_{\mathcal{Z}} \|S(y,z)\|\, q(z|y;\theta)\,\mathrm{d}\mu(z) \;<\; +\infty\,.$$

Let $\ell\colon \Theta \to \mathbb{R}$ and $L\colon \mathcal{S}\times\Theta \to \mathbb{R}$ defined as,

$$\begin{aligned} \text{for all } y\in\mathcal{Y}, \qquad \ell\colon\ \theta \;&\mapsto\; \int_{\mathcal{Z}} q(y,z;\theta)\,\mathrm{d}\mu(z) \\ \text{and} \qquad L\colon\ (s,\theta) \;&\mapsto\; -\psi(\theta) + \Big\langle s\,\Big|\,\phi(\theta)\Big\rangle\,. \end{aligned}$$

(M2) The two functions $\psi\colon \Theta \to \mathbb{R}$ and $\phi\colon \Theta \to \mathcal{S}$ are twice continuously differentiable on Θ.

(M3) The function $\bar{s} : \Theta \to \mathcal{S}$ is continuously differentiable on Θ where $\bar{s}$ is defined as:

$$\forall y\in\mathcal{Y}, \qquad \bar{s}\colon\ \theta \;\mapsto\; \int_{\mathcal{Z}} S(y,z) q(z|y;\theta)\,\mathrm{d}\mu(z) \;=\; \mathbb{E}_\theta\left[S(Z)\right]\,;$$

(M4) The function $\ell\colon \Theta \to \mathbb{R}$ is continuously differentiable and for all $y \in \mathcal{Y}$ and $\theta \in \Theta$

$$\partial_\theta \int_{\mathcal{Z}} q(y,z;\theta)\,\mathrm{d}\mu(z) \;=\; \int_{\mathcal{Z}} \partial_\theta\, q(y,z;\theta)\,\mathrm{d}\mu(z)\,;$$

(M5) There exists a continuously differentiable function $\hat{\theta}\colon \mathcal{S} \to \Theta$ such that

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

Let, for all $\theta \in \Theta$ and closed set $E \subset \Theta$, $d(\theta, E)$ denotes the distance of θ to E.

Theorem 6.1 (*Convergence of the EM – Delyon et al. (1999)*)

Assume that (M1 - 5) hold. Assume in addition that for any $\theta \in \Theta$, clos $(\mathcal{L}(\theta))$ is a compact subset of Θ. Then, for any initial point $\theta_0 = \theta$, the sequence $(\ell(\theta_k))_{k\in\mathbb{N}}$ is increasing and

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

Note that the monotonicity of the EM algorithm guarantees that as EM iterates, its guesses won't get worse in terms of their likelihood.

I – 2. Variants of the EM Algorithm

We give here some variants of the EM algorithm, but postpone the stochastic approximation of the EM (SAEM) algorithm to a dedicated section (Section II). Variants can be split into three groups : the ones which are mainly motivated by convergence speed considerations, the ones which are more concerned with limitations concerning the M-step and, last, the ones which cope with limitations about the E-step. The SAEM algorithm falls into this third category.

a. Speeding up the EM Algorithm

The convergence rate of the EM algorithm is, in the general form, only linear and governed by the fraction of missing data. In particular, the EM algorithm has relatively slow convergence compared to numerical optimization approaches like Newton–Raphson updates. In order to speed up the convergence of the EM algorithm, many variants have been proposed. Several examples and a pedagogic introduction to convergence rates and speeding-up methods are given in the book by McLachlan and Krishnan (2007), in the tutorial by Roche (2012) and in the tutorial by Gupta and Chen (2011).

b. The Generalized EM Algorithm – GEM Algorithm

The vanilla M-step of the EM algorithm consists in maximizing the conditional expected log-likelihood $Q(\,\cdot\,|\theta_k)$. It is however possible to relax the requirement of maximization to just increasing $Q(\,\cdot\,|\theta_k)$, *i.e.* to substitute the vanilla M-step for:

M-step: Finding θ_{k+1} such that $Q(\theta_{k+1}|\theta_k) \geqslant Q(\theta_k|\theta_k)$,

where θ_k denotes the current parameter. This form of the algorithm is called *Generalized EM* (GEM) and is also guaranteed to converge (Delyon et al., 1999).

Also to cope with the problem of intractable maximization, Lange (1995) introduced a gradient algorithm that is closely related to the EM algorithm : The *Gradient-EM*. This algorithm solves the M-step of the EM algorithm by one iteration of the Newton-Raphson's method.

c. The Stochastic EM Algorithm

When the expectation of the conditional log-likelihood $Q(\cdot|\theta_k)$ cannot be analytically computed, an alternative is to introduce some stochasticity is the estimation procedure.

The simple idea of the *stochastic EM* algorithm (Celeux and Diebolt, 1985) is to replace the computation and the maximization of $Q(\cdot|\theta_k)$ by the much simpler computation of conditional distribution $q(z|y;\theta_k)$ and the simulation of an unobserved sample z_k. Afterwards, it maximizes the complete-data log-likelihood only for that fixed assignment. This leads to the following modified iteration: given a current estimate θ_k;

S-step: Draw an unobserved sample z_k from $q(\,\cdot\,|y;\theta_k)\,$;

M-step Find $\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\; q(y, z_k;\theta)\,.$

Theorem 6.1 ensures the convergence of the EM algorithm towards stationary points of the log-likelihood. However, the limiting position of the EM algorithm greatly depends on its initial position. The Stochastic EM algorithm aims to reduce the dependency on this initial position. Its inherent randomness technically allows the SEM algorithm to escape from a saddle point or an undesired local maximum of the likelihood function. However, its convergence is only proven in mean.

d. The Monte-Carlo EM Algorithm

The Monte-Carlo EM (MCEM) algorithm (Wei and Tanner, 1990) proposes a Monte-Carlo implementation of the expectation in the E-step. Thus, the MCEM algorithm can be seen as a generalization of the SEM algorithm in which we draw m independent samplings of the latent variable Z instead of just one. In other words, given a current estimate θ_k, the MCEM algorithm writes:

S-step: Draw m samples $z_k^j \sim q(\,\cdot\,|y;\theta_k)\,$, where $j \in [\![1, m]\!]\,$;

E-step: Compute the Monte-Carlo estimation $Q_k(\theta) = \dfrac{1}{m}\displaystyle\sum_{j=1}^{m} \log q(y, z_k^j;\theta)\,$;

M-step: Maximize Q_k, *i.e.* find $\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\; Q_k(\theta)\,.$

In order Q_k to well-approximate the expectation $Q(\cdot|\theta_k)$, we have to choose a large m, significantly increasing the computation time. However, a large m reduce the stochasticity of the MCEM which will then have the same drawbacks than the EM algorithm. As already said, for $m = 1$ the MCEM reduces to the SEM algorithm. Thus an efficient strategy is to consider the "simulated annealing" version of the MCEM algorithm: Wei and Tanner (1990) recommend to start with small values of m and then to increase m to infinity at suitable rate of performance. Then we go from pure SEM ($m = 1$) to pure EM ($m = +\infty$).

Biscarat (1994) and Celeux et al. (1995) have proved the convergence of the simulated annealing MCEM toward local minima of the log-likelihood, for a suitable sequence of m and in the mixture context. More broadly, Fort and Moulines (2003) proved the almost-sure convergence of the MCEM algorithm. Moreover, they proved the convergence of the MCMC-MCEM algorithm, *i.e.* of the MCEM algorithm in combination with Markov chain Monte Carlo simulation methods for the S-step, under weak conditions on the simulation kernel.

II. The Stochastic Approximation EM Algorithm

The stochastic approximation version of the EM (SAEM) algorithm (Delyon et al., 1999) consists in replacing the E-step by a stochastic approximation obtained by the use of simulated data. Thus, the update of the current parameter is benefiting from the previous simulation. More precisely, given an initial approximation of the expected log-likelihood Q_0 and a sequence of positive step-size $(\gamma_k)_{k\in\mathbb{N}}$, the iteration k of the SAEM algorithm consists of the three following steps:

S-step: Sample the latent variable z_k under the conditional density $q(\cdot|y;\theta_k)$;

SA-step: Update $Q_k(\theta)$ as $Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k\big(\log q(y, z_k;\theta) - Q_k(\theta)\big)$;

M-step: Maximize Q_{k+1} in the feasible set Θ, *i.e.* find $\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q_{k+1}(\theta)$.

In practice, we set Q_0 to be the null function over Θ.

Note that if $\gamma_k = 1$ for any $k \in \mathbb{N}$, *i.e.* if there is no memory in the stochastic approximation, the SAEM reduces to the SEM. A major advantage of the SAEM algorithm with respect to the MCEM is the number of simulations required to ensure the theoretical convergence (see the next paragraph) of the procedure: only one simulation against m tending to infinity.

II – 1. Convergence of the SAEM Algorithm Toward Local Maxima

Let $\mathcal{F} = \{\mathcal{F}_k\}_{k\in\mathbb{N}}$ be the natural filtration with respect to the process $(z_k)_{k\in\mathbb{N}}$, where $z_k \sim q(\,\cdot\,|y;\theta_{k-1})$ for all k.

$$\forall k \in \mathbb{N}, \quad \mathcal{F}_k = \sigma\left(\{z_1, z_2, \ldots, z_k\}\right) = \sigma(\{z_1^{-1}(A), z_2^{-1}(A), \ldots z^{-1}(A) | A \in \mathcal{B}(\mathcal{Z})\})\,.$$

We assume that (M1 - 5) hold. In particular, as the complete log-likelihood belongs to the curved exponential family (M1), the SA-step is more conveniently (and equivalently) replaced by an update of the estimation of the conditional expectation of the sufficient statistics. Hence, the SA-step rewrites

SA-step Update $s_k(\theta)$ as $s_{k+1}(\theta) \;=\; s_k(\theta) + \gamma_k(S(y, z_k) - s_k(\theta))\,,$

where for all $\theta \in \Theta$, $s_0(\theta) = 0$.

Let consider the following assumptions which are taken from Delyon et al. (1999):

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

(SAEM2) The two functions $\psi\colon \Theta \to \mathbb{R}$ and $\phi\colon \Theta \to \mathcal{S}$ are n_s times differentiable;

(SAEM3) For all positive Borel functions ϕ, for all $k \in \mathbb{N}$ and all $y \in \mathcal{Y}$,

$$\mathbb{E}[\phi(Z_{k+1})|\mathcal{F}_k] \;=\; \int_{\mathcal{Z}} \phi(z) q(z|y;\theta_k)\,\mathrm{d}\mu(z)\,;$$

(SAEM4) For all $\theta \in \Theta$, all $y \in \mathcal{Y}$ and all $k \in \mathbb{N}$,

$$\int_{\mathcal{Z}} \|S(y, z)\|^2\, q(y, z;\theta)\,\mathrm{d}\mu(z) \;<\; +\infty\,.$$

Theorem 6.2 (*Convergence of the SAEM – Delyon et al. (1999)*)

Assume that (M1 - 5) and (SAEM1 - 4) hold. Assume in addition that, with probability 1, $\mathrm{clos}\left(\{s_k\}_{k\in\mathbb{N}^*}\right)$ is a compact subset of $\mathcal{S}$. Then, with probability 1,

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

This theorem ensures the convergence of the SAEM algorithm only toward a stationary point of the log-likelihood. To ensure the convergence of the algorithm toward a local *maximum* we have to assume at least local convexity of the log-likelihood. Some conditions upon which the convergence toward local *maxima* is guaranteed are given in Section 7 of Delyon et al. (1999). As these conditions are model-dependent, we do not focus on this aspect in this chapter. Basically, this is equivalent to assuming that the observed likelihood is convex.

II – 2. The Monte-Carlo Markov-Chain SAEM Algorithm

In the S-step of the SAEM algorithm, we assume that we are able to sample from the conditional density $q(\cdot\,|y;\theta)$ for any $\theta \in \Theta$. In many practical situations, it will not be possible to generate the latent variable data z exactly. To overcome this issue, Kuhn and Lavielle (2004) proposed to use a Monte-Carlo Markov chain procedure in the S-step of SAEM. More precisely, we assume that for all $\theta \in \Theta$, the conditional distribution $q(\cdot\,|y;\theta)$ is the unique limiting distribution of a transition kernel Π_θ and we draw a new z_k with respect to the kernel Π_{θ_k}, starting at z_{k-1}. Hence, the k^{th} iteration of the MCMC-SAEM consists of the three following steps:

S-step: Generate a realization $z_k \sim \Pi_{\theta_k}(z_{k-1}, \cdot)$;

SA-step: Update $Q_k(\theta)$ as $Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k\big(\log q(y, z_k;\theta) - Q_k(\theta)\big)$;

M-step: Maximize $Q_{k+1}(\theta)$ in the feasible set Θ, *i.e.* find $\theta_{k+1} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q_{k+1}(\theta)$.

Note that under Assumption (M1), as well as for the SAEM algorithm, the stochastic approximation can be applied to the sufficient statistics.

Kuhn and Lavielle (2004) proved the convergence of the MCMC-SAEM algorithm but they assumed in their demonstration that the chain $(z_k)_{k\in\mathbb{N}}$ generated during the estimation procedure takes its values in a compact subset. However, this condition is hardly ever met. Allassonnière et al. (2010) relaxed this assumption and proved the convergence of the MCMC-SAEM toward local maxima in a more general setting.

The numerical efficiency of the MCMC-SAEM algorithm have been demonstrated in numerous practical situations. As an example, it is the core of the Monolix software.

Despite appealing features, the SAEM algorithm suffers from two main drawbacks. Firstly, it requires to be able to sample from the conditional distribution $q(\cdot\,|y;\theta)$ or at least to be able to target this distribution by a Markov chain. In practice, one may want to relax this constraint. Secondly, the convergence of the SAEM is only local and the limit point of this algorithm can strongly depend on its initialization.

In the following chapter, we propose a new class of stochastic approximation of the EM algorithm to overcome this two issues.

Appendix A. The Theorem 2 and Lemma 2 of Delyon et al. (1999)

In Chapter 7, we prove the convergence of a generalization of the SAEM algorithm. As well as Theorem 6.2, our proof is mainly based on the following theorem and lemma. The

theorem establish the convergence of Robin-Monroe type approximation procedures, *i.e.* the convergence of sequences defined recursively as

$$\forall k \in \mathbb{N}, \qquad s_k \;=\; s_{k-1} + \gamma_k\big(h(s_k) + r_k + e_k\big)\,.$$

In order to keep consistent appellation, in Chapter 7, we will refer to this theorem as *Theorem 2* of Delyon et al. (1999).

Theorem 6.3 (*Theorem 2 of Delyon et al. (1999)*)

Assume that

(SA0) With probability 1, for all $k \in \mathbb{N}$, $s_k \in \mathcal{S}$.

(SA1) $(\gamma_k)_{k\in\mathbb{N}^*}$ is a decreasing sequence of positive numbers such that $\sum_{k=1}^{\infty} \gamma_k = \infty$.

(SA2) The vector field h is continuous on $\mathcal{S}$ and there exists a continuously differentiable function $V : \mathcal{S} \to \mathbb{R}$ such that :

(i) for all $s \in \mathcal{S}$, $F(s) \;=\; \Big\langle \mathrm{d}_s V(z) \,\Big|\, h(s) \Big\rangle \leqslant 0\,,$

(ii) $\mathrm{int}\,(V(\mathcal{L})) \;=\; \varnothing$ where $\mathcal{L} = \{s \in \mathcal{S} | F(s = 0)\}\,.$

(SA3) With probability 1, $\mathrm{clos}\,(\{s_k\}_{k\in\mathbb{N}})$ is a compact subset of $\mathcal{S}$.

(SA4) With probability 1, $\sum \gamma_k e_k$ exists and is finite, $\lim r_k = 0$.

Then, with probability 1, $\overline{\lim}\, d(s_k, \mathcal{L}) \;=\; 0$.

Similarly, we will refer to this lemma as *Lemma 2* of Delyon et al. (1999). It demonstrates the existence of a Lyapunov function for the vector field $h\colon\; s \mapsto \bar{s}(\hat{\theta}(s)) - s$.

Lemma 6.4.1. Assume (M1 - 5) and (SAEM2). Then (SA2) is satisfied with $V = -\ell \circ \hat{\theta}$. Moreover,

$$\begin{aligned} \{s \in \mathcal{S} | F(s) = 0\} \;&=\; \{s \in \mathcal{S} |\, \mathrm{d}_s V(s) \;=\; 0\} \\ \text{and} \qquad \hat{\theta}\left(\{s \in \mathcal{S} | F(s) = 0\}\right) \;&=\; \{\theta \in \Theta |\, \mathrm{d}_\theta \ell(\theta) = 0\}\,, \end{aligned}$$

where $F\colon\; s \mapsto \Big\langle \mathrm{d}_s V(s) \,\Big|\, h(s) \Big\rangle$.

– Chapter VII –

A New Class of SAEM Algorithms. Escaping Local Minima and Handling Intractable Sampling

Malgré la performance numérique de l'algorithme SAEM, du fait de la complexité de notre modèle, nos propres expériences se sont heurtées aux limites de ce dernier. En particulier, l'algorithme SAEM est très sensible à ses conditions initiales et, malgré la stochasticité de la procédure induite par l'approximation stochastique, peut rester piégé dans des minima locaux. De plus, l'algorithme SAEM suppose que l'on est à même de simuler la loi conditionnelle des variables latentes sachant les observations (avec la terminologie des modèles à variables latentes), éventuellement par une méthode de type MCMC, ce qui n'est pas toujours le cas.

Nous proposons dans ce chapitre une nouvelle classe d'algorithmes SAEM : les algorithmes SAEM approchés, ou approximated-SAEM en anglais, dont on démontre la convergence vers des minima locaux sous des hypothèses standards. Cette classe repose sur la simulation par une loi approchée, en un sens à définir, de la vraie loi conditionnelle dans l'étape de simulation. En particulier, on englobe des algorithmes pré-existants tel que l'ABC-SAEM (Picchini et Samson, 2018) dont l'efficacité numérique avait été établie mais dont la convergence théorique n'avait pas été démontrée.

Enfin, en se basant sur des techniques de recuit simulé, on propose une version tempérée de l'algorithme SAEM afin de favoriser sa convergence vers des minima globaux. Dans cette version, on approche la loi conditionnelle en la tempérant suivant un schéma de températures sinusoïdal amorti. Nous appliquons cette méthode à l'estimation des paramètres dans les modèles de mélange gaussien et en illustrons ainsi la supériorité numérique sur l'algorithme SAEM. Cet algorithme est également appliqué à la séparation de sources *via* l'analyse en facteurs indépendants.

Contents

DESPITE appealing features, the limit position of the SAEM algorithm can strongly depend on its starting position. Moreover, sampling from the posterior distribution may be intractable or have a high computational cost. To cope with this two issues, we propose here a new stochastic approximation version of the EM in which we do not sample from the exact distribution in the expectation phase of the procedure.

We first prove the convergence of this algorithm toward local maxima of the observed likelihood. Then, we propose an instantiation of this general procedure to favor convergence toward global maxima. Experiments on synthetic and real data highlight the performance of this algorithm in comparison to the SAEM.

I. Maximum Likelihood Estimation through an EM-Like Algorithm

We use in the sequel the classical terminology of the missing data problem, even though the approaches developed here apply to a more general context.

Let $\mathcal{Y} \subset \mathbb{R}^{n_y}$ denote the set of observations, $\mathcal{Z} \subset \mathbb{R}^{n_z}$ the set of latent variables and $\Theta \subset \mathbb{R}^{n_\theta}$ the set of admissible parameters. Let μ be a σ-finite positive Borel measure on $\mathcal{Z}$. For sake of simplicity, we will use the notation q for different likelihoods, specifying their variables in brackets. In particular, for all $(y;\theta) \in \mathcal{Y} \times \Theta$, $q(y,\,\cdot\,;\theta)$ is the complete likelihood given the observation y and parameter θ and we assume it is integrable with respect to the measure μ. As for, we note $q(y;\theta) = \int_{\mathcal{Z}} q(y,z;\theta)\,\mathrm{d}\mu(z)$ the observed likelihood and $q(z|y;\theta) = \frac{q(y,z;\theta)}{q(y;\theta)}$ the posterior distribution of the missing data z given the observed data y. Our goal is to estimate the parameters that maximize the likelihood of the observations of n independent samples of a random variable Y, *i.e.* that maximize the observed data likelihood.

I – 1. A New Stochastic Approximation Version of the EM Algorithm

We propose in this contribution a generalization of the SAEM algorithm, referred to as approximated-SAEM. Similar to the SAEM, the basic idea is to split the E-step into a simulation step and a stochastic averaging procedure. This averaging concerns the conditional expected log-likelihood

$$\theta \mapsto Q(\theta|\theta_k) = \int_{\mathcal{Z}} \log q(y,z|\theta) q(z|y,\theta_k)\ \mathrm{d}z\,,$$

where θ_k denotes the current optimal parameter. Starting from $Q_0(\theta) = 0$ for all $\theta \in \Theta$, we build an approximation of $\theta \mapsto Q(\theta|\theta_k)$ through stochastic approximation of the compete log-likelihood. We denote Q_k this approximation. In the original SAEM, the S-step consists in generating realizations of the missing data vector under the posterior distribution $q(\cdot|y;\theta)$. Here, we propose to sample under *approximation* of the posterior distribution. The following paragraph describes this new algorithm.

Let $\gamma = (\gamma_k)_{k\in\mathbb{N}}$ be a sequence of positive step-size for the stochastic approximation, and $\tilde{q} = (\tilde{q}_k)_{k\in\mathbb{N}}$ be a sequence of *approximated* distributions on $\mathcal{Z} \times \Theta$ such that for all $k \in \mathbb{N}$ and all $\theta \in \Theta$, $\tilde{q}_k(\cdot;\theta)$ is integrable on $\mathcal{Z}$ with respect to the measure μ. As in the SAEM, once the step size γ_k decreases, we can consider a constant number of simulations. In practice (and from now on to avoid cumbersome notations), as the S-step is generally the most computationally costly, we set this number to one. Then, the approximated-SAEM iterates the following three steps:

S-step: Sample the latent variable $\tilde{z}_k$ under the approximated density $\tilde{q}_k(\cdot;\theta_k)$;

SA-step: Update $Q_k(\theta)$ according to

$$Q_k(\theta) \;=\; Q_{k-1}(\theta) + \gamma_k\big(\log q(y,\tilde{z}_k;\theta) - Q_{k-1}(\theta)\big)\,;$$

M-step: Maximize Q_k in the feasible set Θ, *i.e.* find $\theta_{k+1} \in \Theta$ such that

$$\forall\theta \in \Theta, \quad Q_k(\theta_{k+1}) \;\geqslant\; Q_k(\theta)\,.$$

Note that without approximation, *i.e.* if the approximated densities $\tilde{q}_k$ match with the correct posterior distribution, we feature the classical SAEM. Moreover, the approximated densities $\tilde{q}_k$ may depend on the observations y, as in variational Bayesian methods or may be done by ABC samplers as in ABC-SAEM. In Section I.2, we propose a way to build a sequence $\tilde{q}$ leading to good properties in practice and theoretical guarantees are given in the following section.

a. Curved Exponential Family

Before establishing the convergence of this procedure, we briefly recall the hypothesis required to prove the convergence of the EM. More precisely, we restrict our attention to models for which the complete data likelihood belongs to the curved exponential family. In this paragraph and the following, we keep the notations of Delyon et al. (1999): an hypothesis stated with a ($\star$) means that it is a direct generalization of the corresponding one in Delyon et al. (1999); on the contrary, hypothesis stated without are unchanged compared to the original one. See Section I.1 of Chapter 6 for a reminder of the original assumptions.

(M1*) The parameter space Θ is an open subset of $\mathbb{R}^{n_\theta}$. For all $y \in \mathcal{Y}$, $z \in \mathcal{Z}$ and $\theta \in \Theta$, the complete data likelihood function can be expressed as

$$q(y,z;\theta) \;=\; \exp\Big(-\psi(\theta) + \Big\langle S(y,z)\,\Big|\,\phi(\theta)\Big\rangle\Big)$$

where $S : \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_s}$ is a Borel function and $\mathcal{S}$ is an open subset of $\mathbb{R}^{n_s}$. The convex hull of $S(\mathbb{R}^{n_z})$ is included in $\mathcal{S}$. For all $\theta \in \Theta$, all $y \in \mathcal{Y}$ and all $k \in \mathbb{N}$, we have

$$\int_{\mathcal{Z}} \|S(y,z)\|\, \tilde{q}_k(z;\theta)\, \mathrm{d}\mu(z) \;<\; +\infty \qquad \text{and} \qquad \int_{\mathcal{Z}} \|S(y,z)\|\, q(z|y;\theta)\, \mathrm{d}\mu(z) \;<\; +\infty\,.$$

Let $\ell\colon \Theta \to \mathbb{R}$ and $L\colon \mathcal{S} \times \Theta \to \mathbb{R}$ defined as,

$$\text{for all } y \in \mathcal{Y}, \qquad \ell\colon\ \theta \;\mapsto\; \int_{\mathcal{Z}} q(y,z;\theta)\, \mathrm{d}\mu(z)$$
$$\text{and} \qquad L\colon\ (s,\theta) \;\mapsto\; -\psi(\theta) + \left\langle s \,\middle|\, \phi(\theta) \right\rangle .$$

Note that if the function ℓ is defined as above, it implicitly depends on y. However, as the observations y are assumed to be i.i.d, the counterpart function that would be defined on the whole set $\mathcal{Y}$ would be a sum over y of the defined ℓ.

(M2) The two functions $\psi\colon \Theta \to \mathbb{R}$ and $\phi\colon \Theta \to \mathcal{S}$ are twice continuously differentiable on Θ.

(M3) The function $\bar{s} : \Theta \to \mathcal{S}$ is continuously differentiable on Θ where $\bar{s}$ is defined as:

$$\forall y \in \mathcal{Y}, \qquad \bar{s}\colon\ \theta \;\mapsto\; \int_{\mathcal{Z}} S(y,z) q(z|y;\theta)\, \mathrm{d}\mu(z) \;=\; \mathbb{E}_\theta\left[S(Z)\right]\,;$$

(M4) The function $\ell\colon \Theta \to \mathbb{R}$ is continuously differentiable and for all $y \in \mathcal{Y}$ and $\theta \in \Theta$

$$\partial_\theta \int_{\mathcal{Z}} q(y,z;\theta)\, \mathrm{d}\mu(z) \;=\; \int_{\mathcal{Z}} \partial_\theta\, q(y,z;\theta)\, \mathrm{d}\mu(z)\,;$$

(M5) There exists a continuously differentiable function $\hat{\theta}\colon \mathcal{S} \to \Theta$ such that

$$\forall \theta \in \Theta, \quad \forall s \in \mathcal{S}, \quad L(s, \hat{\theta}(s)) \;\geqslant\; L(s,\theta)\,.$$

Hypothesis (M1*) differs from (M1) as we do not only require the function $z \mapsto \|S(z;\theta)\|$ to be integrable with respect to the posterior measure $q(\cdot|y;\theta)\,\mathrm{d}\mu$, but also with respect to all approximated distributions $\tilde{q}_k(\cdot;\theta)\,\mathrm{d}\mu$, for all parameters $\theta \in \Theta$, all observations $y \in \mathcal{Y}$ and all iterations $k \in \mathbb{N}$. For most models of practical interest (see for instance Section II.2), the function $L(s;\cdot)$ has a unique global *maximum* and the existence and the differentiability of $\hat{\theta}$ is a direct consequence of the implicit function theorem.

For exponential families, the SA-step is more conveniently (and equivalently) replaced by an update of the estimation of the conditional expectation of the sufficient statistics. Let s_k denote the k^{th} approximation of the conditional expectation of the sufficient statistics. Then, the k-th iteration of the approximated-SAEM summarizes in:

$$\begin{cases} s_k \;=\; s_{k-1} + \gamma_k(S(y,\tilde{z}_k) - s_{k-1}) \\ \theta_k \;=\; \hat{\theta}(s_k) \end{cases} \qquad \text{where} \qquad \tilde{z}_k \sim \tilde{q}_k(\cdot;\theta_{k-1})\,. \tag{7.1}$$

where s_k is initialized to zero: $s_0(\theta) = 0$ for all $\theta \in \Theta$.

b. Convergence Toward Local Maxima

Let $\widetilde{\mathcal{F}} = \{\widetilde{\mathcal{F}}_k\}_{k\in\mathbb{N}}$ the natural filtration with respect to the process $(\tilde{z}_k)_{k\in\mathbb{N}}$ and $\mathcal{F} = \{\mathcal{F}_k\}_{k\in\mathbb{N}}$ the natural filtration with respect to the process $(z_k)_{k\in\mathbb{N}}$ where $z_k \sim q(\cdot|y;\theta_{k-1})$ for all k. Let, for all set $\mathcal{X}$, $\mathrm{clos}(\mathcal{X})$ denotes the closure of $\mathcal{X}$ and consider the following assumptions which are generalization of the ones of Delyon et al. (1999):

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

(SAEM2) The two functions $\psi\colon \Theta \to \mathbb{R}$ and $\phi\colon \Theta \to \mathcal{S}$ are n_s times differentiable;

(SAEM3*) For all positive Borel functions ϕ, for all $k \in \mathbb{N}$ and all $y \in \mathcal{Y}$,

$$\begin{aligned} & \mathbb{E}[\phi(Z_{k+1})|\widetilde{\mathcal{F}}_k] \;=\; \int_{\mathcal{Z}} \phi(z)\tilde{q}_k(z;\theta_k)\,\mathrm{d}\mu(z) \\ \text{and} \qquad & \mathbb{E}[\phi(Z_{k+1})|\mathcal{F}_k] \;=\; \int_{\mathcal{Z}} \phi(z)q(z|y;\theta_k)\,\mathrm{d}\mu(z)\,; \end{aligned}$$

(SAEM4*) For all $\theta \in \Theta$, all $y \in \mathcal{Y}$ and all $k \in \mathbb{N}$,

$$\int_{\mathcal{Z}} \|S(y,z)\|^2\, \tilde{q}_k(z;\theta)\,\mathrm{d}\mu(z) \;<\; +\infty\,.$$

Likewise, see Assumption (SAEM1) is characteristic of stochastic approximation procedures in which the step-size have to decrease not too fast. Like Assumption (M1*), (SAEM3*) is similar to (SAEM3), except that we assume that, given $\theta_0,\ldots,\theta_k$, *both* simulated latent variables $\tilde{z}_1,\ldots,\tilde{z}_k$ and $z_1,\ldots,z_k$ are conditionally independent, given their respective natural filtration. In Assumption (SAEM4*), we demand the integrability of $z \mapsto \|S(y,z)\|^2$ with respect to the measures $\tilde{q}_k(z;\theta)\,\mathrm{d}\mu$.

The following theorem ensures the convergence of our new stochastic approximation version of the EM algorithm. This theorem is the approximated counterpart of Theorem 5 of Delyon et al. (1999).

Theorem 7.1 (*Convergence of the approximated-SAEM*)

Assume that (M1*), (M2 - 5), (SAEM1), (SAEM2), (SAEM3*) and (SAEM4*) hold. Assume in addition that:

(A) For all $y \in \mathcal{Y}$, the sequence $\left(\tilde{q}_k(\cdot;\theta)\right)_{k\in\mathbb{N}}$ converge in mean on every compact subset of Θ for the measure $S.\mu$ to $q(\cdot|y;\theta)$, that is to say for all observations $y \in \mathcal{Y}$ and all compact $\mathcal{K} \subset \Theta$,

$$\lim_{k\to\infty} \left\{ \sup_{\theta\in\mathcal{K}} \int_{\mathcal{Z}} S(y,z)\left(\tilde{q}_k(z;\theta) - q(z|y;\theta)\right)\mathrm{d}\mu(z) \right\} = 0\,;$$

(B) With probability 1, $\mathrm{clos}\left(\{s_k\}_{k\in\mathbb{N}^*}\right)$ is a compact subset of $\mathcal{S}$.

Let $\mathcal{L} = \{\theta \in \Theta | \partial_\theta \ell(\theta) = 0\}$. Then, with probability 1,

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

Hypothesis (A) makes explicit what we mean by sequence of approximated densities. In particular, it allows a wide variety of numerical schemes; we propose an example of practical interest in Section I.2. Note that (SAEM4*) and (A) ensure the function $z \mapsto \|S(y,z)\|^2$ to be integrable with respect to the measure $q(y,z\theta)\,\mathrm{d}\mu$.

In practice, checking the compactness condition (B) may be intractable. In that case, we have to recourse to a stabilization procedure. We proceed as in Andrieu et al. (2006). Let $(\mathcal{K}_n)_{n\in\mathbb{N}}$ be an exhaustion by compact sets of the space $\mathcal{S}$, *i.e.* be a sequence of compact subsets of $\mathcal{S}$ such that

$$\bigcup_{n\in\mathbb{N}} \mathcal{K}_n = \mathcal{S} \qquad \text{and} \qquad \forall k \in \mathbb{N}, \quad \mathcal{K}_n \subset \mathrm{int}(\mathcal{K}_{n+1})\,,$$

where $\mathrm{int}(A)$ denotes the interior of the set A. The main idea is to reset the sequence s_k to an arbitrary point every time s_k wanders out of the compact subset $\mathcal{K}_{n_k}$, where n_k is the number of projections up to the k-th iteration. Let $\varepsilon = (\varepsilon_k)_{k\in\mathbb{N}}$ be a monotone non-increasing sequence of positive numbers and let K be a subset of $\mathcal{Z}$. Last, let $\Pi\colon \mathcal{Z}\times\mathcal{S} \to K \times \mathcal{K}_0$ be a measurable function (See Andrieu et al. (2006) for details about the way to choose Π). The stochastic approximation with truncation on random boundaries summarizes as:

Algorithm 2: Stochastic approximation with truncation on random boundaries.

Set $n_0 = 0$, $s_0 \in \mathcal{K}_0$ and $\tilde{z}_0 \in K$
for *all* $k \in \mathbb{N}$ **do**
  Sample $\tilde{z}^* \sim \tilde{q}_k(\cdot;\theta_{k-1})$
  Compute $s^* = s_{k-1} + \gamma_k(S(y,\tilde{z}^*) - s_{k-1})$
  if $s^* \in \mathcal{K}_{n_{k-1}}$ **then**
    Set $(\tilde{z}_k, s_k) = (\tilde{z}^*, s^*)$
  else
    Set $(\tilde{z}_k, s_k) = \Pi(\tilde{z}_{k-1}, s_{k-1})$ and $n_k = n_{k-1} + 1$
  end
  Set $\theta_k = \hat{\theta}(s_k)$
end

Note that the statement of Theorem 7.1 is very similar to the corresponding one of Delyon et al. (1999), namely Theorem 5 which establish the convergence of the SAEM. In other words, approximate the posterior distribution in the S-step does not require supplementary considerations to still guarantee the convergence of the sequence $(\theta_k)_{k\in\mathbb{N}}$. Thus,

the scope of application of the approximated-SAEM algorithm is at least as unrestrictive as the one of the SAEM.

The proof of the theorem consists in applying Theorem 2 of Delyon et al. (1999). We recall this theorem in Appendix A of Chapter 6. In particular, (SA0 - 4) refer to their hypothesis (SA0 - 4). Moreover, since it is used in our demonstration, we also recall Lemma 2 from the same paper. For sake of simplicity, we prove the convergence of the approximated-SAEM under the compactness condition (B). However, the result remains true even if (B) is not satisfied, on condition of having recourse to this truncation on random boundaries procedure (Algorithm 2).

Proof: As for all $k \in \mathbb{N}$, $\gamma_k \in [0,1]$, (SA0) is verified under (M1*) and (SAEM1). Moreover, (SA1) is implied by (SAEM1) and (SA3) by (B). Note that under Assumption (B), there exists, with probability 1, a compact set K such that for all $k \in \mathbb{N}$, $s_k \in K$.

Let, for all $s \in \mathcal{S}$ and $k \in \mathbb{N}$, $h(s) = \bar{s}(\hat{\theta}(s)) - s$,

$$\begin{aligned} e_k &= S(y, \tilde{z}_k) - \mathbb{E}\left[S(y, \tilde{z}_k)|\widetilde{\mathcal{F}}_{k-1}\right] \\ \text{and} \qquad r_k &= \mathbb{E}\left[S(y, \tilde{z}_k)|\widetilde{\mathcal{F}}_{k-1}\right] - \bar{s}(\hat{\theta}(s_{k-1})) \end{aligned}$$

such that Equation (7.1) writes on Robbins-Monro type approximation procedure.

As Lemma 2 of Delyon et al. (1999) (see Appendix A) depends only of the meanfield of the model, it can be applied as it is. Thus, (SA2.i) is satisfied with the Lyapunov function $V = -\ell \circ \hat{\theta}$ and

$$\begin{aligned} \{s \in \mathcal{S} | F(s) = 0\} &= \{s \in \mathcal{S} | \partial_s V(s) = 0\}, \\ \hat{\theta}\left(\{s \in \mathcal{S} | F(s) = 0\}\right) &= \{\theta \in \Theta | \partial_\theta \ell(\theta) = 0\} = \mathcal{L} \end{aligned}$$

Moreover, (SA2.ii) is satisfied due to the Sard theorem and (SAEM2). We only need to focus on (SA4).

Set for all $n \in \mathbb{N}^*$, $E_n = \sum_{k=1}^n \gamma_k e_k$. The sequence $(E_n)_{n\in\mathbb{N}^*}$ is a $\widetilde{\mathcal{F}}$-martingale: for all $m > n$, $\mathbb{E}[E_m|\widetilde{\mathcal{F}}_n] = E_n$ as for all $k > n$, $\widetilde{\mathcal{F}}_n \subset \widetilde{\mathcal{F}}_{k-1}$. Moreover, for all $n \in \mathbb{N}$, -

$$\mathbb{E}\left[\left\|S(y, \tilde{z}_{n+1}) - \mathbb{E}\left[S(y, \tilde{z}_{n+1})|\widetilde{\mathcal{F}}_{n+1}\right]\right\|^2 \middle| \widetilde{\mathcal{F}}_n\right] \leqslant \mathbb{E}\left[\|S(y, \tilde{z}_{n+1})\|^2 \middle| \widetilde{\mathcal{F}}_n\right] < \infty \text{ a.s.}$$

since by (B) and (M5), with probability 1, $\hat{\theta}(s_n)$ is in the compact set $\hat{\theta}(K) \subset \Theta$. So,

$$\sum_{n=1}^{\infty} \mathbb{E}\left[\|E_{n+1} - E_n\|^2 \middle| \widetilde{\mathcal{F}}_n\right] \leqslant \sum_{n=1}^{\infty} \gamma_{n+1}^2 \mathbb{E}\left[\|S(\tilde{z}_{n+1})\|^2 \middle| \widetilde{\mathcal{F}}_n\right] < \infty \text{ a.s.}.$$

According to Theorem 2.15 of Hall and Heyde (1980), with probability 1, $\lim_{n\to\infty} E_n$ exists. Moreover,

$$r_n \;=\; \int_Z S(y,z)\left(q(z|y,\hat\theta(s_{n-1})) - \tilde q_n(z,\hat\theta(s_{n-1}))\right)\mathrm{d}\mu(z)$$

for all $n \in \mathbb{N}$, which converge to 0 according to hypothesis (A), proving (SA4).

Thus, Theorem 2 of Delyon et al. (1999) applies and

$$\limsup_{k\to\infty} d(s_k, \{s\in\mathcal{S}|\partial_s V(s)=0\}) \;=\; \limsup_{k\to\infty} d(s_k,\{s\in\mathcal{S}|F(s)=0\}) \;=\; 0\,.$$

Lastly, by continuity of $\hat\theta\colon \mathcal{S}\to\Theta$,

$$\limsup_{k\to\infty} d\left(\hat\theta(s_k), \hat\theta(\{s\in\mathcal{S}|F(s)=0\})\right) \;=\; \limsup_{k\to\infty} d(\theta_k,\mathcal{L}) \;=\; 0\,.$$

□

The obtained results demonstrate that, under appropriate conditions, the sequence $(\theta_k)_{k\in\mathbb{N}}$ converges to a connected component of the set $\mathcal{L}$ of stationary points of ℓ. Moreover, some conditions upon which the convergence toward local *maxima* is guaranteed are given in Section 7 of Delyon et al. (1999). As this conditions only depend on the design of the model and not on the definition of the optimizing sequence $(\theta_k)_{k\in\mathbb{N}}$, the corresponding theorems remain exact in our context leading to classical hypothesis ensuring convergence toward local *maxima*.

I – 2. A Tempering Version of the SAEM

We focus in the following on an instantiation of the approximated-SAEM, leading to the *tempering-SAEM*. Let $(T_k)_{k\in\mathbb{N}}$ be a sequence of positive numbers such that $\lim_{k\to\infty} T_k = 1$. We set, for all $y\in\mathcal{Y}$, all $z\in\mathcal{Z}$, all $\theta\in\Theta$ and all $k\in\mathbb{N}$,

$$\tilde q_k(z;\theta) = \frac{1}{c_\theta(T_k)}\, q(z|y;\theta)^{1/T_k}\,,$$

where $c_\theta(T_k)$ is a scaling constant.

Let $y\in\mathcal{Y}$ and $\mathcal{K}\subset\Theta$ compact. Then, by continuity of the function $\theta\mapsto q(z|y;\theta)$, it exists $M\in\mathbb{R}$ such that

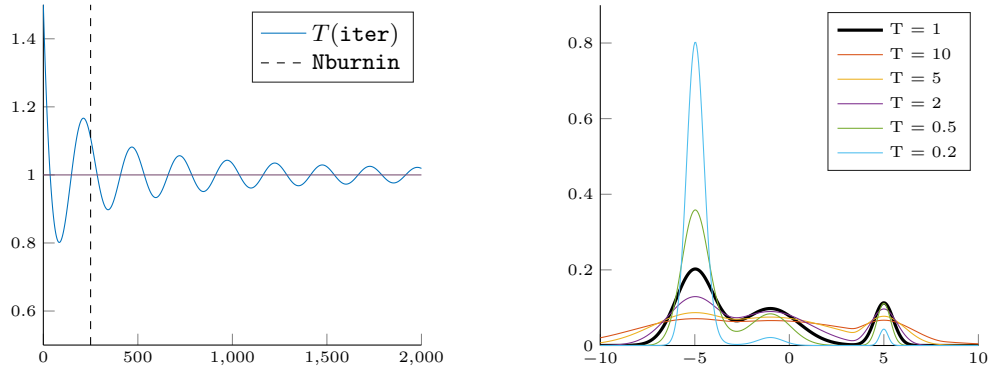
$$\sup_{\theta\in\mathcal{K}} |S(y,z)\,(\tilde q_k(z;\theta) - q(z|y;\theta))| \;\leqslant\; \sup_{\theta\in\mathcal{K}} M\left|1 - \frac{1}{c_\theta(T_k)}\exp\left(-\Big(1-\frac{1}{T_k}\Big)q(z|y,\theta_k)\right)\right|\,.$$

Thus, as $\mathcal{K}$ is compact, (A) is satisfied.

Note that our tempering-SAEM differs from the simulated annealing version of Lavielle and Moulines (1997) as we do not modify the model but only the sampling-step of the estimation algorithm.

a. Escape Local Maxima

This scheme has been built with the intuition of the simulated annealing: the sequence $(T_k)_{k\in\mathbb{N}}$ has to be interpreted as a sequence of temperatures. The higher T_k, the more the corresponding distribution $\tilde{q}_k$ lies flat and the (approximated) hidden variable z_k is able explore all the set $\mathcal{Z}$. On the contrary, a low temperature will freeze the exploration of z_k (see Figure 7.1b). Thus, finding an appropriate sequence $(T_k)_{k\in\mathbb{N}}$ to keep a balance between both behaviors is a great methodological challenge.



(a) Tempering scheme *(b) Tempering distributions*

***Figure 7.1** – Construction of the temperature scheme.*
Fig. 7.1a: Evolution of the temperature over iteration for the tempering-SAEM.
Fig. 7.1b: Influence of the temperature over the pattern of the distribution.

We propose here an oscillatory tempering pattern which oscillate around one with decreasing amplitude. In other words given the decreasing and amplitude rate a and b, the scaling-parameter r and the delay c, we define our sequence of temperatures by: for all $k \in \mathbb{N}$,

$$T_k \;=\; 1 + a^{\kappa} + b\,\frac{\sin(\kappa)}{\kappa}\,, \qquad \text{where} \qquad \kappa \;=\; \frac{k + c \times r}{r}\,.$$

We design this scheme to decrease, with an exponential rate toward 1, with dampened oscillations. In this form, the tempering scheme includes the tempering scheme used by Monolix. Even so, the experiments conducted in section II tend to show that the exponential decrease is not necessary. In particular, a is set to zero for all the experiments.

In order the tempering scheme to converge toward 1, we just need to require that the exponential rate $a \in [0, 1[$. In particular, the parameter b can be chosen independently negative or positive. A positive b will flattened the distribution at the beginning of the optimization procedure. On the contrary, a negative b will make the profile of the distribution look more prickly. It can be interesting to enforce the distinction of two close modes, as in Section II.1.b.

Due to the oscillations of the temperature, the latent variable z_k will explore and gather in turns. Thus, in case of multimodal density, the latent variable will be able to

switch from one mode to an other during the heating steps and to explore these same modes during the cooling phases. In particular, during the optimization, the tempering-SAEM may escape from local minima in which the SAEM would get stuck. Figures 7.8 and 7.9 effectively illustrate this phenomenon. In this way, the local *maxima* of the likelihood can be avoided. Moreover, as the approximated distributions regularly gather around the modes of the posterior distribution $q(\cdot|y;\theta_k)$, the exploration of z will stabilize and the algorithm will converge.

Although the analysis of this algorithm is heuristic, the simulations (see the following section) confirm the intuition and give good results. A theoretical analysis is an ongoing problem.

II. Application and Experiments

As explained in the previous paragraph, the tempering-SAEM allows us to escape from local *maxima*. To illustrate this phenomenon, we propose two applications: cluster analysis through Gaussian mixture model and independent factor analysis which can lead to blind source separation (Allassonnière and Younes, 2012; Attias, 1999; Moulines et al., 1997).

II – 1. Multivariate Gaussian Mixture Models

Before considering a more realistic application, we first present an application of the tempering-SAEM to multivariate Gaussian mixture model (GMM). Actually, in spite of an apparent simplicity, this model illustrates well the main features of our algorithm.

Let $y = (y_i)_{i\in[\![1,n]\!]} \in \mathbb{R}^{nd}$ be a n-sample of $\mathbb{R}^d$. We assume that y is distributed under a weighted sum of m d-dimensional Gaussians: Given the weights $\alpha = (\alpha_j)_{j\in[\![1,m]\!]}[0,1]^m$ such that $\sum_{j=1}^m \alpha_j = 1$, the centroids $\mu = (\mu_j)_{j\in[\![1,m]\!]} \in \mathbb{R}^{md}$ and the covariance matrices $\Sigma = (\Sigma_j)_{j\in[\![1,m]\!]} \in (\mathcal{S}_d\mathbb{R})^m$, we assume that

$$y|z,\theta \;\sim\; \bigotimes_{i=1}^{n} \mathcal{N}\left(\mu_{z_i}, \Sigma_{z_i}\right) \qquad \text{and} \qquad z|\theta \;\sim\; \sum_{j=1}^{m} \alpha_j \delta_j \,,$$

where $\theta = (\alpha, \mu, \Sigma)$ and $z = (z_i)_{i\in[\![1,n]\!]}$ is the latent variable specifying the identity of the mixture component of each observation. In the following, we compare the efficiency of the EM, the SAEM and the tempering-SAEM algorithms to produce a *maximum* likelihood estimate of the parameters with the *a priori* given exact number of components m.

Classically, as closed-form expressions are possible for finite GMM, the EM algorithm is a very popular technique used to produce the *maximum* likelihood estimation of the parameters (McLachlan and Peel, 2000). However, the computational cost can be prohibitive. A faster procedure is to use the SAEM algorithm. Nevertheless, both

algorithms are very sensitive to its initial position: solutions can highly depend on their starting point and consequently produce sub-optimal *maximum* likelihood estimates (Biernacki et al., 2003). The tempering-SAEM appears as a way to escape from local *maxima* and reach global *maxima* more often.

a. Insensitivity of the tempering-SAEM to Initialization

To estimate the sensitivity of the tempering-SAEM algorithm to its initial position, we generate a synthetic dataset (Figure 7.2) and perform the estimation 500 times for the three algorithms, with the same sequences of points chosen at random within the dataset.

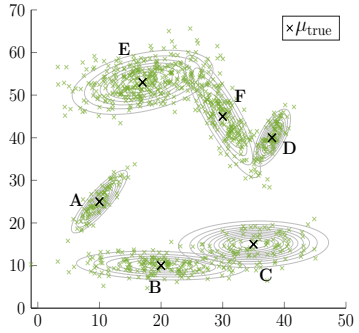
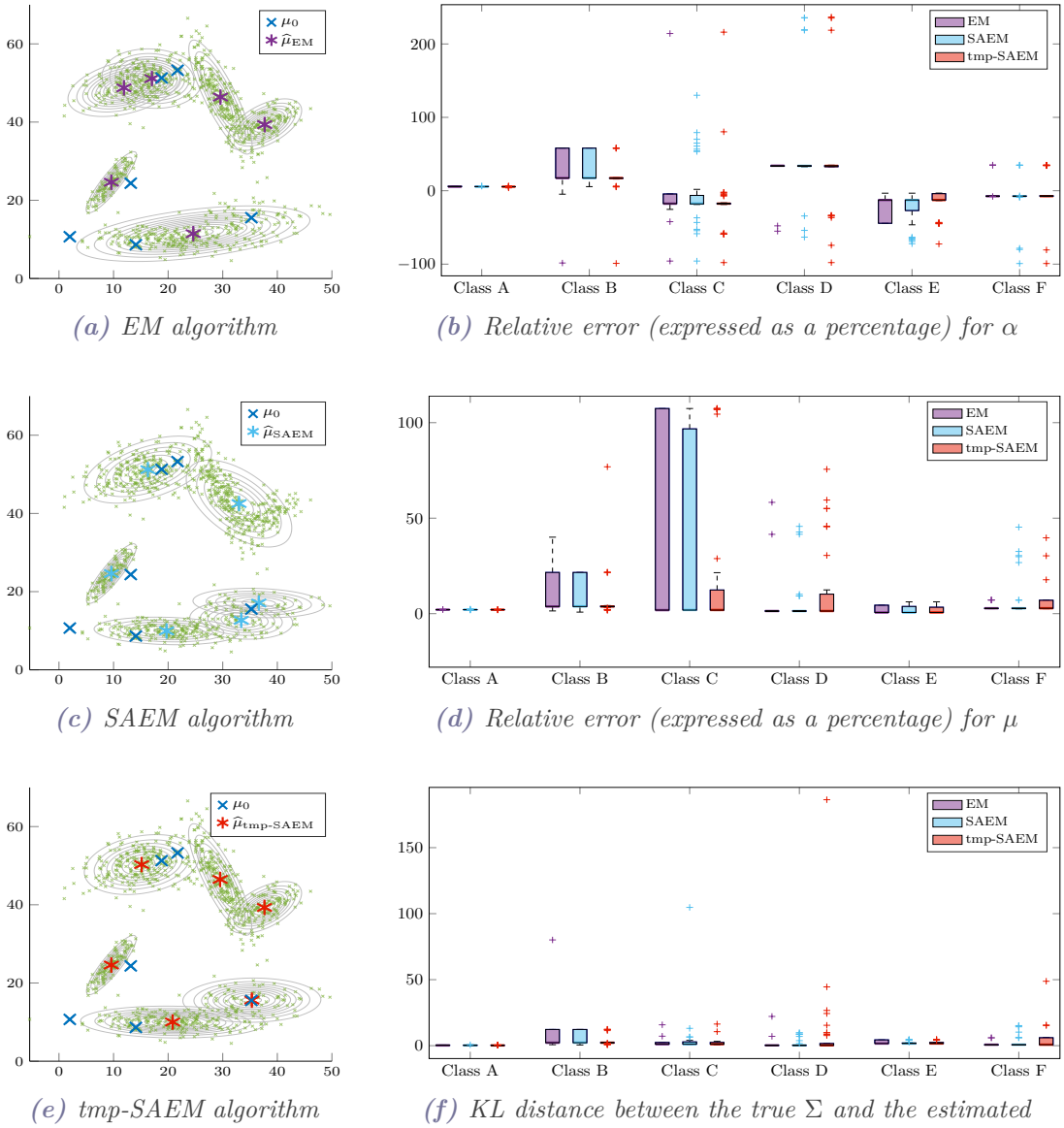


***Figure 7.2** – Learning dataset.*
Synthetic dataset used to perform the experience regarding Section II.1.a.

As in Chapter 5, the Kullback-Leibler divergence between two covariance matrices refers to the one between the corresponding centered Gaussian distributions. L The relative errors for α and μ and the Kullback-Leibler divergence between the true covariance matrices Σ and the estimated one are compiled in Figures 7.3b, 7.3d and 7.3f. The class refer to the ones of Figure 7.2. We consider the algebraic relative error for α so that we can deduce if the studied algorithm tend to empty (class E) or overfill (class B) the classes. First, the tempering-SAEM is always competitive with the EM and the SAEM and most of the time greater. In other words, the global *maximum* is more often reached while tempering the posterior distribution. Moreover, while EM and SAEM achieve fairly identical results, the tempering-SAEM is able to discriminate overlapped classes. Figures 7.3a, 7.3c and 7.3e displays the result of a type run for each of the three algorithms, with the same initial points (the blue crosses). Class A, which is the only isolated class, is seemingly the best learned. The EM and SAEM seem to empty the class C for the benefit of the class B and merge them together on a "super-class" as if there were only 5 components in the Gaussian mixture.

The three procedures are detailed in Appendix A.



(a) EM algorithm *(b) Relative error (expressed as a percentage) for α*

(c) SAEM algorithm *(d) Relative error (expressed as a percentage) for μ*

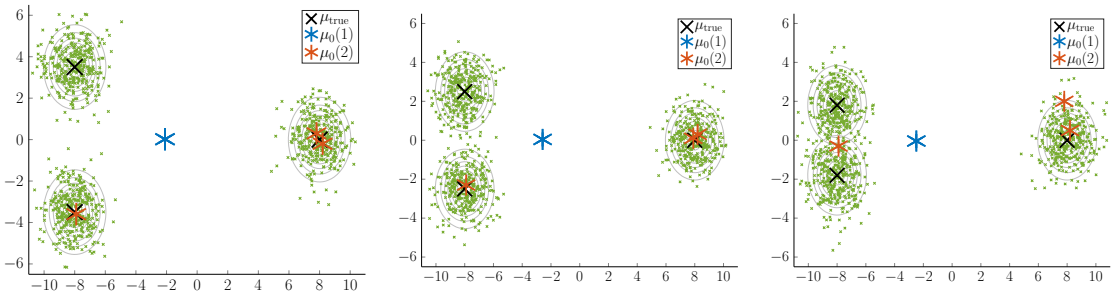
(e) tmp-SAEM algorithm *(f) KL distance between the true Σ and the estimated*

***Figure 7.3** – Multivariate Gaussian mixture model.*
Figs. 7.3a, 7.3c and 7.3e: Qualitative comparison of the *maximum* likelihood estimation of the parameters. The estimation is performed with the same initial points (in orange). Figs. 7.3b, 7.3d and 7.3f: Relative error (expressed as a percentage) for the weights α and the centroïds μ. Kullback-Leibler distance between the true covariance matrices Σ and the estimated ones, for 500 runs and $n = 1000$.

b. Escaping Local Minima

We then consider a situation known to be badly managed by the EM algorithm. Namely, we consider a three clusters' dataset. One cluster is on the far right side, two are on

the far left side and all the three clusters are equiprobable. Moreover, we want to study the influence of the distance between the two left clusters. So we build three datasets: one where the two left clusters are properly distinct, one where they are close and a last one where they are almost merged. The three datasets are displayed at Figure 7.4. For each dataset, we perform the optimization for two different initial positions, referred as initialization 1 and 2 in the following. In the first case, the three centroïdes μ are initialized at the barycenter of the observed data. In the second one, we initialize two means on the right side and one on the left side.



(a) Dataset I *(b) Dataset II* *(c) Dataset III*

***Figure 7.4** – Three scattered clusters' datasets*
The datasets used to perform the experiences regarding Section II.1.b. For each of them, we consider two possible initial positions for the means μ: either all at the barycenter of the dataset (the blue asterisk) or two of them in the single right cluster and the last mean on the left side (the orange asterisks).

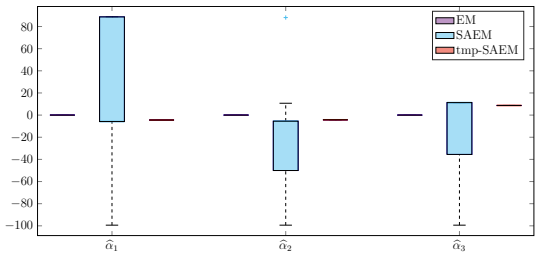
For each situation, we perform the estimation through all the three algorithms. We present at Tables 7.1, 7.2 and 7.3 the relative errors for the different parameters. As previously, we consider the algebraic relative error for the weights α. To better understand the behavior of the different algorithms, we also provide a box plot of these relative errors at Figures 7.5 a-b, 7.6 a-b and 7.7 a-b. The SAEM algorithm tends to empty classes for the benefit of other(s). It seems to be less the case for the tempering-SAEM. Note that, whatever the dataset, if the mean parameters are initialized to the mean of the dataset, the EM algorithm does not move. Thus, the error concerning the mixture proportion α seems to be very small, but this is only due to the initialization of the parameters α in favor to an equiprobable mixture.

Figures 7.5 c-d, 7.6 c-d and 7.7 c-d display the average of the estimated $\widehat{\mu}$ and $\widehat{\Sigma}$ by the three algorithms (EM vs SAEM vs tempering-SAEM), for each dataset and each initial position.

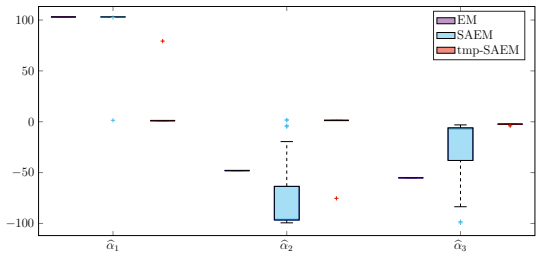
The tempering-SAEM succeed to accurately estimate all the parameters related to the first and second datasets (Figure 7.5, Figure 7.6, Table 7.1 and Table 7.2), even when two of the mean parameters μ are initialized in the right cluster (Initialization 2). When the parameters μ are initialized to the barycenter of the dataset, the tempering-SAEM still accurately estimates the different parameters, including for the dataset III where

***Table 7.1** – Quantitative performance of the estimation for the **dataset I**.*
Mean (standard deviation) relative errors (expressed as a percentage) for the estimated parameters of the GMM within the dataset I. Over 100 runs.

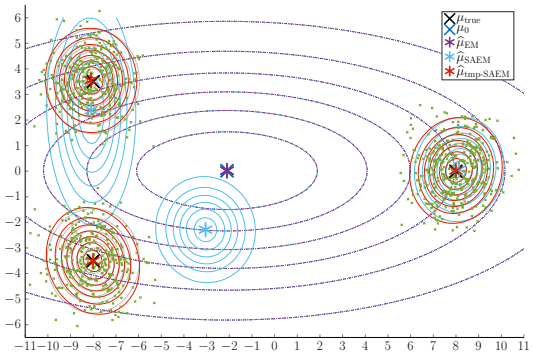
	EM - 1	**EM** - 2	**SAEM** - 1	**SAEM** - 2	**tmp-SAEM** - 1	**tmp-SAEM** - 2
$\widehat{\alpha}_1$	0.00	103.10	24.83 (46.24)	99.46 (18.99)	−4.46 (0.00)	2.01 (8.54)
$\widehat{\alpha}_2$	0.00	−48.02	−19.41 (25.89)	−77.72 (28.72)	−4.23 (0.00)	0.39 (8.35)
$\widehat{\alpha}_3$	0.00	−55.08	−5.42 (26.22)	−21.87 (24.94)	8.69 (0.00)	−2.40 (0.18)
$\widehat{\mu}_2$	78.46	39.44	14.28 (18.28)	38.07 (7.15)	1.24 (0.00)	1.62 (4.18)
$\widehat{\mu}_2$	78.46	185.93	58.54 (84.14)	168.49 (37.28)	0.17 (0.00)	2.56 (17.03)
$\widehat{\mu}_3$	126.23	0.73	2.86 (4.06)	2.94 (4.40)	0.34 (0.00)	1.03 (0.01)
$\widehat{\Sigma}_1$	1503.22	306.00	104.94 (216.86)	295.19 (31.90)	0.99 (0.00)	7.08 (33.01)
$\widehat{\Sigma}_2$	1503.22	7.26	19.04 (98.85)	18.85 (5.96)	4.78 (0.00)	2.16 (2.41)
$\widehat{\Sigma}_3$	1503.22	8.90	5.19 (0.20)	6.07 (0.21)	2.35 (0.00)	1.52 (0.27)



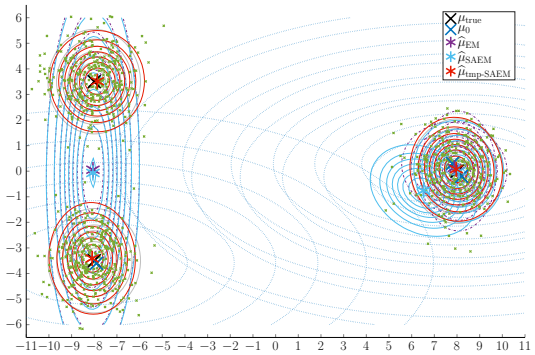
(a) Init. 1 – Relative error for α (in %)



(b) Init. 2 – Relative error for α (in %)



(c) Init. 1 – Qualitative performance

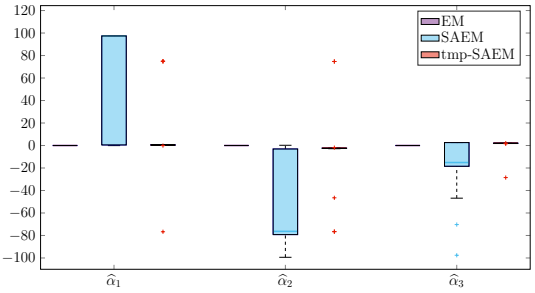


(d) Init. 2 – Qualitative performance

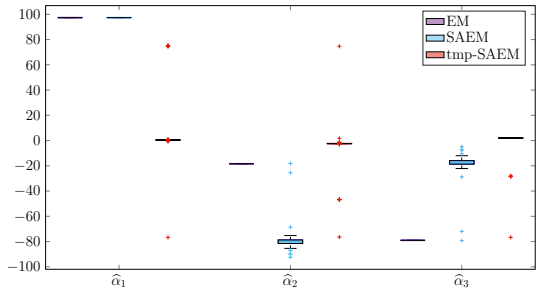
***Figure 7.5** – Performance of the estimation for the **dataset I**.*
Average centroïdes and covariance matrices estimated by the EM, the SAEM and the tempering-SAEM algorithms within the dataset I, according to the initialization. In purple dashed lines, the covariance matrix estimated by the EM; in blue plain lines, the one estimated by the SAEM and in bold red lines the one estimated by the tempering-SAEM. In dotted blue lines the initial covariance matrices associated to the two different initial means (the blue crosses).
Tempering scheme: $a = 0$, $b = -10$, $c = 2$, $r = 10$.

***Table 7.2** – Quantitative performance of the estimation for the **dataset II**.*
Mean (standard deviation) relative errors (expressed as a percentage) for the estimated parameters of the GMM within the dataset II. Over 100 runs.

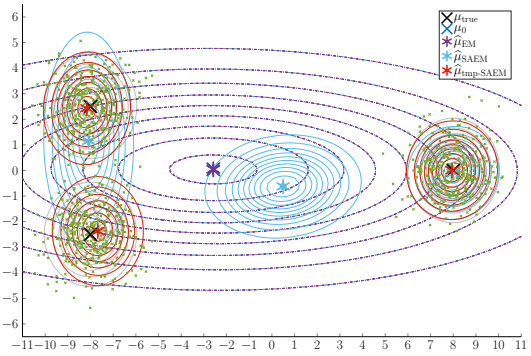
	EM - 1	**EM** - 2	**SAEM** - 1	**SAEM** - 2	**tmp-SAEM** - 1	**tmp-SAEM** - 2
$\widehat{\alpha}_1$	0.00	97.4	55.16 (48.27)	97.4 (0.00)	0.34 (19.25)	3.81 (18.33)
$\widehat{\alpha}_2$	0.00	−18.36	−44.85 (38.38)	−79.44 (7.78)	−2.01 (18.19)	−3.67 (12.17)
$\widehat{\alpha}_3$	0.00	−79.04	−10.31 (15.68)	−17.96 (7.78)	1.67 (3.15)	−0.14 (10.95)
$\widehat{\mu}_2$	70.97	29.12	17.05 (13.80)	29.12 (0.00)	3.34 (7.31)	3.21 (6.94)
$\widehat{\mu}_2$	70.97	192.96	104.28 (92.43)	187.28 (1.11)	5.31 (25.61)	9.47 (39.60)
$\widehat{\mu}_3$	132.29	13.28	2.15 (1.82)	2.94 (1.09)	0.79 (0.48)	1.45 (5.57)
$\widehat{\Sigma}_1$	1438.21	154.34	88.00 (57.50)	154.34 (0.00)	7.81 (27.00)	10.60 (34.47)
$\widehat{\Sigma}_2$	1438.21	10.58	44.17 (608.01)	13.68 (0.01)	7.28 (27.00)	3.48 (13.66)
$\widehat{\Sigma}_3$	1438.21	13.60	7.64 (0.13)	10.12 (0.00)	4.14 (0.90)	4.63 (2.97)



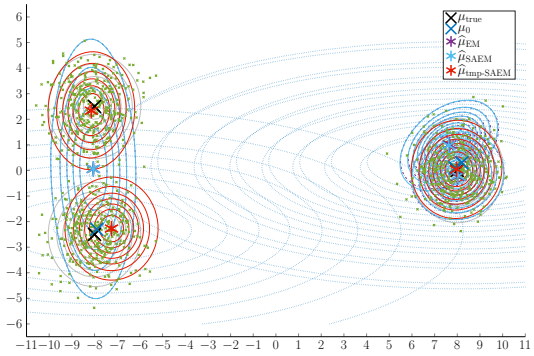
(a) Init. 1 – Relative error for α (in %)



(b) Init. 2 – Relative error for α (in %)



(c) Init. 1 – Qualitative performance

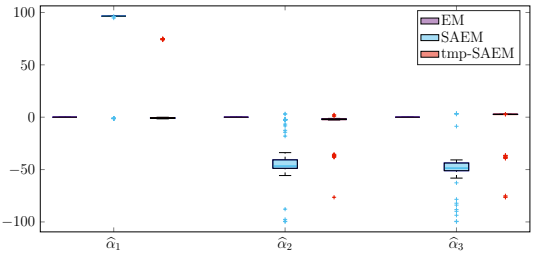


(d) Init. 2 – Qualitative performance

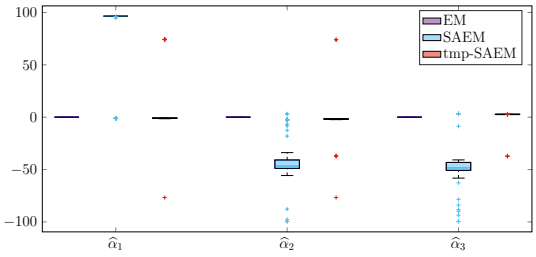
***Figure 7.6** – Performance of the estimation for the **dataset II**.*
Mean centroïdes and covariance matrices estimated by the EM, the SAEM and the tempering-SAEM algorithms within the dataset II, according to the initial position of the means. Same conventions as for the previous figure.
Tempering scheme: $a = 0$, $b = -4.5$, $c = 1$, $r = 4.8$.

***Table 7.3** – Quantitative performance of the estimation for the **dataset III**.*
Mean (standard deviation) relative errors (expressed as a percentage) for the estimated parameters of the GMM within the dataset III. Over 100 runs.

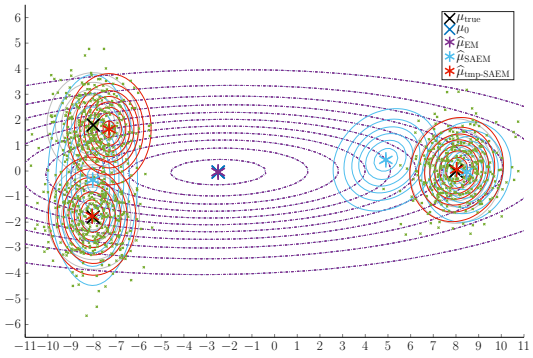
	EM - 1	**EM** - 2	**SAEM** - 1	**SAEM** - 2	**tmp-SAEM** - 1	**tmp-SAEM** - 2
$\widehat{\alpha}_1$	0.00	96.5	82.77 (34.01)	94.91 (12.40)	2.99 (22.04)	68.43 (20.77)
$\widehat{\alpha}_2$	0.00	−24.85	−41.62 (23.97)	−44.71 (18.13)	−4.64 (19.36)	−33.88 (9.76)
$\widehat{\alpha}_3$	0.00	−71.65	−41.15 (25.54)	−50.20 (18.62)	1.65 (6.42)	−34.55 (11.24)
$\widehat{\mu}_2$	70.56	20.82	18.11 (6.74)	20.51 (2.46)	3.58 (7.16)	19.26 (5.28)
$\widehat{\mu}_2$	70.56	196.07	158.34 (69.90)	187.25 (24.20)	9.84 (38.15)	174.04 (51.80)
$\widehat{\mu}_3$	131.10	5.64	6.46 (4.50)	7.43 (3.48)	0.95 (1.04)	7.10 (1.93)
$\widehat{\Sigma}_1$	1451.58	87.14	75.49 (8.35)	85.79 (1.10)	10.40 (21.34)	80.38 (22.82)
$\widehat{\Sigma}_2$	1451.58	5.51	29.18 (194.51)	12.84 (1.76)	6.42 (11.62)	11.61 (2.48)
$\widehat{\Sigma}_3$	1451.58	6.99	7.07 (0.21)	7.92 (0.21)	3.06 (0.67)	7.49 (1.49)



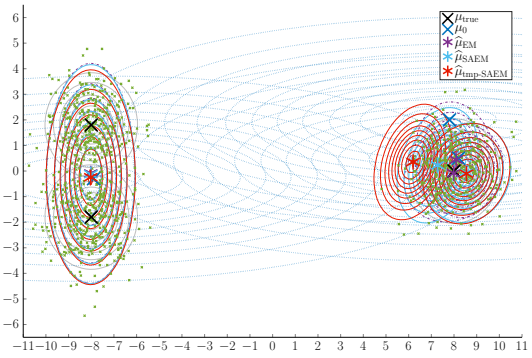
(a) Init. 1 – Relative error for α (in %)



(b) Init. 2 – Relative error for α (in %)



(c) Init. 1 – Qualitative performance



(d) Init. 2 – Qualitative performance

***Figure 7.7** – Performance of the estimation for the **dataset III**.*
Mean centroïdes and covariance matrices estimated by the EM, the SAEM and the tempering-SAEM algorithms within the dataset III, according to the initial position of the means. Same conventions as for the previous figure.
Tempering scheme: $a = 0$, $b = -4.7$, $c = 1$, $r = 5$.

the left clusters are merged (Figure 7.7c). However, when two of the mean parameters are initialized within the single right cluster, the tempering-SAEM does not succeed to capture the two left classes if the left clusters are to close (Figure 7.7d), but this can be easily explained by the distribution of the observations (Figure 7.4c). Still concerning initialization 2 and dataset 3, the tempering-SAEM is nevertheless at least competitive with the SAEM algorithm. Most interesting behavior: even though the tempering-SAEM does not always explain the whole distribution, the relative error may fall to zero with the tempering-SAEM, whereas it is never the case for the SAEM algorithm. In other words, the tempering-SAEM favor the convergence toward global maxima, and may succeed to almost surely reach them as in the first dataset, that is exactly the expected behavior of this algorithm.

Last, we present at Figures 7.8 and 7.9 the evolution of the means and their associated covariance matrices. The lines 6 and 7 of Figure 7.8 illustrate the capacity of the tempering-SAEM to distinguish two close classes. On the contrary, the SAEM algorithm does not seem to be able to do so and remains trapped in a local minimum.

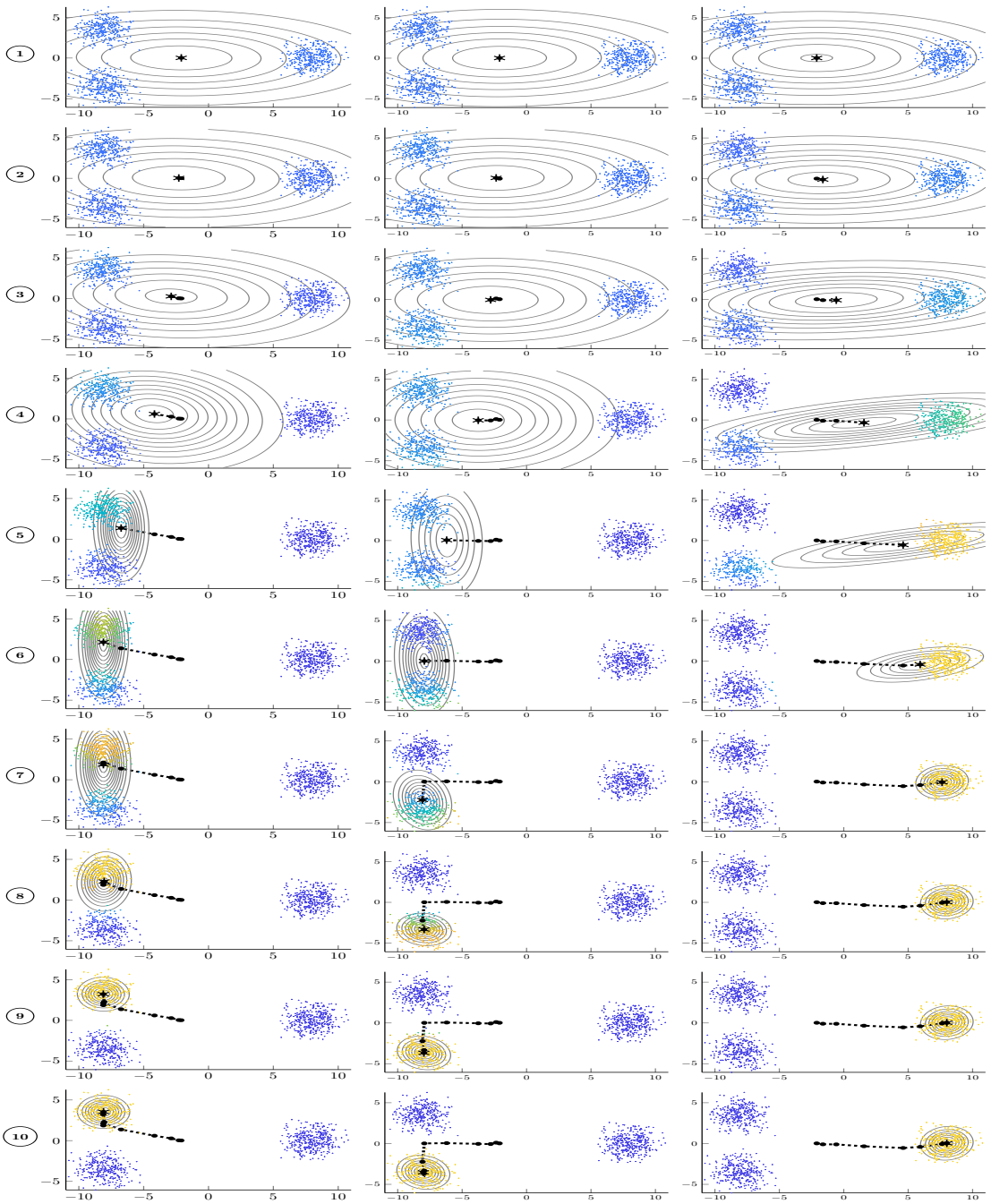
This experiment also highlights the benefits of an oscillating temperature scheme compared to only warm up the conditional distribution throughout the first iterations. Indeed, initialized the centroïds to the mean of the dataset may be interpreted as the limit case of heating the conditional for the first iterations. However, our experiments show that the SAEM initialized at the mean (Initialization 1) behave less well than the tempering-SAEM whatever the initialization.

II – 2. Independent Factor Analysis

The decomposition of a sample of multi-variable data on a relevant subspace is a recurrent problem in many different fields from source separation problem in acoustic signals to computer vision and medical image analysis. Independent component analysis has become one of the standard approaches. This technique relies upon a data augmentation scheme, where the (unobserved) input are viewed as the missing data. We observe multivariable data y which are measured by n sensors and supposed to arise from m source signals x, that are linearly mixed together by some linear transformation H, and corrupted by an additive Gaussian noise ε. Simply put, we observe $y = (y^{(t)})_{t\in[\![1,T]\!]}$, where each measurement is a point of $\mathbb{R}^n$ and assumed to be given by $y^{(t)} = Hx^{(t)} + \varepsilon^{(t)}$, where $H \in \mathcal{M}_{n,m}\mathbb{R}$, $x^{(t)} \in \mathbb{R}^m$ and $\varepsilon^{(t)} \overset{i.i.d}{\sim} \mathcal{N}(0, \lambda I_n)$, $\lambda \in \mathbb{R}$. The suitability of the SAEM algorithm in this context has been demonstrated in Moulines et al. (1997) and Allassonnière and Younes (2012). We propose here to modify the learning principle to make the procedure less susceptible to trapping states.

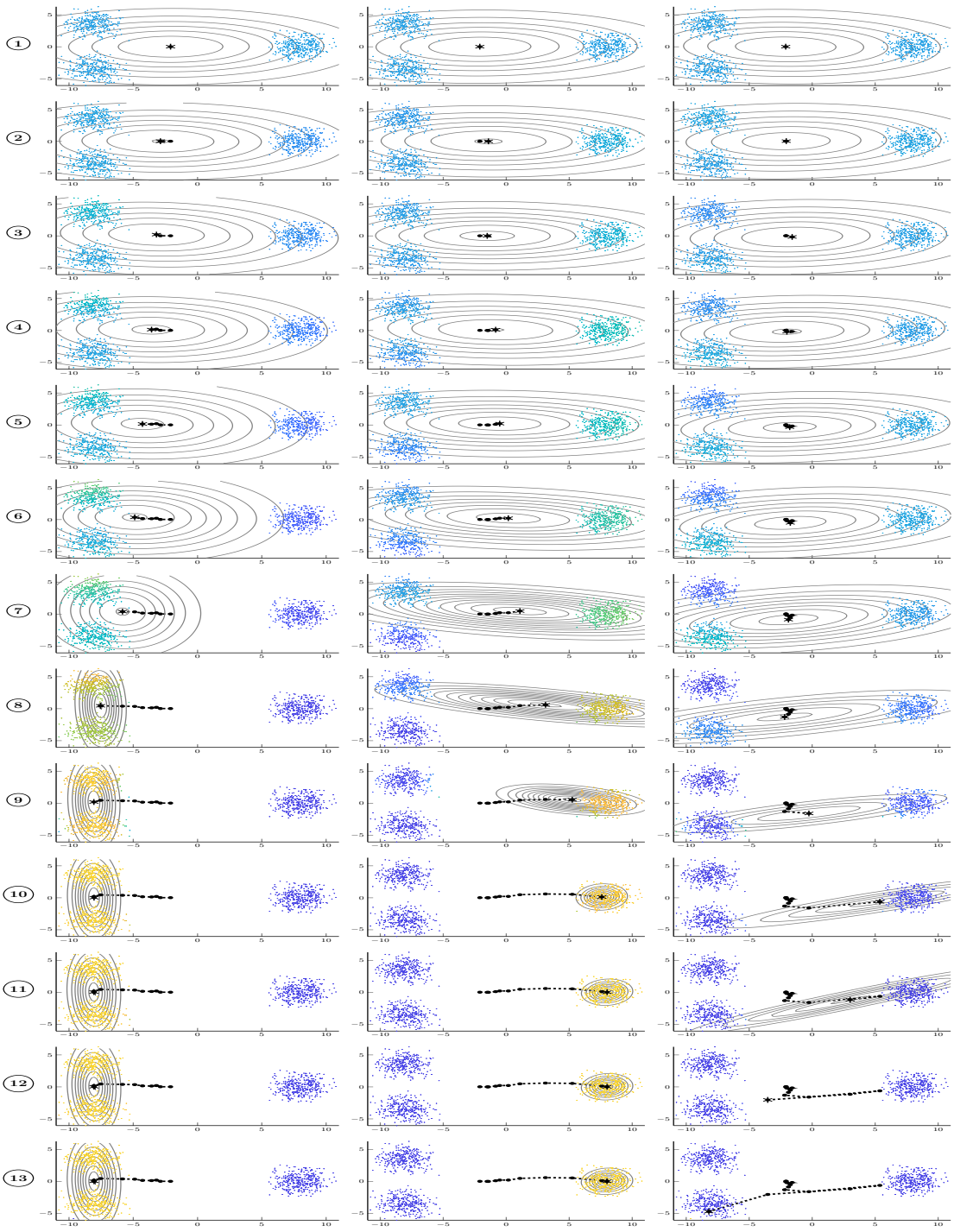
As in Moulines et al. (1997) and Attias (1999), we assume that:

1. $(x^{(t)})_{t\in[\![1,T]\!]}$ and $(\varepsilon^{(t)})_{t\in[\![1,T]\!]}$ are independent;
2. $(x^{(t)})_{t\in[\![1,T]\!]}$ is an i.i.d sequence of random vectors, with independent component.



(a) Classe A *(b) Classe B* *(c) Classe C*

***Figure 7.8** – Estimation by the tempering-SAEM algorithm.*
Evolution of the parameters throughout the estimation by the tempering-SAEM algorithm, within the dataset I, with initial means at the barycenter of the dataset, for a type run. For each class, we plot the trajectories of the mean μ and the evolution of the associated covariance. The observed data are colored according to their probability to belonging to the classes: from 0 in dark blue to 1 in yellow.



(a) Classe A & B *(b) Classe C* *(c) Empty classe*

***Figure 7.9** – Estimation by the SAEM algorithm.*
Evolution of the parameters throughout the estimation by the SAEM algorithm, within the dataset I, with initial means at the barycenter of the dataset, for a type run. We keep the conventions of the previous figure.

Each component $x_i^{(t)}$ is given by a mixture of k Gaussians indexed by $z_i^{(t)} \in [\![1,k]\!]$ with means $\mu_{z_i^{(t)}}$, variances $\sigma^2_{z_i^{(t)}}$ and mixing proportions $\alpha_{z_i^{(t)}}$:

$$\begin{aligned} q(x_i^{(t)};\theta_i^{(t)}) &= \sum_{z_i^{(t)}=1}^{k} \alpha_{z_i^{(t)}}\, \mathcal{G}\left(x_i^{(t)} - \mu_{z_i^{(t)}}; \sigma^2_{z_i^{(t)}}\right) , \\ \theta_i^{(t)} &= \left(\alpha_{z_i^{(t)}}, \mu_{z_i^{(t)}}, \sigma^2_{z_i^{(t)}}\right) , \end{aligned}$$

where for all vectors x and μ and all symmetric matrix Σ, $\mathcal{G}(x-\mu,\Sigma)$ refers to the (multivariate) Gaussian distribution.

This model is called independent factor analysis (IFA). The problem is to find the value of the parameter $W = (H, \lambda, \theta)$ given y. Identifiability in this model is discussed in Comon (1994). Basically, the sources are defined only to within an order permutation and scaling. To avoid trivialities, we fix the variances $(\sigma_j^2)_{j\in[\![1,k]\!]}$ to one (Allassonnière and Younes, 2012). Note that this definition of the IFA model is somewhat less general that the one introduced by Attias (1999) in which the components are supposed to be independent but not necessarily identically distributed. Nevertheless, it has been shown that restrictive IFA models can perform well in practice (Allassonnière and Younes, 2012).

The likelihood of the IFA can be put in exponential form using the sufficient statistics, for all $j \in [\![1,k]\!]$,

$$\begin{aligned} S_{1,j}(x,y,z) &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{z_i=j\}} \;; & S_4(x,y,z) &= y\,{}^t y \;; \\ S_{2,j}(x,y,z) &= \frac{1}{m}\sum_{i=1}^{m} x_i\, \mathbb{1}_{\{z_i=j\}} \;; & S_5(x,y,z) &= y\,{}^t x \;; \\ S_{3,j}(x,y,z) &= \frac{1}{m}\sum_{i=1}^{m} x_i^2\, \mathbb{1}_{\{z_i=j\}} \;; & S_6(x,y,z) &= x\,{}^t x \;. \end{aligned}$$

The M-step is then given by

$$H = [S_5]\,([S_6])^{-1} \;; \qquad \alpha = [S_1] \;; \qquad \mu = \frac{[S_2]}{[S_1]} \;; \qquad \sigma^2 = \mathbf{1}_k \;;$$
$$\lambda = \|\,[S_6]\,\|_2^2 - 2\left\langle H \,\middle|\, [S_5] \right\rangle + \left\langle {}^t H H \,\middle|\, [S_6] \right\rangle ,$$

where $\mathbf{1}_k$ stands for the k-vector off all 1 and the brackets denote the empirical-average. Moreover, it is possible to compute the conditional distribution of the hidden variable (x,z) given observed values of y and the E-step can be computed exactly (Attias, 1999): For all $\zeta \in [\![1,k]\!]^m$,

$$\mathbb{P}(z=\zeta|y;W) = \frac{\alpha_\zeta \mathcal{G}\left(y - H\mu_\zeta; H\Delta_\zeta\,{}^t H + \lambda I_n\right)}{\sum_z \alpha_z \mathcal{G}\left(y - H\mu_z; H\Delta_z\,{}^t H + \lambda I_n\right)}$$

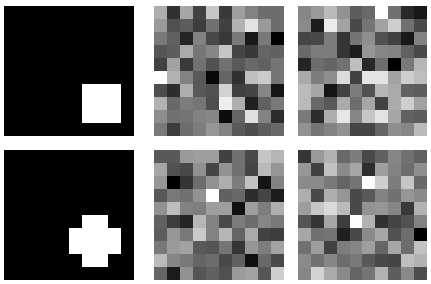
$$\text{and} \qquad q(x|y,z;W) = \mathcal{G}\left(x - \nu_{y,z}; \Sigma_z\right) ,$$

where

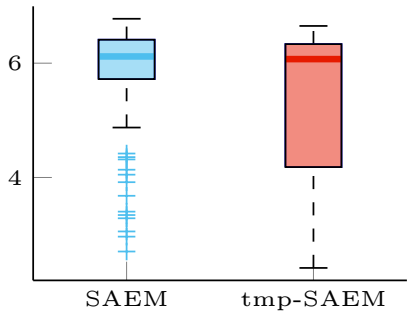
$$\alpha_z = \prod_{i=1}^{m} \alpha_{z_i} \; ; \qquad \mu_z = (\mu_{z_i})_i \; ; \qquad \Delta_z = \mathcal{D}iag\left((\sigma_{z_i}^2)_i\right) \; ;$$
$$\Sigma_z = \left(\frac{1}{\lambda}\, {}^tHH + \Delta_z^{-1}\right)^{-1} \; ; \qquad \nu_{y,z} = \Sigma_z\left(\frac{1}{\lambda}\, {}^tHy + \Delta_z^{-1}\mu_z\right) .$$

Thus, as well as for the GMM, we can compare the efficiency of SAEM vs tempering-SAEM algorithms in this context.

In Section II.1, we were interested in the performance of our algorithm for data generated according to the true model. We relax here this assumption and observe $T = 100$ images distributed in accordance with the Bernoulli-Gaussian model (BG-ICA, Allassonnière and Younes (2012)), with two components. The components are represented as two-dimensional binary images. The first one is a black image with a white cross in the top left corner. The second one has a white square in the bottom right corner. At Figure 7.10, we present the two decomposition images, 4 typical observations and the renormalized $\mathbb{L}^2$ norm between the true H (in the BG-ICA model) and the estimated one for 100 runs.



(a) Decomposition images (left) and samples of the dataset.



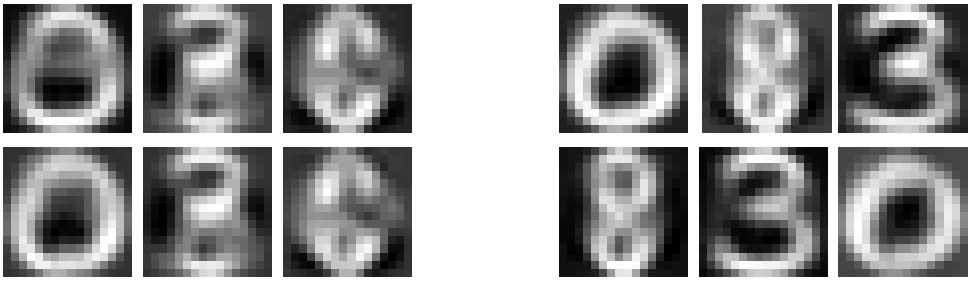
(b) $\mathbb{L}^2$ *norm.*

Figure 7.10 *– IFA estimation within images distributed according to BG-ICA.* Renormalized $\mathbb{L}^2$ norm between the source matrix H used to build the dataset and the estimated one. The dataset consists of 100 images distributed in accordance with the two-components Bernouilli-Gaussian model build from the square and the cross binary images.

This experience confirms the robustness of the tempering-SAEM. Moreover, one could have feared that the augmentation of the number of hyper-parameters due to the choice of the temperature scheme would increase the variance. Figure 7.10 eliminates this assumption. However, the context is very favorable to the SAEM algorithm which obtain very good and hard to outperformed results. To measure the efficiency of the tempering-SAEM, we test it on the USPS database, which contains gray-level images of handwritten digits.

We consider a balanced mix of the digits 0, 3 and 8, which consists of 50 samples for each of the three digits. We then run both the SAEM and the tempering-SAEM.

We present at Figure 7.11 two typical runs (in line). If the two of them succeed in discriminate 0 against 3 and 8, the tempering-SAEM outperform the SAEM algorithm concerning 3 versus 8. Thus, the tempering-SAEM produces meaningful sources, which could be the result of a clustering procedure, while the SAEM runs into difficulties. Hence, this experience suggests that the tempering-SAEM can indeed escape from local *maxima* in which the SAEM can be trapped.



(a) SAEM *(b) tempering-SAEM*

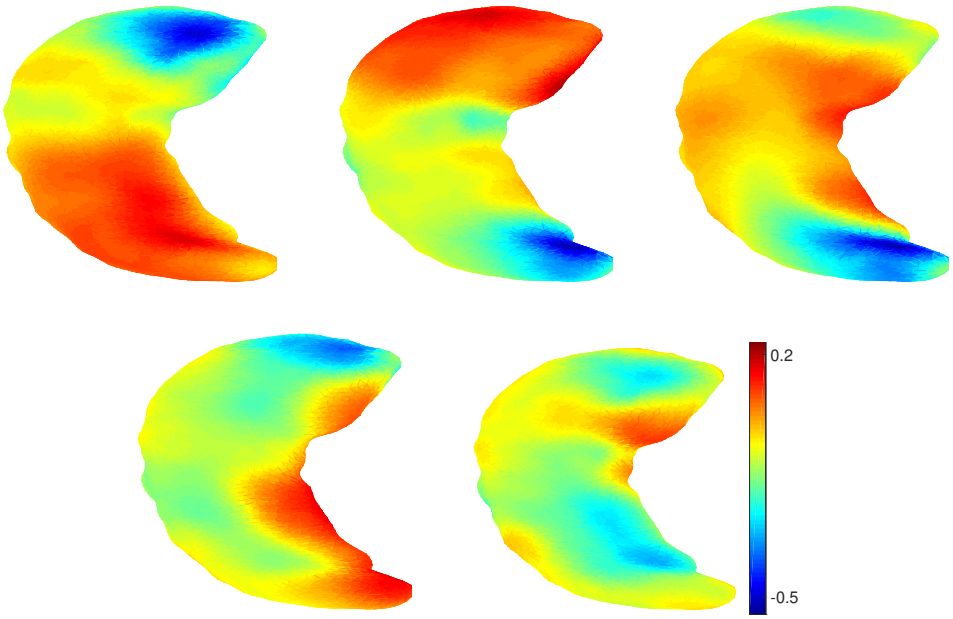
***Figure 7.11** – Independent factor estimation within the USPS dataset.*
Results of the independent factor estimation on a balanced mix of digits 0, 3 and 8 from the USPS database. The dataset is composed of 50 samples of each digits.

Last, we consider a dataset consisting of 101 hippocampi surfaces. The subjects of the dataset can be split in three groups of size 57, 32 and 12 respectively. The first group corresponds to healthy patients; the next two groups correspond to patients with Alzheimer's disease, at two stages of advancement (mild and advanced). Over each hippocampus, a scalar field represent the deformation of the considered hippocampus regarding a template one. Thus we can study the diversity of atrophy patterns, depending on the patient's state of health. We have computed $m = 5$ decomposition vectors based on the complete data set. Figure 7.12 presents these decomposition vectors mapped on the meshed hippocampus for both SAEM and tempering-SAEM algorithms. For comparison purpose, we enforce the same colorbar for both experiments and all hippocampi. Then, it seems that the two algorithms behave in much the same way, at least visually. This experiment attests to the reliability of the tempering-SAEM.

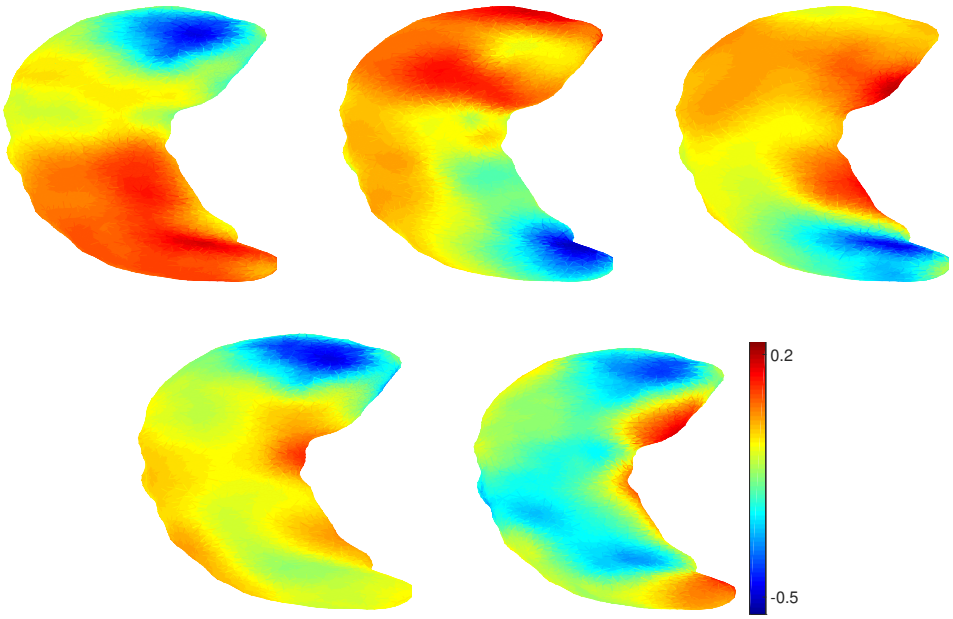
***Table 7.4** – p-values obtained from the IFA estimation of the hippocampi dataset.*
Mean (standard deviation) of the p-values for the five decomposition vectors presented at Figure 7.12, over 50 runs.

		SAEM	tmp-SAEM
Groups 1 vs 2&3	$10^{-3}\times$	0.43 (0.31)	0.37 (0.27)
Groups 1 vs 2	$10^{-3}\times$	11.76 (7.43)	11.33 (7.18)

At Table 7.4, we provide the p-values obtained from the comparison of the five columns of H among the three subgroups. The test is based on a Hoteling T-statistic evaluated on the coefficients, the p-value being computed using permutation sampling.



(a) SAEM



(b) tempering-SAEM

***Figure 7.12** – Independent factor estimation within the hippocampi dataset.* Results of the independent factor estimation on a corpus of 101 hippocampi. Atrophy patterns of the hippocampi in the context of Alzheimer's disease.

Following Allassonnière and Younes (2012), we compute the p-value for two different comparisons: the first one compare the healthy patients with respect to Alzheimer's and dementia patients (the two last groups). The second test compare the healthy patient with respect to the mild Alzheimer's patients (the second group). Due to the stochasticity of the SAEM algorithm, we computed an average and a standard deviation of the p-values over 50 runs, with the same initial conditions. Thus, the tempering-SAEM algorithm always behaves at least as well as the SAEM algorithm.

Finally, applying the tempering-SAEM for independent factor analysis aims to check that the advantages of the tempering-SAEM over the SAEM can improve or at least does not deteriorate the results of maximum likelihood estimation in complex hierarchical models.

II – 3. Discussion and Perspective

We propose in this chapter a new stochastic approximation version of the EM algorithm. The benefit of this general procedure is twofold: we can deal with the problem of intractable or difficult sampling in one hand and favor convergence toward global *maxima* in the other hand.

Our first contribution is theoretical with the proof of the convergence of the approximated-SAEM toward local *maxima*. This result gives an *a posteriori* justification for some existent schemes like the ABC-SAEM (Picchini and Samson, 2018) or Monolix. Moreover, our general framework is versatile enough to encompass a wide range of algorithms. Our second contribution goes this way by proposing an instantiation of this general procedure to prevent convergence toward local *maxima*, referred to as tempering-SAEM. This tempering-SAEM method is the one used in the Monolix software. We have applied this algorithm in both synthetic and real data frameworks and obtained improved results with respect to the state of the art algorithms in both cases.

This opens up new perspectives. Typically, now that we have ensured of the convergence of the approximated-SAEM, a natural opening concerns the study of the convergence of the approximated-MCMC-SAEM. Indeed, although the convergence of this algorithm has not yet been demonstrated, the tempering-MCMC-SAEM has already shown its numerical efficiency, especially in the case of medical applications (Debavelaere et al., 2019).

Appendix A. Multivariate Gaussian Mixture Model

We give here some details about the estimation procedure in the multivariate Gaussian mixture model. The complete log-likelihood of the GMM model is

$$\log q(y,z;\theta) \;=\; -n\log 2\pi - \sum_{j=1}^{m}\sum_{i=1}^{n}\left(\frac{1}{2}\log|\Sigma_j| - \log\alpha_j + {}^t(y_i-\mu_j)\,\Sigma_j^{-1}\,(y_i-\mu_j)\right)\mathbb{1}_{\{z_i=j\}}\,.$$

A – 1. Estimation through the EM Algorithm

Let $t\in\mathbb{N}$ index the current iteration. The general EM algorithm iterates the following two steps:

E-step: Compute $Q(\theta|\theta^t)=\mathbb{E}\left[\log q(y,z;\theta)|y,\theta^t\right]$;

M-step: Set $\theta^{t+1}=\operatorname{argmax}_{\theta\in\Theta}Q(\theta|\theta^t)$.

For all $(i,j)\in[\![1,n]\!]\times[\![1,m]\!]$, set $\tau_{i,j}=\mathbb{P}\left[z_i=j|y_i,\theta^t\right]$. Then,

$$Q(\theta|\theta^t) \;=\; -n\log 2\pi - \sum_{j=1}^{m}\sum_{i=1}^{n}\left(\frac{1}{2}\log|\Sigma_j| - \log\alpha_j + {}^t(y_i-\mu_j)\,\Sigma_j^{-1}\,(y_i-\mu_j)\right)\tau_{i,j}\,.$$

According to Bayes' rule,

$$\tau_{i,j} \;=\; \frac{\alpha_j\mathcal{G}(y_i-\mu_j;\Sigma_j)}{\sum_{j=1}^{m}\alpha_j\mathcal{G}(y_i-\mu_j;\Sigma_j)}\,,$$

where $\mathcal{G}(y-\mu;\Sigma)$ refers to the Gaussian distribution with mean μ and covariance matrix Σ. Lastly, a straightforward computation gives

$$\alpha_j^{t+1} \;=\; \frac{1}{n}\sum_{i=1}^{n}\tau_{i,j}\quad,\qquad \mu_j^{t+1} \;=\; \frac{\sum_{i=1}^{n}\tau_{i,j}y_i}{\sum_{i=1}^{n}\tau_{i,j}}$$

and

$$\Sigma_j^{t+1} \;=\; \frac{\sum_{i=1}^{n}\tau_{i,j}(y_i-\mu_j^{t+1})\,{}^t(y_i-\mu_j^{t+1})}{\sum_{i=1}^{n}\tau_{i,j}}\,.$$

A – 2. Estimation through the SAEM Algorithm

Given a sequence of positive step-size for the stochastic approximation $\gamma=(\gamma_t)_{t\in\mathbb{N}}$, the general SAEM algorithm iterates the following two steps:

SAE-step: Sample a new hidden variable z^{t+1} according to the conditional distribution $q(z|y,\theta^t)$ and compute

$$Q_{t+1}(\theta) \;=\; Q_t(\theta)+\gamma_t\left(\log q(y,z;\theta^t)-Q_t(\theta)\right)\,;$$

M-step: Set $\theta^{t+1} = \text{argmax}_{\theta\in\Theta} Q_{t+1}(\theta)$.

The GMM belongs to the curved exponential family. Actually, for all y, z and θ,

$$\log q(y,z;\theta) = -n\log(2\pi) + \sum_{j=1}^{m}\left(\log\alpha_j - \frac{1}{2}\log|\Sigma_j| + \left\langle \mu_j\,{}^t\mu_j \,\middle|\, \Sigma_j^{-1}\right\rangle_{\mathcal{F}}\right) S_{1,j}(y,z) + \sum_{j=1}^{m}\left[\left\langle \Sigma_j^{-1}\,\middle|\, S_{3,j}(y,z)\right\rangle_{\mathcal{F}} - 2\left\langle \Sigma_j^{-1}\mu_j\,\middle|\, S_{2,j}(y,z)\right\rangle\right]$$

where, for all $j \in [\![1,m]\!]$,

$$S_{1,j}(y,z) = \sum_{i=1}^{n} \mathbb{1}_{z_i=j} \quad ; \quad S_{2,j}(y,z) = \sum_{i=1}^{n} y_i\,\mathbb{1}_{z_i=j} \qquad \text{and} \qquad S_{3,j}(y,z) = \sum_{i=1}^{n} y_i\,{}^t y_i\,\mathbb{1}_{z_i=j}\,.$$

So, the SAE-step is replaced by an update of the estimation of the conditional expectation of the sufficient statistics, namely, for all $\ell \in \{1,2,3\}$, and all j,

$$S_{\ell,j}^{t+1} = S_{\ell,j}^{t} + \gamma_t\left(S_{\ell,j}(y,z^{t+1}) - S_{\ell,j}^{t}\right)$$

where, for all i, z_i^{t+1} is sampled from the discrete law $\sum_{j=1}^{m}\tau_{i,j}\delta_j$ where $\tau_{i,j} = \mathbb{P}\left[z_i = j|y_i,\theta^t\right]$ as in the EM-case.

The M-step can also be computed in close-form:

$$\alpha_j^{t+1} = \frac{1}{n}S_{1,j}\,, \qquad \mu_j^{t+1} = \frac{S_{2,j}}{S_{1,j}} \qquad \text{and} \qquad \Sigma_j^{t+1} = \frac{S_{3,j} - S_{2,j}\,{}^t\mu_j^{t+1}}{S_{1,j}}\,.$$

A – 3. Estimation through the tmp-SAEM Algorithm

The previous computation remain true except that the hidden variables z_i^{t+1} are now sampled from the tempered conditional distribution

$$\frac{1}{c(T_t)}\sum_{j=1}^{m}\tau_{i,j}^{1/T_t}\delta_j\,, \qquad \text{where} \qquad c(T_t) = \sum_{j=1}^{m}\tau_{i,j}^{1/T_t}$$

and T_t is defined in Section I.2.

To stabilize the convergence of both SAEM and tempering-SAEM, we may use inverse Wishart priors for the variances and Gaussian priors for the weights.

– Quatrième partie –

Conclusion & Perspectives

– Chapitre VIII –

Conclusion et perspectives

Le présent manuscrit comporte deux volets : tout d'abord, nous avons proposé un modèle non linéaire à effets mixtes pour l'analyse statistique de données longitudinales à dynamiques multiples et à valeurs sur des variétés riemanniennes. Ce modèle a été conçu avec en tête des applications à l'anatomie computationnelle. Nous avons ensuite concentré nos efforts sur un aspect plus numérique en proposant une nouvelle classe d'approximations stochastiques de l'algorithme EM. Nous profitons de ce chapitre pour passer en revue notre contribution en ce qui concerne, d'une part, l'analyse statistique de données longitudinales et, d'autre part, l'optimisation stochastique.

Sur l'analyse spatio-temporelle de données longitudinales

L'analyse statistique de données longitudinales à valeurs sur des variétés est un domaine en pleine mutation. Une idée porteuse et qui a fait ses preuves pour ce type de données est de tirer partie de la structure géométrique forte des variétés riemanniennes et de découpler transformations spatiale et temporelle. On est ainsi à même de quantifier et d'expliquer la variabilité inter-individuelle aussi bien en terme de déformations géométriques que de changements de dynamique d'évolution. À notre connaissance, le premier modèle du type remonte aux travaux de Durrleman et al. (2013). Plus récemment, l'approche générique développée par Schiratti et al. (2015, 2017) permet d'aligner temporellement les différents individus de manière efficiente et avec une complexité algorithmique maitrisée. Ce modèle possède intrinsèquement une grande applicabilité mais une hypothèse très forte sur la dynamique du phénomène observé est faite, ce qui en réduit du même coup la portée : tous les individus sont alignés selon une dynamique unique et supposée monotone, ce qui est en pratique peu réaliste. Mathématiquement, cette hypothèse se traduit à l'échelle macroscopique par la construction d'une trajectoire représentative géodésique et, à l'échelle individuelle, par la construction d'un couple de déformations spatiale et temporelle pour chacun des sujets.

Dans cette thèse, et on nous basant sur cette approche générique, nous avons développé un modèle permettant de traiter le cas de dynamiques non-monotones. Pour cela, on suppose que la trajectoire représentative de l'évolution n'est plus seulement géodé-

sique mais géodésique par morceaux et on construit autant de déformations spatiales et temporelles qu'il y a de géodésiques constituant la courbe représentative. Ces déformations sont construites avec une certaine autonomie les unes par rapport aux autres afin de pouvoir traiter d'une grande variété de situations pratiques, tout en assurant une évolution continue pour chaque sujet. De fait, le modèle de Schiratti et al. (2015, 2017) apparait comme un cas particulier de notre modèle générique et on conserve par rapport à ce modèle original une grande applicabilité, en étant plus générique.

En particulier, nous avons appliqué ce modèle au suivi de chimiothérapie, situation typique dans laquelle la dynamique d'évolution est amenée à changer. En effet, la mise en place d'un nouveau traitement induit généralement trois phases d'évolution distinctes pour un même patient : une phase de réponse au traitement dans laquelle la taille de ses tumeurs va décroitre, une phase dite stable où la taille des tumeurs reste inchangée et, dans la plupart des cas, une phase de progression de la maladie dans laquelle les tumeurs vont de nouveau croître, ce qui nécessite la mise en place d'un nouveau traitement dans les délais les plus brefs. Pouvoir estimer avec précision le temps d'échappement au traitement est donc crucial dans ce contexte.

Plus précisément, dans le cadre du suivi chimiothérapeutique, nous avons proposé deux instanciation du modèle générique : d'une part, pour le suivi de score RECIST et, d'autre part, pour le suivi de forme anatomique en 3 dimension que l'on apparente à des tumeurs segmentées. Ce premier modèle pour les scores RECIST est le fruit d'une collaboration en cours avec des oncologues et radiologues de l'HEGP.

Nous nous sommes également intéressés à démontrer la consistance de l'estimateur du *maximum a posteriori* (MAP) pour ce modèle, à savoir à démontrer que toute suite produite par l'estimateur du MAP allait être asymptotiquement aussi proche que souhaité de l'ensemble des paramètres admissibles, *i.e.* de l'ensemble des paramètres induisant le MAP. Cette démonstration repose sur le livre séminal de van der Vaart (2000) ; contrairement aux résultats de consistance classiques, nous avons également démontré que l'ensemble des paramètres admissibles était non-vide. Nous avons par ailleurs formulé les hypothèses de notre théorème de manière à traiter du modèle de Schiratti et al. (2015, 2017) et de ses variantes (Bône et al., 2018; Koval et al., 2018) ; cette démonstration fournit donc des garanties théoriques à plusieurs modèles pré-existants.

Cependant, et malgré son caractère très générique, le modèle tel que proposé dans ce manuscrit ne permet pas l'étude de populations dans lesquelles les comportements de certaines sous-populations diffèrent. Pour cela, il conviendrait, en se basant sur des modèles de mélanges usuels, d'introduire une sur-couche dans la modélisation que nous proposons ici afin d'aboutir à un modèle de mélange pour l'étude statistique de données longitudinales à valeurs sur des variétés riemanniennes. Les travaux récents de Debavelaere et al. (2019) vont dans ce sens.

Une autre limitation réside dans le choix de la métrique sur la variété template M_0. En effet, nous supposons ce choix comme naturel dans la construction de notre modèle. Pour autant, lorsque l'on s'intéresse à des données complexes ou en grande dimension, il n'est généralement pas possible de concevoir une métrique riemannienne pertinente. En se basant sur des réseaux de neurones, Louis et al. (2017, 2019) ont proposé une méthode pour apprendre la métrique riemannienne idéale pour le modèle de Schiratti et al. (2015, 2017). Ces travaux ne s'appliquent cependant pas en l'état à notre modèle, notre trajectoire représentative étant géodésique par morceaux et pas seulement géodésique. Ainsi, une première perspective possible réside dans le fait de généraliser les travaux de Louis et al. (2017, 2019) à des dynamiques multiples pour pouvoir traiter de notre modèle.

En outre, nous imposons comme hyper-paramètre du modèle le nombre de composantes m dans la trajectoire représentative géodésique par morceaux. Dans le cas du suivi chimiothérapeutique, la dynamique d'évolution médicale sous-jacente étant évidente, cette contrainte n'est pas très restrictive. Pour autant, on peut très bien imaginer des situations pratiques dans lesquelles ce n'est pas le cas. Par exemple, la forme rémittente de la sclérose en plaque évolue par poussées, caractérisées par l'apparition de troubles en quelques jours, pouvant régresser complètement ou non en quelques semaines ; ce nombre de poussées est bien évidement inconnu. La sélection de modèles pourrait résoudre ce problème mais des investigations plus poussées sont nécessaires pour s'en convaincre.

Enfin, notre modèle suppose que chaque morceau de la trajectoire représentative est construit à partir d'une même variété template M_0, comme une déformation spatiale d'une géodésique de cette variété template. Cette variété template s'interprète comme encodant le type de données que l'on observe : un score, des scores, des images, *etc.* Pour autant, en pratique, cette variété peut être amenée à varier d'un tronçon à l'autre. Pour reprendre l'exemple du suivi de chimiothérapie, les oncologues peuvent vouloir étudier la progression tumorale dans n_1 différents organes lors d'une première phase d'évolution de la maladie puis de modifier le choix ou le nombre de ces organes dans une seconde phase, conduisant à $n_2 \neq n_1$ organes suivis. Alors, la variété M_0 consisterait dans la première phase en un produit cartésien de n_1 variétés et dans la deuxième phase en un produit cartésien de n_2 variétés *a priori* distinctes de celles utilisée dans la première phase.

Sur les approximations stochastiques de l'algorithme EM

Le modèle décrit ci-dessus, de par sa complexité nécessite une algorithmique performante. L'algorithme de Metropolis-Hastings classique ne permet notamment pas une estimation efficace du MAP du fait de la présence de nombreux minima locaux dans la vraisemblance du modèle. Dans ce but, nous avons proposé une nouvelle classe d'approximations stochastiques de l'algorithme EM : les algorithmes SAEM approchés, ou approximated-SAEM en anglais. Cette classe repose sur la simulation par une loi approchée de la vraie loi conditionnelle dans l'étape de simulation. Plus précisément, on dit d'une suite de loi de probabilités $\left(\tilde{q}_k(\cdot;\theta)\right)_{k\in\mathbb{N}}$ qu'elle approche la loi $q(\cdot;\theta)$ si elle

converge en moyenne sur tout compact de Θ vers $q(\cdot;\theta)$. Avec le formalisme des modèles à données manquantes, pour toute observation $y \in \mathcal{Y}$, on écrira donc que la suite $\left(\tilde{q}_k(\cdot;\theta)\right)_{k\in\mathbb{N}}$ approche la loi conditionnelle $q(\cdot|y;\theta)$ si, pour tout compact $\mathcal{K} \subset \Theta$,

$$\lim_{k\to\infty} \left\{ \sup_{\theta\in\mathcal{K}} \int_{\mathcal{Z}} S(y,z)\left(\tilde{q}_k(z;\theta) - q(z|y;\theta)\right) \mathrm{d}\mu(z) \right\} = 0\,.$$

En particulier, un tel schéma numérique englobe des algorithmes pré-existants tel que l'ABC-SAEM (Picchini et Samson, 2018) dont l'efficacité a été établie mais dont la convergence théorique n'avait pas été démontrée.

Nous avons démontré la convergence locale d'un telle classe d'approximations stochastiques. Cette démonstration nous a permis de mettre en exergue la similitude entre les hypothèses requises par cette nouvelle classes d'approximation stochastiques et la classe usuelle de Delyon et al. (1999). En d'autre termes, approcher la distribution conditionnelle dans l'étape de simulation n'exige pas de considérations supplémentaires pour continuer à garantir la convergence de la séquence générée par l'algorithme SAEM. Pour autant, on se limite donc au cas de modèles appartenant à la famille exponentielle ; une première amélioration possible serait de s'affranchir de cette hypothèse.

En nous basant sur des techniques de recuit simulé, nous avons proposé une version tempérée de l'algorithme SAEM. Cette version a été conçue dans le but de favoriser la convergence de l'algorithme SAEM vers des minima globaux et ce pour les raisons expliquées ci-dessus. Pour ce faire, on construit une suite approchante de la loi conditionnelle en tempérant cette dernière suivant un schéma de températures sinusoïdal amorti. En effet, une température élevée va avoir tendance à lisser le profil de la densité de probabilité tandis qu'une température plus basse va au contraire rendre cette dernière plus piquée. Ainsi, en cas de distribution multimodales, en alternant des phases de chaud et de froid, on autorise la variable latente à switcher de mode pendant les phases de chauffe et à se fixer pendant les phases de refroidissement. Les expériences que nous avons conduites sur le modèle de mélange gaussien et l'analyse en facteurs indépendants confirment cette heuristique. Pour autant même si nos expériences numériques tendent à démontrer la véracité de cette heuristique, la démonstration de la convergence du SAEM tempéré vers un minima global reste un problème ouvert.

Une autre limitation concernant le SAEM tempéré relève du choix des paramètres décrivant la suite de températures. En effet, afin de permettre une grande flexibilité en fonction des situations étudiées, nous avons proposé une suite de températures dépendant de paramètres multiples, à savoir

$$T_k \;=\; 1 + a^{\kappa} + b\,\frac{\sin(\kappa)}{\kappa}\,, \qquad \text{where} \qquad \kappa \;=\; \frac{k + c\times r}{r}\,.$$

Le choix des paramètres optimaux se révèle épineux en pratique. Une amélioration naturelle serait de proposer un schéma de températures adaptatif, à la manière des schémas adaptatifs pour faire varier la variance de la loi de proposition dans le cas de l'algorithme

de Metropolis-Hastings (Atchadé, 2006; Roberts et Rosenthal, 2007, 2009).

Enfin, en nous basant sur les travaux de Andrieu et al. (2006) et Allassonnière et al. (2010), nous travaillons actuellement à la démonstration de la convergence d'algorithmes de type MCMC-SAEM approchés. En effet, tel que proposé dans cette dissertation, nous supposons que nous sommes à même de simuler contre chacun des termes de la suite approchante $\left(\tilde{q}_k(\cdot;\theta)\right)_{k\in\mathbb{N}}$, ce qui s'avère ne pas être le cas la plupart du temps et en particulier en ce qui concerne le SAEM tempéré. Debavelaere et al. (2019) ont d'ores et déjà appliqué le MCMC-SAEM tempéré sur un modèle similaire à celui qui est introduit dans notre manuscrit (Partie Part I.) et ils ont obtenu des résultats numériques très encourageants. De plus, le logiciel Monolix repose sur une algorithmique similaire au MCMC-SAEM tempéré sur ses premières itérations.

Vers un outil d'aide à la décision pour les médecins

Un des buts du modèle qui nous avons introduit est d'être utilisé pour faire de la prédiction : étant donné des observations pour un individus donné, typiquement les premières mesures lors d'un suivi médical, on souhaiterait pouvoir conclure quant au devenir de ce sujet afin de permettre par exemple la détection précoce de changement ou de survenue d'un événement. L'objectif sous-jacent est de pouvoir proposer aux praticiens un outil d'aide à la décision qui leur permettrait de choisir le plus rapidement possible le traitement le plus adapté à leur patient, en fonction de la réaction de celui-ci aux premières étapes de la prise en charge et de l'évolution depuis la détection de la maladie.

Le modèle pour les scores RECIST correspond en fait à la premières étape dans la réalisation d'un tel outil. Dans cette première étape, nous nous sommes intéressés à l'effet d'une molécule sur l'évolution tumorale et ce sans tenir compte ni des antécédents médicaux, ni des données omiques pour chacun des patients. Plusieurs améliorations se présentent alors naturellement à nous.

Un première amélioration se rapporte au type de données traitées. En effet, nous avons également proposé un modèle pour l'étude des tumeurs 3D segmentées et ce modèle nous apparait comme plus précis en terme de suivis médical, le score RECIST n'étant pas aussi informatif que la forme exacte des tumeurs. Pour autant, il repose sur une segmentation manuelle des tumeurs et donc sur un choix humain arbitraire des tumeurs à étudier. Afin de palier des biais mentaux auxquels ne serait, par nature, pas sensible un algorithme d'apprentissage, il nous apparait donc comme judicieux de proposer une instanciation du modèle générique pour des données images.

Une seconde amélioration serait de ne plus se limiter à l'étude de chacun des médicaments indépendamment mais à l'effet d'un protocole de soin constitué d'une succession de plusieurs cures médicamenteuses. En pratique, dans le cas du cancer du rein métastatique, en considérant les traitements anti-angiogéniques et dits de thérapie génique, les praticiens disposent de neuf molécules différentes. D'une part, le choix mais, éga-

lement, l'ordre dans lequel ces différents médicaments sont administrés semblent avoir une importance. Cependant, aucune méta-étude n'ayant été menée sur ces protocoles de soin, il est pour le moment impossible de conclure quant à la séquence de soin optimale, à patient donné. Coupler ce premier modèle, qui permet d'étudier des comportements fluctuant au cours du temps, à des techniques de sélection de modèles permettrait de fournir une réponse d'une utilité pratique manifeste, tout en ouvrant de nombreuses perspectives mathématiques. En effet, même à supposer que le nombre de molécules soit constant dans les différents protocoles de soin, le tirage avec remise des différentes substances actives engendrerait un nombre trop important de configurations à tester, pour un nombre trop faible de patients. Ainsi, on ne peut se contenter d'une procédure naïve consistant à tester l'ensemble des configurations possibles et de choisir la « meilleure ».

Un autre aspect d'amélioration de notre modèle est la meilleure prise en charge des données manquantes. Les données médicales, et plus encore en cancérologie où les prises de mesures sont coûteuses et éprouvantes pour le patient, sont souvent incomplètes ou bruitées et les outils existants répondent encore mal à la nécessité d'extrapoler efficacement malgré les informations manquantes. Ainsi, développer des méthodes aptes à pallier l'aspect parcellaire des données dans le cadre de nos modèles longitudinaux s'avère crucial pour une approche cohérente et efficace de la problématique médicale. De plus, intégrer les données omiques quand celles-ci sont disponibles permettrait une meilleure prise en charge. Outre la difficulté de traitement mathématique de ces données qui devrait aboutir à des questions d'ordre géométrique, ce genre de données ne sont pas toutes disponibles pour l'ensemble des patients, induisant par là même un challenge en terme de traitement statistique.

Dans la même idée, nous avons dans ce premier modèle étudié le score RECIST de chacun des patients. Or, les différentes tumeurs se développant dans des organes distincts, on peut associer à chacun des patients non pas un mais des scores RECIST : un par organe atteint par le cancer. Comme annoncé à la page 175, une manière simple d'étudier plusieurs organes en parallèle est de considérer que la variété template M_0 est en fait un produit cartésien de variétés correspondant pour chacune à un organe. Cependant, outre le caractère fluctuant des organes suivis par les oncologues au cours des différentes phases de progression de la maladie, ceux-ci peuvent même être distincts d'individus à l'autre, ce qui complexifie encore l'exercice de modélisation.

Par suite, cette thèse a été l'occasion pour moi de proposer un modèle pour l'analyse statistique de données longitudinales dont la dynamique d'évolutions est non-monotone et à valeurs sur des variétés riemanniennes. Elle a également été l'occasion de proposer une algorithmique efficiente afin de rendre possible une estimation performante dans nos modèles compliqués. Enfin, j'ai collaboré avec des oncologues de l'HEGP afin de poser les premières briques dans la réalisation d'un outil d'aide à la prise en charge médicale. Cette collaboration m'a permise de me familiariser avec le domaine des mathématiques appliquées à la médecine de manière réellement effective.

– Chapter IX –

Conclusion and Perspectives

Our work in this dissertation is twofold: we first proposed a nonlinear mixed effects model for the statistical analysis of longitudinal Riemannian manifold valued data with multiple dynamic. This model was designed with the aim of application to computational anatomy. We then focused on a more numerical aspect by proposing a new class of stochastic approximations of the EM algorithm. We use the pretext of the present chapter to review our contribution in the domain of statistical analysis of longitudinal data and stochastic optimization.

On the Saptio-Temporal Analysis of Longitudinal Data

The statistical analysis of longitudinal Riemannian-valued data is undergoing tremendous changes. A promising idea is to take advantage of the strong geometrical structure of the Riemannian shape spaces and to uncouple spatial and temporal deformations. We are then able to quantify and explain inter-individual variability both in terms of geometrical deformations and changes in the evolution's dynamic. To our knowledge, the first such model is from the work of Durrleman et al. (2013). More recently, the generic approach developed by Schiratti et al. (2015, 2017) allows to temporally align the individuals in an efficient way and with control over the algorithmic complexity. This model posses a wide intrinsic applicability but a strong hypothesis is made on the dynamic of the observed phenomenon, thus reducing its scope: all the individuals are aligned on a unique monotonous dynamic by assumption, which can be unrealistic. This hypothesis is mathematically reflected at macroscopic scale by construction of a geodesic representative trajectory and at microscopic scale by construction of a couple of spatial and temporal deformations for each subject.

In this PhD work, and building on the generic approach, we developed a model allowing to handle non-monotonous dynamics. To do so, we assume the representative trajectory of the evolution to no longer be geodesic but piecewise-geodesic instead and we construct as many spatial and temporal deformations as there are pieces of geodesics in the representative curve. These deformations are constructed with some autonomy relatively to each other in order to treat a wide range of practical situations although

enforcing a continuous evolution for each subject. Actually, the model of Schiratti et al. (2015, 2017) appears as a particular case of our generic model and we inherit from its broad applicability, being more generic.

In particular, we applied this model to chemotherapy monitoring, a typical situation in which the evolution dynamic is subject to change. Indeed, the introduction of a new treatment leads to three distinct phases of evolution for a patient: a first phase of response to treatment with decreasing of tumors size, a second stable phases with tumors size unchanged and, in most cases, a last phase of progression of the disease with a new increase in tumors size. There is then an urgent necessity for a new treatment. Therefore, being able to estimate precisely the escape time associated to the treatment is crucial in this context.

Plus précisément, dans le cadre du suivi chimiothérapeutique, nous avons proposé deux instanciation du modèle générique : d'une part, pour le suivi de score RECIST et, d'autre part, pour le suivi de forme anatomique en 3 dimension que l'on apparente à des tumeurs segmentées. Ce premier modèle pour les scores RECIST est le fruit d'une collaboration en cours avec des oncologues et radiologues de l'HEGP.

We also focused on proving consistency for the estimator of the *maximum a posteriori* (MAP) for this model, that is to say proving that all sequences produced by the estimator of the MAP is asymptotically as close as possible to the set of admissible parameters, *i.e.* the set of parameters associated with the MAP. This proof relies on the seminal book of van der Vaart (2000); unlike classical consistency results, we also proved that the set of admissible parameters is non-empty. We also formulated our hypotheses in order to include the model of Schiratti et al. (2015, 2017) and its variants Bône et al. (2018); Koval et al. (2018); this proof thus provides theoretical guarantees to previous models.

However, and despite it generic nature, the model proposed in this manuscript is not suited for populations in which sub-populations present different behaviors. Encompassing these situations would require to introduce, building on usual mixture models, another modeling layer. The recent work of Debavelaere et al. (2019) head in this direction.

Another limitation is the choice of the metric over the template manifold M_0. Indeed, we assume this choice as natural in the construction of our model. However, when dealing with high-dimensional or complex data, it is in general not possible to design a Riemannian geometry of relevance. Based on neural networks, Louis et al. (2017, 2019) proposed a method to learn the ideal Riemannian metric in the context of the model of Schiratti et al. (2015, 2017). However, this work does not apply to our model as it stands, since our representative trajectory is piecewise-geodetic and not only geodetic. Thus, a first possible perspective is to generalize Louis et al. (2017, 2019)s work to multiple

dynamics in order to be able to deal with our model.

Moreover, we impose as hyper-parameters the number of components m in the piecewise-geodesic representative trajectory. In the case of chemotherapy monitoring its not a very restrictive constraint as the evolution dynamic is standard. But there are more complex situations where its not the case. For example, the relapsing form of multiple sclerosis tends to come and go, with the characteristic apparition of trouble in a few days, sometimes dwindling in some weeks, completely or not; this number of phases is indeed unknown. Model selection could overcome this issue but a more thorough investigation is required to be sure.

Lastly, our model assumes each piece of the representative trajectory to be constructed within the same template manifold M_0, as a spatial deformation of a geodesic of this template manifold. This manifold is to be interpreted as encoding the type of observed data: a score, many scores, images, *etc.* But still, in practice, this array can vary from a piece of trajectory to another. Back to chemotherapy monitoring context, oncologists could want to study tumors progression in n_1 different organs in a first evolution phase of the disease and to study afterwards a higher number $n_2 \neq n_1$ of possibly different organs in a second phase. Then, the manifold M_0 will be a Cartesian product of n_1 manifolds during the first phase and a Cartesian product of n_2 manifolds, distinct from the previous ones, during the second phase.

On the Stochastic Approximations of the EM Algorithm

Because of its complexity, the model described above requires a high-performance algorithm. In particular, the traditional Metropolis-Hastings algorithm does not allow an efficient estimation of the MAP due to the presence of many local minima in the likelihood of the model. To this end, we have proposed a new class of stochastic approximations of the EM algorithm: the approximated-SAEM algorithms. This class relies on sampling from an approximation of the real conditional law in the simulation step. More specifically, a sequence of probability distributions $\left(\tilde{q}_k(\cdot;\theta)\right)_{k\in\mathbb{N}}$ is said to approach a distribution $q(\cdot;\theta)$ if it converges in mean on every compact subset of Θ toward $q(\cdot;\theta)$. With the formalism of missing data models, for all observation $y \in \mathcal{Y}$, we will therefore write that the sequence $\left(\tilde{q}_k(\cdot;\theta)\right)_{k\in\mathbb{N}}$ approach the conditional distribution $q(\cdot|y;\theta)$ if, for every compact $\mathcal{K} \subset \Theta$,

$$\lim_{k\to\infty} \left\{ \sup_{\theta\in\mathcal{K}} \int_{\mathcal{Z}} S(y,z)\big(\tilde{q}_k(z;\theta) - q(z|y;\theta)\big)\,\mathrm{d}\mu(z) \right\} = 0\,.$$

In particular, we encompass pre-existent algorithms like the ABC-SAEM (Picchini and Samson, 2018) whose numerical efficiency was empirically established but whose theoretical convergence was not proved.

We proved the local convergence of such a class of stochastic approximations. This demonstration allowed us to highlight the similarity between the assumptions required by this new stochastic approximations of the EM algorithm and the usual one of Delyon

et al. (1999). In other words, approximate the conditional distribution in the sampling step does not require supplementary considerations to still guarantee the convergence of the sequence generated by the SAEM algorithm. Nevertheless, we limit ourselves to models for which the the complete data likelihood belongs to the curved exponential family; a first possible improvement would be to relax this assumption.

Building on simulated annealing techniques, we propose a tempered version of the SAEM algorithm. This version was designed to favor the convergence of the SAEM toward global minima for the reasons explained above. To do this, we build a sequence of approximated distributions by tempering the conditional law according to a damped sinusoidal temperature scheme. Indeed, a high temperature will tend to flat the profile of the distribution, while a lower temperature will make this profile look more prickly. Thus, in case of multimodal density, by alternating hot and cold phases, the latent variable will be able to switch from one mode to an other during the heating steps and to explore these same modes during the cooling phases. The experiments we conducted on the Gaussian mixing model and the analysis in independent factors confirm this heuristic. Even if our numerical experiments tend to prove the truth of this heuristic, the demonstration of the convergence of tempering SAEM toward a global minimum remains an ongoing problem.

Another limitation concerning tempering SAEM is the choice of parameters describing the temperature sequence. Indeed, in order to allow a great flexibility according to the studied situations, we proposed a temperature scheme depending on multiple parameters, namely

$$T_k \;=\; 1 + a^{\kappa} + b\,\frac{\sin(\kappa)}{\kappa}\,, \qquad \text{where} \qquad \kappa \;=\; \frac{k + c \times r}{r}\,.$$

Choosing the optimal parameters is tricky in practice. A natural improvement would be to propose an adaptive temperature scheme, in the same way as adaptive schemes for variances of the proposal law in sampling algorithms such as Metropolis-Hastings algorithm (Atchadé, 2006; Roberts and Rosenthal, 2007, 2009).

Lastly, based on the work of Andrieu et al. (2006) and Allassonnière et al. (2010), we are currently working on the proof of the convergence of approximated MCMC-SAEM. Indeed, as proposed in this dissertation, we assume that we are able to draw samples from each of the terms of the sequence of approximated distributions $\left(\tilde{q}_k(\cdot;\theta)\right)_{k\in\mathbb{N}}$, which is not the case most of the time and in particular with regard to tempering SAEM. Debavelaere et al. (2019) already applied the tempering MCMC-SAEM on a model similar to the one introduced in our manuscript (Part Part II) and obtained very promising numerical results. In addition, the Monolix software is based on a similar tempering MCMC-SAEM on its first iterations.

Toward a Decision-Making Tool for Physicians

One of the aims of the model that we introduced is to be used to make predictions: given observations for a given individual, typically the first measurements during a medical monitoring, we would like to be able to conclude on the future of this subject in order to allow for example the early detection of a change or the occurrence of an event. The underlying objective is to be able to provide physicians with a decision support tool that would allow them to choose the most appropriate treatment for their patient as quickly as possible, based on the patient's response to the early stages of medical care and progression since the disease was detected.

The model for RECIST scores is in fact the first step in the development of such a tool. In this first step, we were interested in the effect of a molecule on tumor progression, without taking into account either medical history or omics data for each patient. Several improvements are then naturally available to us.

A first improvement relates to the type of data processed. Indeed, we have also proposed a model for the study of segmented 3D tumors and this model appears to be more precise in terms of medical monitoring, the RECIST score being not as informative as the exact shape of the tumors. However, it is based on a manual segmentation of tumors and therefore on a human and arbitrary choice of tumors to be studied. In order to compensate for mental biases to which a learning algorithm would not be sensitive by nature, it therefore seems appropriate to propose an instantiation of the generic model for image data.

A second improvement would be to no longer restricted to study each drug independently but rather to study the effect of a care protocol consisting of a succession of several drugs. In practice, in the context of metastatic kidney cancer, when considering anti-angiogenic and gene therapy treatments, physicians have nine different molecules. On the one hand, the choice but also the order in which these different drugs are administered seems to be important. However, as no meta-study has been conducted on these care protocols, it is currently impossible to conclude on the optimal sequence of care for a given patient. Coupling this first model, which allow for the study of behaviors that fluctuate over time, with model selection techniques would provide a response of obvious practical utility, while opening up many mathematical perspectives. Indeed, even if the number of molecules were to be constant in the different care protocols, the sampling with replacement of the different active substances would generate too many configurations to be tested, for a too small number of patients. Thus, one cannot be satisfied with a naive procedure consisting in testing all possible configurations and choosing the "best".

Another aspect of improving our model is to better manage missing data. Medical data, and even more in oncology where measurements are costly and strenuous for the patient, are often incomplete or noisy. Existing tools still do not respond well to the need to extrapolate effectively despite missing information. Thus, developing methods to overcome the fragmented aspect of data in our longitudinal models is crucial for a

coherent and effective approach to medical problems. In addition, integrating omics data when available would allow for better monitoring. Apart from the difficulty of mathematical processing of these data, which should lead to geometric questions, this type of data is not all available for all patients, thus creating a challenge in terms of statistical processing.

In the same vein, in this first model we studied the RECIST score of each patient. However, since the different tumors develop in different organs, not one but RECIST scores can be associated with each patient: one per organ affected by cancer. As announced on page 181, a simple way to study several organs in parallel is to consider that the template manifold is in fact a Cartesian product of manifolds corresponding for each one to an organ. However, in addition to the fluctuating nature of the organs monitored by oncologists during the different phases of disease progression, they may even be distinct from one individual to another, making the modeling exercise even more complex.

Consequently, this PhD was an opportunity for me to propose a model for the statistical analysis of longitudinal Riemannian manifold valued data whose dynamic of evolution are non-monotonic. It was also the opportunity to propose an effective algorithm to allow for efficient estimation in our complicated models. Finally, I collaborated with oncologists form the HEGP to take the first steps in the development of a tool to assist in medical management. This collaboration gave me the opportunity to familiarize myself with the field of mathematics applied to medicine in a truly effective way.

– Cinquième partie –

Annexes

– Annexe A –

Des notions de géométrie riemannienne

La présente annexe redonne, en français, des résultats élémentaires de géométrie différentielle et riemannienne, utiles à la construction de notre modèle Part II.

Contents

On rappelle ici quelques résultats fondamentaux de géométrie différentielle et riemannienne. Pour une explication plus complète sur la géométrie différentielle, on peut consulter le livre de Lafontaine (1996) et sur la géométrie riemannienne, parmi les références classiques, les livres de Do Carmo (1976), Jost (2002) ou Gallot et al. (2004) ou, pour un exposé plus accessible, l'ouvrage de Rouvière et Debreil (2016).

I. Variétés différentielles

I – 1. Vecteurs tangents et application différentielle

a. Fonctions différentiables

Soit M et N deux variétés, de classe $\mathscr{C}^k$, de dimensions respectives n_M et n_N.

Definition A.1. *Une application continue $f : M \to N$ est dite* de classe $\mathscr{C}^k$ *si pour tout $m \in M$ il existe une carte $(\mathcal{U}, \phi)$ de M avec $m \in \mathcal{U}$, une carte $(\mathcal{V}, \psi)$ de N avec $f(m) \in N$ telle que l'application*

$$\psi \circ f \circ \phi^{-1} \ : \ \phi\left(f^{-1}(\mathcal{V}) \cap \mathcal{U}\right) \to \psi(\mathcal{V})$$

soit de classe $\mathscr{C}^k$, en tant que fonction de $\mathbb{R}^{n_M}$ dans $\mathbb{R}^{n_N}$.

Autrement dit, on considère le diagramme commutatif suivant :

$$\begin{array}{ccc} \tilde{\mathcal{U}} & \xrightarrow{\quad f \quad} & \mathcal{V} \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\psi} \\ \phi\left(\tilde{\mathcal{U}}\right) & \xrightarrow{\psi\circ f\circ\phi^{-1}} & \psi(\mathcal{V}) \end{array}$$

où $\tilde{U} = f^{-1}(\mathcal{V}) \cap \mathcal{U}$ pour que les compositions d'applications soient bien définies.

Proposition A.2

Toute composée d'applications $\mathscr{C}^k$ est de classe $\mathscr{C}^k$.

b. Espaces tangents

Notons $\mathscr{C}_m^M$ l'ensemble des courbes lisses $\gamma\colon I \to M$ définies sur un intervalle ouvert I contenant 0 et tel que $\gamma(0) = m$.

Definition A.3. *Deux courbes* $\gamma_1 : I_1 \to M$ *et* $\gamma_2 : I_2 \to M$ *de* $\mathscr{C}^M_m$ *sont dites* tangentes en m *si* $\gamma_1(0) = \gamma_2(0) = m$ *et il existe une carte* $(\mathcal{U}, \phi)$ *telle que* $m \in \mathcal{U}$ *et*

$$(\phi \circ \gamma_1)'(0) = (\phi \circ \gamma_2)'(0) \quad (\in \mathbb{R}^{n_M}) .$$

Intuitivement, cela revient à demander à ce que les courbes γ_1 et γ_2 aient la même vitesse. Par ailleurs, cette définition ne dépend pas de la carte $(\mathcal{U}, \phi)$ choisie. En effet si $(\mathcal{V}, \psi)$ est une autre carte telle que $m \in \mathcal{V}$, on a :

$$(\psi \circ \gamma_i)'(0) = \Big(\psi \circ \phi^{-1} \circ \phi \circ \gamma_i\Big)'(0) = \mathcal{D}_{\phi(m)} \underbrace{\Big(\psi \circ \phi^{-1}\Big)}_{\mathscr{C}^k-\text{difféo}} \cdot (\phi \circ \gamma_i)'(0) . \qquad (\star)$$

On définit donc une relation d'équivalence sur $\mathscr{C}^M_m$ ce qui amène à la définition :

Definition A.4. *Un* vecteur tangent *à* M *en un point* m *de* M *est une classe d'équivalence de la relation ci-dessus. L'ensemble des vecteurs tangents à* M *en* m *forme l'*espace tangent *à* M *en* m*, on le note* T_mM*.*

Example : Dans le cas dégénéré ou M est un espace vectoriel alors toute carte de M s'obtient par restriction de (M, Id). En particulier, deux courbes $\gamma_1 : I_1 \to M$ et $\gamma_2 : I_2 \to M$ sont tangentes en un point $m := \gamma_i(0) \in M$ ssi $\gamma_1'(0) = \gamma_2'(0)$. Autrement dit, ssi les courbes γ_1 et γ_2 admettent la même tangente – même vecteur directeur $\gamma_i'(0)$ et même point d'application $\gamma_i(0)$ – au sens usuel du terme et le vecteur tangent n'est autre que le vecteur directeur de cette unique tangente.

Proposition A.5

T_mM admet une structure d'espace vectoriel réel de dimension n_M.

Enfin, on conclut ce paragraphe en définissant l'application tangente entre deux variétés :

Definition A.6. *Soit* $f : M \to N$ *une application lisse et* $m \in M$*. On définit l'*application linéaire tangente *de* f *en* m*, notée* T_mf*, par passage au quotient de l'application* $\gamma \mapsto f \circ \gamma$ *de* $\mathscr{C}^M_m$ *dans* $\mathscr{C}^N_{f(m)}$*.*

Example : Reprenons l'exemple précédent en supposant que M et N sont des espaces vectoriels. Dans ce cas, l'application T_mf est la différentielle $\mathcal{D}_mf$ de f en m.
En effet, dans ce cas, $(f \circ \gamma)'(0) = \mathcal{D}_mf \cdot \gamma'(0)$ d'après la formule de dérivation composée d'où le résultat en quotientant.

Proposition A.7

Si $f : M \to N$ et $g : N \to P$ sont deux applications lisses entre variétés, alors

$$\forall m \in M, \quad T_m(g \circ f) = T_{f(m)}g \circ T_mf .$$

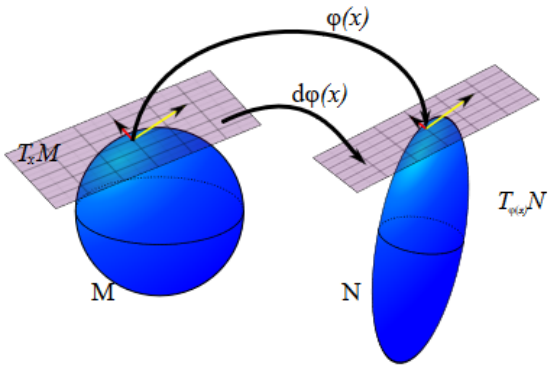


Figure A.1 – Vecteur tangent, plant tangent et application tangente.

c. Fibrés tangents

Definition A.8. *On appelle* fibré tangent *et on note TM l'union disjointe des espaces tangents à M en chacun de ses points.*

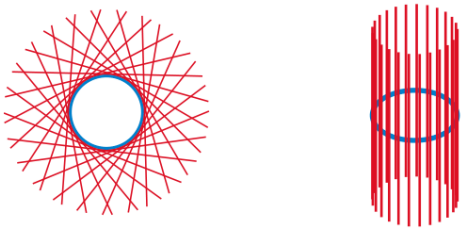


Figure A.2 – Deux représentations du fibré tangent du cercle.

Pour le moment, le fibré tangent ainsi défini est une somme ensembliste mais on peut le munir d'une topologie comme le précise la proposition suivante :

Proposition A.9

1. Si M est de classe $\mathscr{C}^k$, $k \geqslant 1$, alors son fibré tangent TM peut être muni d'une structure canonique de variété différentielle, de dimension $2n_M$ et de classe $\mathscr{C}^{k-1}$.
2. La projection canonique p de TM sur M est une fibration.

En fait, le fibré tangent est même un fibré vectoriel réel de rang n_M.

Definition A.10. *Une variété dont le fibré tangent est trivialisable est dite* parallélisable.

Example : Le cercle $\mathbb{S}^1$ est parallélisable comme l'illustre la figure A.2. A droite, on a trivialisé le fibré tangent.

I – 2. Champs de vecteurs

a. Dérivation globale

Definition A.11. *On appelle* champ de vecteur *sur une variété une subsection (lisse) de son fibré tangent. On note $\mathscr{C}^\infty(TM)$ l'ensemble des champs de vecteurs sur M.*

On peut munir $\mathscr{C}^\infty(TM)$ d'une structure d'espace vectoriel en posant :

$$\forall s,t \in \mathscr{C}^\infty(TM), \quad \forall x \in M, \quad (s +_{\mathscr{C}^\infty(TM)} t)(x) = s(x) +_{TM_x} t(x)$$

où l'opération $+_{TM_x}$ est bien définie de par la structure d'espace vectoriel des fibres TM_x.

Example : Dans le cas où M est un espace vectoriel, son fibré tangent est trivial et en particulier l'ensemble des subsections de TM sont de la forme $m \in M \mapsto (m, f(m)) \in M \times \mathbb{R}^{n_M}$ et s'identifient aux applications de classe $\mathscr{C}^\infty$ de M dans $\mathbb{R}^{n_M}$: $\mathscr{C}^\infty(M, \mathbb{R}^{n_M})$.
La définition précédente est donc cohérente avec le cadre vectoriel.

Par la suite, si $X \in \mathscr{C}^\infty(TM)$ est un champ de vecteur sur M, on note X_x sa valeur en un point x.

Definition A.12. *On appelle* dérivation *sur une variété M une application linéaire de $\mathscr{C}^\infty(M)$ dans lui-même vérifiant l'identité de Leibniz :*

$$\forall f,g \in \mathscr{C}^\infty(M), \quad \delta(f.g) = f.\delta(g) + g.\delta(f)\,.$$

On note $\mathcal{D}(M)$ l'ensemble des dérivations de M.

On associe à tout champ de vecteur X de M l'application

$$\begin{array}{rccc} L_X\ : & \mathscr{C}^\infty(M) & \to & \mathscr{C}^\infty(M) \\ & f & \mapsto & \Big(L_X f\ :\ x\ \mapsto\ L_X f(x) = T_x f \cdot X_x\Big) \end{array}$$

dite *dérivation associée à X*.

Theorem A.13

L'application $L\ :\ X \mapsto L_X$ est une bijection de $\mathscr{C}^\infty(TM)$ sur l'ensemble des dérivations de M.

Ainsi, on peut voir les champs de vecteurs comme des dérivations. Ce point de vue justifie l'abus de notation classique suivant : Soit $(\mathcal{U}, \phi)$ une carte de M. On notera encore $\frac{\partial}{\partial x_i}$ – ou parfois $\mathrm{d}x_i$ – le champ de vecteur sur $\mathcal{U}$ image par ϕ^{-1} du champ de

vecteur $\frac{\partial}{\partial x_i}$ sur $\phi(\mathcal{U})$. Cette écriture, choix de la carte dépendante, permet d'écrire tout champ de vecteur X sur $\mathcal{U}$ de manière unique comme :

$$X = \sum_{i=1}^{n_M} X^i \frac{\partial}{\partial x_i} \quad \text{où} \quad X^i \in \mathscr{C}^\infty(\mathcal{U})\,.$$

En particulier, on récupère une écriture en terme de coordonnées des champs de vecteurs sur une variétés.

L'ensemble des dérivations $\mathcal{D}(M)$ est naturellement muni d'une structure linéaire. Néanmoins, la composée de deux dérivations n'est pas nécessairement une dérivation. En effet, soit δ et δ' deux dérivations sur M. Alors :

$$\begin{aligned} \forall f, g \in \mathcal{C}^\infty(M), \quad \delta\delta'(fg) &= \delta\left(f.\delta'(g) + \delta'(f).g\right) \\ &= \delta(f).\delta'(g) + f.\delta\delta'(g) + \delta\delta'(f).g + \delta'(f).\delta(g)\,. \end{aligned}$$

Mais, on a le résultat suivant :

Lemma 1.14.1. Si δ et δ' sont deux dérivations sur M alors $\delta \circ \delta' - \delta' \circ \delta$ est une dérivation sur M.

Definition A.14. *Le* crochet *de deux champs de vecteurs X et Y, noté $[X, Y]$ est le champ de vecteur associé à la dérivation $L_X L_Y - L_Y L_X$.*

Proposition A.15

Soit X et Y deux champs de vecteurs sur M s'écrivant dans des coordonnées locales

$$X = \sum_{i=1}^{n_M} X^i \frac{\partial}{\partial x_i} \quad \text{et} \quad Y = \sum_{i=1}^{n_M} Y^i \frac{\partial}{\partial x_i}\,.$$

Alors,

$$[X, Y] = \sum_{i=1}^{n_M} Z^i \frac{\partial}{\partial x_i} \quad \text{où} \quad Z^i = \sum_{j=1}^{n_M} \left(X^j \frac{\partial Y^i}{\partial x^j} - Y^j \frac{\partial X^j}{\partial x^j} \right)\,.$$

Lemma 1.16.1 (Lemme de Jacobi)**.** Pour tout champ de vecteurs X, Y et Z sur M,

$$[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0\,.$$

II. Cadre riemannien

Dans tout ce qui suit M désigne une variété différentiable de dimension n.

II – 1. Métriques riemanniennes

Pour tout $m \in M$, T_mM étant un espace vectoriel, on peut le munir d'un produit scalaire $\left\langle . \middle| . \right\rangle_m$. Ainsi, on peut espérer transporter la structure courbe de la variété M sur l'espace linéaire T_mM et récupérer la notion de distance que l'on connait bien. La notion de variété riemannienne repose sur cette idée.

On rappelle que si $X : \mathscr{C}^\infty(M) \to \mathscr{C}^\infty(M)$ est un champ de vecteur sur M alors l'application $X_m : \mathscr{C}^\infty(M) \to \mathbb{R}$ définie par

$$\forall \phi \in \mathscr{C}^\infty(M), \quad X_m(\phi) = (X(\phi))(m)$$

est un vecteur tangent, *i.e* $X_m \in T_mM$.

Definition A.16. *Une* structure riemannienne *sur M est la donnée, en chaque point m de M d'un produit scalaire $\langle . \,|\, . \rangle_m$ de T_mM tel que, pour tout champ de vecteur X, Y de M l'application $m \mapsto \langle X_m \,|\, Y_m \rangle_m$ est une fonction de classe $\mathscr{C}^k$, $k \in \mathcal{N}$.*

Ici, on prend $k = +\infty$ pour ne pas s'embêter avec des questions de régularités et on note classiquement $\|.\|_m$ la norme issue du produit scalaire $\langle \,|\, \rangle_m$.

Soit $(\mathcal{U}, \phi)$ une carte de M dans laquelle tout point de $\mathcal{U}$ se décompose en les coordonnées $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. Alors, tout vecteur tangent en $m \in \mathcal{U}$ s'exprime comme combinaison linéaire des $\partial_{x_i}(m) := \phi^{-1}\left(\frac{\partial \phi}{\partial x_i}(m)\right)$ où $\frac{\partial}{\partial x_i}$ désigne la dérivée par rapport au $i^{\text{ème}}$ vecteur de base dans $\mathbb{R}^n$, *i.e* $\forall \xi, \zeta \in T_mM$,

$$\exists (\xi_i)_{i\in[\![1,n]\!]}, (\zeta_i)_{i\in[\![1,n]\!]}, \quad \xi = \sum_{i=1}^n \xi_i \partial_{x_i}(m) \quad ; \quad \zeta = \sum_{i=1}^n \zeta_i \partial_{x_i}(m)\,.$$

La métrique en m est alors donnée par :

$$g_m(\xi, \zeta) = \sum_{i,j=1}^n g_{i,j}(m)\xi_i\zeta_j \quad \text{avec} \quad g_{i,j}(m) = g\left(\partial_{x_i}(m), \partial_{x_j}(m)\right)\,.$$

Autrement dit :

$$g = \sum_{i,j=1}^n g_{i,j}\partial_{x_i} \otimes \partial_{x_j} \quad \text{noté abusivement par la suite} \quad \sum_{i,j=1}^n g_{i,j}\partial_{x_i}\partial_{x_j}\,.$$

Remarquons que la construction ci-dessus est équivalente à se donner, pour tout $m \in \mathcal{U}$, une matrice symétrique définie positive g_m dont les coefficients sont des fonctions $\mathscr{C}^\infty$ de m telle que :

$$\forall \xi, \zeta \in\ T_mM, \quad \langle \xi \,|\, \zeta \rangle_m = {}^t\xi g_m \zeta$$

Definition A.17. *On appelle* métrique riemannienne *et on note g l'application*

$$\begin{array}{rccl} g: & M & \to & \displaystyle\bigsqcup_{m\in M}\Big(T_mM\times T_mM\Big)^{\mathbb{R}} \\ & m & \mapsto & g_m\colon (\xi,\zeta)\mapsto \langle \xi \,|\, \zeta\rangle_m \text{ tel que } \xi,\zeta\in T_mM\,. \end{array}$$

Theorem A.18

Pour toute variété M, il existe au moins une métrique riemannienne.

Proposition A.19

Soit $(\mathcal{U},\phi)$ et $(\mathcal{U}',\phi')$ deux cartes locales de M dans lesquelles les coordonnées locales s'écrivent respectivement (x_i) et (x'_i) et où est la métrique est donnée par $(g_{i,j})$ et $(g'_{i,j})$. Alors

$$(g'_{i,j}) = {}^t\Phi^{-1}(g_{i,j})\Phi^{-1}$$

où Φ désigne la matrice jacobienne de la fonction de transition $\phi'\circ\phi^{-1}$.

On est ainsi à même de remplir notre objectif, *i.e* de mesurer des déplacements le long de la variété M :

Definition A.20. *Soit $\gamma:[0,a]\to M$ une courbe continue et $\mathscr{C}^1$ par morceaux : soit en particulier $(0=a_0<a_1<\ldots<a_n=a)$ la partition de $[0,a]$ telle que γ soit $\mathscr{C}^1$ sur chacun des $]a_k,a_{k+1}[$. On définit alors la* longueur *de γ par :*

$$l(\gamma)=\sum_{i=0}^{n-1}\int_{a_i}^{a_{i+1}}\left\|\frac{\mathrm{d}\gamma}{\mathrm{d}t}(\gamma(t))\right\|_{\gamma(t)}\mathrm{d}t\,.$$

En particulier, si γ est $\mathscr{C}^1$ sur $]0,a[$, $l(\gamma)=\int_0^a\|\frac{\mathrm{d}\gamma}{\mathrm{d}t}(\gamma(t))\|_{\gamma(t)}\,\mathrm{d}t$.

Example : On se place sur $\mathbb{R}^2$ munit de sa structure euclidienne. Soit $\gamma:[0,\pi]\to\mathbb{R}^2$ définie par $\gamma(t)=e^{it}$, autrement dit γ parcourt $\mathbb{S}^1$ dans le sens trigonométrique. Dans une carte évidente on peut écrire γ comme $\gamma(t)=(\cos(t),\sin(t))$ d'où :

$$\big\langle\gamma'(t)\,\big|\,\gamma'(t)\big\rangle_{\gamma(t)}=\sin^2t+\cos^2t=1\quad\text{et}\quad l(\gamma)=\int_0^\pi\mathrm{d}t=\pi$$

qui est le résultat attendu. Cette définition semble donc, de premier abord, cohérente.

II – 2. Connexion de Levi-Civita

Nous avons dans les subsections précédentes défini des espaces tangents en chacun des points de notre variété. On aimerait à présent pourvoir les relier les uns aux autres afin de pouvoir étudier des champs de vecteurs tangents comme des objets "continus" et notamment les dériver, dans un sens à préciser.

Definition A.21. *Une* connexion *sur une variété (lisse) M est une application bilinéaire ∇ de $\mathscr{C}^\infty(TM)\times\mathscr{C}^\infty(TM)$ dans $\mathscr{C}^\infty(TM)$ telle que :*
$\forall X,Y\in\mathscr{C}^\infty(TM)$ et $f\in\mathscr{C}^\infty(M)$,

$$\nabla_{fX}Y = f\nabla_X Y \quad et \quad \nabla_X(fY) = (X\cdot f)Y + f\nabla_X Y$$

où $X\cdot f$ désigne la dérivée directionnelle de f dans la direction X est définie comme précédemment par :

$$\begin{array}{rccc} X\cdot f\ : & M & \to & M \\ & m & \mapsto & \mathcal{D}_m f\cdot X_m\,. \end{array}$$

On parle également de dérivée covariante.

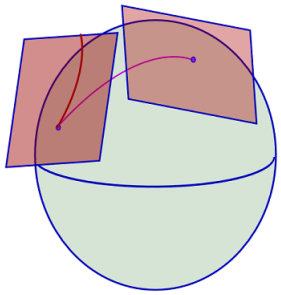


Figure *A.3* – Connexion affine sur la sphère

Definition A.22. *On définit la* torsion *de la connexion ∇ par*

$$\forall X,Y\in\mathscr{C}^\infty(TM),\quad T^\nabla(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y]\,.$$

Si ∇ est de torsion nulle, on dit qu'elle est symétrique *ou sans torsion.*

Theorem A.23

Il existe, pour toute variété riemannienne une unique connexion symétrique consistante avec la métrique, *i.e* telle que pour tout champs de vecteur X, Y, Z,

$$Z\cdot g(X,Y) = g\left(\nabla_Z X, Y\right) + g\left(X, \nabla_Z Y\right)\,.$$

On dit aussi que g est parallèle pour cette connexion et on l'appelle *connexion de Levi-Civita* ou connexion canonique pour la métrique g.

Proposition A.24

Soit g une métrique riemannienne, X et Y deux champs de vecteurs définis

dans des coordonnées locales par :

$$g = \sum_{i,j=1}^{n} g_{i,j}\partial_{x_i}\partial_{x_j} \quad ; \quad X = \sum_{i=1}^{n} X_i \partial x_i \quad ; \quad Y = \sum_{i=1}^{n} Y_i \partial_{x_i}\,.$$

Alors, dans ces mêmes coordonnées,

$$\nabla_X Y = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} X_j \frac{\partial Y_i}{x_j} + \sum_{j,k=1}^{n} \Gamma^i_{jk} X_j Y_k \right) \partial_{x_i}$$

où les Γ^i_{jk} sont définis par la relation $\nabla_{\partial_{x_j}} \partial_{x_k} = \sum_{i=1}^{n} \Gamma^i_{jk} \partial_{x_i}$.
Les Γ^i_{jk} sont appelés *symboles de Christoffel* et, de manière explicite :

$$\Gamma^i_{jk} = \frac{1}{2} \sum_{l=1}^{n} (g_{i,j})^{-1} \left(\frac{\partial g_{k,l}}{\partial x_j} + \frac{\partial g_{l,j}}{\partial x_k} - \frac{\partial g_{j,k}}{\partial x_l} \right) .$$

a. Dérivée covariante le long d'une courbe

Avant d'introduire la notion de géodésique on formalise dans ce paragraphe la notion de déformation infinitésimale d'une courbe.

Definition A.25. *Un* champ de vecteur le long d'une courbe $\gamma : I \subset \mathbb{R} \to M$ *est une courbe* $X : I \to TM$ *telle que* $X(t) \in T_{\gamma(t)}M$ *pour tout* $t \in I$.

Example : De manière naturelle, $\gamma'(t)$ est un champ de vecteur le long de la courbe γ.

Theorem A.26

Soit ∇ la connexion de Levi-Civita associée à M et $\gamma : I \to M$ une courbe sur M. Il existe un unique opérateur que l'on note $\mathcal{D}_t$ défini sur l'espace vectoriel des champs de vecteurs le long de γ tel que, pour tout tel champ X :

1. Pour toute fonction $f : I \to \mathbb{R}$ lisse,
$$\mathcal{D}_t(fX)(t) = f'(t)X(t) + f(t)\mathcal{D}_t X(t)\,.$$
2. S'il existe un voisinage de t_0 dans I tel que X soit la restriction à γ d'un champ de vecteur Y défini sur un voisinage de $\gamma(t_0) \subset M$, alors :
$$\mathcal{D}_t X(t_0) = \left(\nabla_{\gamma'(t_0)} Y\right)_{\gamma(t_0)} .$$

Proposition A.27

Soit X et Y deux champs de vecteurs le long de γ. Alors :

$$\frac{\mathrm{d}t}{\mathrm{d}t} g\left(X(t), Y(t)\right) = g\left(\mathcal{D}_t X(t), Y(t)\right) + g\left(X(t), \mathcal{D}_t Y(t)\right) .$$

II – 3. Géodésiques

Nous commençons par définir formellement la notion de courbe de "plus courte distance", ce qui permet de munir M d'une distance (proposition A.30) compatible avec sa topologie. Enfin, la proposition donne l'équivalence entre la définition suivante et la minimisation d'une fonctionnelle d'énergie.

On reprend les notations du paragraphe précédent et ∇ désigne la connexion de Levi-Civita et $\mathcal{D}_t$ l'opérateur défini en A.26.

Definition A.28. *Une courbe paramétrée γ sur M est une* géodésique *si $\nabla_{\gamma'}\gamma' = 0$, autrement dit $\mathcal{D}_t\gamma' = 0$.*

En particulier, d'après la proposition A.27, la norme de $\gamma'(t)$ est constante.
En effet :

$$\frac{1}{2} g\left(\gamma'(t), \gamma'(t)\right) = g\left(\mathcal{D}_t\gamma', \gamma'\right) = 0 .$$

Et, dans des coordonnées locales où $x(t)$ s'écrit $(x_1(t), \ldots, x_n(t))$, pour tout $i \in [\![1, n]\!]$,

$$\frac{\mathrm{d}^2 x_i}{\mathrm{d}t^2} + \sum_{j,k=1}^{n} \Gamma_{jk}^i\left(x(t)\right) \frac{\mathrm{d}x_k}{\mathrm{d}t} \frac{\mathrm{d}x_j}{\mathrm{d}t} = 0$$

et on a un moyen "simple" de calculer les géodésiques en résolvant l'équation ci-dessus.

Le théorème suivant donne l'existence et l'unicité locale des géodésiques. Non seulement l'existence n'est pas gratuite au vu de la définition mais, surtout, il faut bien prendre garde que cette unicité est uniquement **locale**. Pour s'en convaincre, on peut considérer le cas d'une sphère (figure A.4) : il existe bien-évidement une infinité de géodésiques reliant deux pôles opposés...

Theorem A.29

Soit $m_0 \in M$. Il existe un ouvert $\mathcal{U} \subset M$, $m_0 \in \mathcal{U}$ et $\varepsilon > 0$ tel que pour tout $m \in \mathcal{U}$ et $v \in T_m M$ vérifiant $|v| < \varepsilon$ il existe une unique géodésique $\gamma_v :]-1, 1[\to M$ vérifiant $\gamma_v(0) = m$ et $\gamma_v'(0) = v$.

Example : Dans le cas $M \equiv \mathbb{R}^n$, une courbe γ sur M est une géodésique ssi :

$$\mathcal{D}_t\gamma'(t) = 0 \quad \iff \quad \gamma''(t) = 0 \quad \iff \quad \exists\, x_0, v \in \mathbb{R}^n, \quad \gamma(t) = x_0 + tv\,.$$

Autrement dit, les géodésiques sont les lignes droites paramétrisées à vitesse constance, ce qui est cohérent avec l'idée que l'on se fait de "plus court chemin".

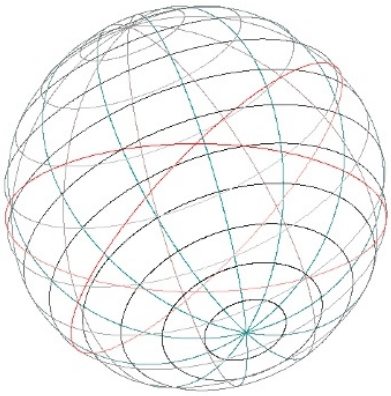


Figure *A.4* – Géodésiques sur la sphère $\mathbb{S}^2$

On rappelle la définition de longueur d'une courbe vu précédemment (définition A.20) et qui, dans le cas d'une courbe $\gamma\colon I \to M$ de classe $\mathscr{C}^1$ donne :

$$l(\gamma) = \int_I \left\| \frac{\mathrm{d}\gamma}{\mathrm{d}t}(\gamma(t)) \right\|_{\gamma(t)} \mathrm{d}t\,.$$

On suppose à présent que M est connexe, ce qui est le cas de l'espace des formes et on définit pour tout $x, y \in M$:

$$d(x,y) := \inf\{l(\gamma) \mid \gamma\ \mathscr{C}^1 \text{ par morceaux reliant } x \text{ à } y\}\,.$$

Proposition A.30

d ainsi définie est une distance sur M qui respecte la topologie de M.

Theorem A.31

Pour tout point m_0 de M il existe un voisinage $\mathcal{U}$ de m_0 et un $\varepsilon > 0$ tel que pour tout $x, y \in \mathcal{U}$ il existe une unique géodésique γ de longueur au plus ε reliant x et y. De plus, $l(\gamma) = d(x,y)$.

Plus précisément, on peut montrer que les géodésiques minimisent *localement* la distance. Encore une fois, ce résultat est uniquement local.

Definition A.32. *Soit $\gamma\colon I \to M$ une courbe $\mathscr{C}^1$ par morceau sur M. On définit l'énergie de γ par la quantité*

$$E(\gamma) = \frac{1}{2}\int_I \left\|\frac{\mathrm{d}\gamma}{\mathrm{d}t}\right\|_{\gamma(t)} \mathrm{d}t$$

quitte à considérer, comme pour la distance, une subdivision sur laquelle γ est $\mathscr{C}^1$...

Proposition A.33

Soit $\gamma_0\colon [a,b] \to M$ une courbe $\mathscr{C}^1$ par morceaux.

1. Si γ_0 minimise la distance de $\gamma_0(a)$ à $\gamma_0(b)$, *i.e* si $d(\gamma_0(a), \gamma_0(b)) = l(\gamma_0)$, et si γ_0 est paramétrisée proportionnellement à "$arclength$", alors γ_0 est une géodésique.

2. Si pour toute courbe γ reliant $\gamma_0(a)$ et $\gamma_0(b)$ $E(\gamma)) \leqslant E(\gamma_0)$, alors γ_0 est une géodésique. En particulier, elle est alors paramétrisée proportionnellement à "$arclength$".

Ce résultat est important pour l'appariement de forme car, le plus souvent, on préfère voir la distance entre deux formes A et B comme un coût de déformation pondérer par une contrainte d'attache aux données.

– Annexe B –

Résultats complémentaires sur les courants

La présente annexe redonne, en français, des résultats élémentaires sur les courants.

Contents

On rappelle ici quelques résultats fondamentaux sur les courants, dont une des utilisations dans le cadre de l'anatomie computationelle est « l'encodage de formes » comme décrit au chapitre 4. On se base principalement sur la thèse de doctorat de Glaunès (2005) et les notes de cours de Viterbo (2013).

I. Formes extérieures

Soit E un espace vectoriel de dimension n sur $\mathbb{R}$. On note $\otimes^p E^*$ l'ensemble des formes p-linéaires sur E. C'est un espace vectoriel sur $\mathbb{R}$.

Definition B.1. *Une* forme *p-linéaire η sur E est dite* alternée *si pour $i \neq j$ on a*

$$\forall x_1, \dots, x_p \in E, \qquad \eta(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_p) \\ = -\eta(x_1, \dots, x_{i-1}, x_j, x_{i+1}, \dots, x_{j-1}, x_i, x_{j+1}, \dots, x_p)\,.$$

On note $\Lambda^p E^*$ l'espace des formes p-linéaires alternées sur E, dites aussi formes extérieures.

Example :
- Par convention, $\Lambda^0 E^* = \mathbb{R}$;
- L'espace $\Lambda^1 E^*$ s'identifie au dual E^* de E, *i.e* à l'espace des formes linéaires sur E. Il est donc en particulier de dimension n ;
- Le déterminant dans une base est un élément de $\Lambda^n E^*$.

Proposition B.2

Une forme p-linéaire η est alternée si et seulement si pour toute famille liée $(x_1, \dots, x_p) \in E^p$, $\eta(x_1, \dots, x_p) = 0$.

Proposition B.3

Soit $\eta \in \Lambda^p E^*$. On a :

$$\forall \sigma \in \mathfrak{S}_p, \quad \forall x_1, \dots, x_p \in E, \quad \eta(x_1, \dots, x_p) = \varepsilon(\sigma)\, \eta(x_{\sigma(1)}, \dots, x_{\sigma(p)})$$

où ε est le morphisme signature.

Definition B.4. *Soit $\alpha \in \otimes^p E^*$, on appelle* partie alternée *de α et on note $P(\alpha)$ la forme extérieure vérifiant :*

$$\forall x_1, \dots, x_p \in E, \quad P(\alpha)(x_1, \dots, x_p) = \frac{1}{p!} \sum_{\sigma \in \mathfrak{S}_p} \varepsilon(\sigma) \alpha(x_{\sigma(1)}, \dots, x_{\sigma(p)})\,.$$

On vérifie facilement que $P(\alpha)$ est bien dans $\Lambda^p E^*$. On peut maintenant définir un « produit » sur les forme extérieures :

Definition B.5. *Soit $\alpha \in \Lambda^p E^*$ et $\beta \in \Lambda^q E^*$. On appelle* produit extérieur *de α et β et on note $\alpha \wedge \beta$ la $(p+q)$-forme extérieure qui vérifie : $\forall x_1, \dots, x_{p+q} \in E$, ,*

$$\alpha \wedge \beta(x_1, \dots, x_{p+q}) = \frac{1}{p!\, q!} \sum_{\sigma \in \mathfrak{S}_{p+q}} \varepsilon(\sigma) \alpha(x_1, \dots, x_p) \beta(x_{p+1}, \dots, x_{p+q}) \,.$$

Proposition B.6

Soit $\alpha \in \Lambda^p E^*$, $\beta \in \Lambda^q E^*$ et $\gamma \in \Lambda^r E^*$. Alors :

$$\alpha \wedge \beta = (-1)^{p\,q} \beta \wedge \alpha \qquad \text{et} \qquad \alpha \wedge (\beta \wedge \gamma) = (\alpha \wedge \beta) \wedge \gamma \,.$$

Le produit extérieur est donc associatif et « presque » commutatif. Il est également distributif. Le produit extérieur permet de doter d'une structure d'algèbre l'ensemble suivant que l'on appelle alors *algèbre extérieure* :

$$\Lambda^\bullet E^* := \bigoplus_{p=0}^{+\infty} \Lambda^p E^* \,.$$

Theorem B.7 (*Structure des formes extérieures*)

Soit $(e_1, \dots, e_p)$ une base de E et $(\mathrm{d}x_1, \dots, \mathrm{d}x_n)$ sa base duale. L'espace vectoriel $\Lambda^p E^*$ est engendré par les éléments de la forme

$$\mathrm{d}x_{i_1} \wedge \cdots \wedge \mathrm{d}x_{i_p}, \quad \text{où} \quad 1 \leqslant i_1 < \cdots < i_p \leqslant n \,.$$

Par conséquent :

$$\dim_{\mathbb{R}}(\Lambda^p E^*) = \binom{n}{p} \,.$$

Example : Le déterminant dans la base $(e_1, \dots, e_n)$ s'écrit $\det = \mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_n$.

II. Formes différentielles et courants

Definition B.8. *On appelle* p-forme différentielle *sur E, ou forme différentielle de degré p sur E, toute application continue*

$$\omega \colon E \to \Lambda^p E^* \,.$$

L'ensemble des p-formes différentielles extérieures se note $\Omega^p(E)$.

Example : 1. $\Omega^0(E)$ est l'ensemble des fonctions continues de E dans $\mathbb{R}$;

2. Si f est une fonction de classe $\mathscr{C}^1$ de E dans $\mathbb{R}$, alors sa différentielle $\mathrm{d}f$ s'assimile à un élément de $\Omega^1(E)$.

D'après le théorème de structure des formes extérieures, un élément $\omega \in \Omega^p(E)$ peut s'écrire sous la forme

$$\omega(y) = \sum_{i_1<\cdots<i_p} g_{i_1,\ldots,i_p}(y)\,\mathrm{d}x_{i_1} \wedge \cdots \wedge \mathrm{d}x_{i_p}\,.$$

Dans le cas d'une fonction continument différentiable f on retrouve alors pour la 1-forme différentielle $\mathrm{d}f$ l'écriture

$$\mathrm{d}f(y) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(y)\,\mathrm{d}x_i\,.$$

Definition B.9. *Notons $\Omega_0^p(E)$ l'ensemble des p-formes différentielles à support compact. On appelle* courant *au sens de De Rham, ou p-courant tout élément de l'espace dual $\left(\Omega_0^p(E)\right)^*$.*

Bibliographie

Citing pages are listed after each reference.

[1] William K Allard. On the first variation of a varifold. *Annals of mathematics*, pages 417–491, 1972. [Page 99.]

[2] Stéphanie Allassonnière and Juliette Chevallier. A new class of EM algorithms. escaping local minima and handling intractable sampling. Preprint, Hal-02044722, 2019. [Pages 28, 55, and 131.]

[3] Stéphanie Allassonnière and Laurent Younes. A stochastic algorithm for probabilistic independent component analysis. *The Annals of Applied Statistics*, 6(1) :125–160, 03 2012. [Pages 153, 160, 163, 164, and 167.]

[4] Stéphanie Allassonnière, Alain Trouvé, and Laurent Younes. Geodesic shooting and diffeomorphic matching via textured meshes. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 365–381. Springer, 2005. [Pages 14 and 42.]

[5] Stéphanie Allassonnière, Yali Amit, and Alain Trouvé. Toward a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69(1) :3–29, 2007. [Page 78.]

[6] Stéphanie Allassonnière, Estelle Kuhn, and Alain Trouvé. Construction of Bayesian deformable models via a stochastic approximation algorithm : A convergence study. *Bernoulli*, 16(3) :641–678, 2010. [Pages 9, 37, 61, 70, 76, 141, 177, and 182.]

[7] Frederick J Almgren. *Plateau's problem : an invitation to varifold geometry*, volume 13. American Mathematical Soc., 1966. [Page 99.]

[8] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2) :5–43, 2003. [Pages 9 and 37.]

[9] Christophe Andrieu, Éric Moulines, and Pierre Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on Control and Optimization*, 44(1) : 283–312, 2006. [Pages 149, 177, and 182.]

[10] Vladimir Arnold. Sur la géométrie différentielle des groupes de lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. In *Annales de l'institut Fourier*, volume 16(1), pages 319–361, 1966. [Pages 16 and 44.]

[11] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3) :337–404, 1950. [Pages 13 and 41.]

[12] Yves F. Atchadé. An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in Applied Probability*, 8(2) :235–254, 2006. [Pages 118, 177, and 182.]

[13] Hagai Attias. Independent factor analysis. *Neural Computation*, 11(4) :803–851, 1999. [Pages 131, 153, 160, and 163.]

[14] Douglas M Bates and Donald G Watts. *Nonlinear regression analysis and its applications*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1988. [Pages 7 and 35.]

[15] 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. [Pages 11, 12, 39, and 114.]

[16] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3) :561 – 575, 2003. [Page 154.]

[17] Murat Bilgel, Jerry L Prince, Dean F Wong, Susan M Resnick, and Bruno M Jedynak. A multivariate nonlinear mixed effects model for longitudinal image analysis : Application to amyloid imaging. *Neuroimage*, 134 :658–670, 2016. [Pages 23 and 50.]

[18] Jean Claude Biscarat. Almost sure convergence of a class of stochastic algorithms. *Stochastic processes and their applications*, 50(1) :83–99, 1994. [Page 139.]

[19] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference : A review for statisticians. *Journal of the American Statistical Association*, 112(518) :859–877, 2017. [Page 130.]

[20] Salomon Bochner. Integration von funktionen, deren werte die elemente eines vektorraumes sind. *Fundamenta Mathematicae*, 20(1) :262–176, 1933. [Pages 12 and 40.]

[21] Alexandre Bône, Olivier Colliot, and Stanley Durrleman. Learning distributions of shape trajectories from longitudinal datasets : A hierarchical model on a manifold of diffeomorphisms. In *Computer Vision and Pattern Recognition*, Salt Lake City, United States, 2018. [Pages 24, 51, 69, 113, 114, 115, 174, and 180.]

[22] 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, 2018. [Pages 18 and 46.]

[23] Alexandre Bône, Maxime Louis, Olivier Colliot, and Stanley Durrleman. Learning low-dimensional representations of shape data sets with diffeomorphic autoencoders. In *International Conference on Information Processing in Medical Imaging*, pages 195–207. Springer, 2019. [Pages 25 and 52.]

[24] William M Boothby. *An introduction to differentiable manifolds and Riemannian geometry*, volume 12. Academic press, 1986. [Page 94.]

[25] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011. [Pages 9 and 37.]

[26] Olivier Cappé, Éric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, 2005. [Page 129.]

[27] Joshua Cates, P Thomas Fletcher, Martin Styner, Martha Shenton, and Ross Whitaker. Shape modeling and analysis with entropy-based particle systems. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 333–345. Springer, 2007. [Pages 17 and 44.]

[28] Gilles Celeux and Jean Diebolt. The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational statistics quarterly*, 2 :73–82, 1985. [Pages 8, 36, 129, and 138.]

[29] Gilles Celeux, Didier Chauveau, and Jean Diebolt. On stochastic versions of the EM algorithm. Technical report, INRIA, 1995. [Pages 9, 36, and 139.]

[30] Phylinda LS Chan, Philippe Jacqmin, Marc Lavielle, Lynn McFadyen, and Barry Weatherley. The use of the SAEM algorithm in MONOLIX software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic HIV subjects. *Journal of pharmacokinetics and pharmacodynamics*, 38(1) :41–61, 2011. [Pages 10, 37, and 130.]

[31] Benjamin Charlier, Nicolas Charon, and Alain Trouvé. The fshape framework for the variability analysis of functional shapes. *Foundations of Computational Mathematics*, 17 (2) :287–357, 2017. [Pages 13, 41, and 60.]

[32] Nicolas Charon. *Analysis of geometric and functional shapes withextensions of currents : applications to registration and atlas estimation*. PhD thesis, École Normale Supérieure de Cachan, 2013. [Pages 99, 101, and 102.]

[33] Nicolas Charon and Alain Trouvé. The varifold representation of nonoriented shapes for diffeomorphic registration. *SIAM Journal on Imaging Sciences*, 6(4) :2547–2580, 2013. [Pages 13, 27, 41, 53, 54, 60, 91, 98, 99, 102, 103, 104, 105, and 113.]

[34] Nicolas Charon and Alain Trouvé. Functional currents : a new mathematical tool to model and analyse functional shapes. *Journal of mathematical imaging and vision*, 48(3) :413–431, 2014. [Page 96.]

[35] Juliette Chevallier, Stéphane Oudard, and Stéphanie Allassonnière. Learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. In *Neural Information Processing Systems*, Long Beach, CA, USA, 2017. [Pages 27, 28, 53, 55, and 65.]

[36] Juliette Chevallier, Vianney Debavelaere, and Stéphanie Allassonnière. A coherent framework for learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. Preprint, Hal-01646298, 2019. [Pages 27, 28, 53, 54, 55, 61, and 65.]

[37] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. Applied multiple regression. *Correlation Analysis for the Behavioral Sciences*, 2, 1983. [Pages 6 and 34.]

[38] Thierry Colin, Angelo Iollo, Damiano Lombardi, and Olivier Saut. Prediction of the evolution of thyroidal lung nodules using a mathematical model. *ERCIM News*, 82 :37–38, 2010. [Page 59.]

[39] Emmanuelle Comets, Karl Brendel, and France Mentré. Model evaluation in nonlinear mixed effect models, with applications to pharmacokinetics. *Journal de la Société Française de Statistique*, 151(1) :106–128, 2010. [Pages 7 and 35.]

[40] Pierre Comon. Independent component analysis, a new concept ? *Signal Processing*, 36 (3) :287 – 314, 1994. Higher Order Statistics. [Page 163.]

[41] Wentworth D'Arcy Thompson. *On Growth and Form.* Cambridge university press, 1942. [Pages 10, 11, and 38.]

[42] Manasi Datar, Joshua Cates, P Thomas Fletcher, Sylvain Gouttard, Guido Gerig, and Ross Whitaker. Particle based shape regression of open surfaces with applications to developmental neuroimaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 167–174. Springer, 2009. [Pages 17 and 44.]

[43] Manasi Datar, Prasanna Muralidharan, Abhishek Kumar, Sylvain Gouttard, Joseph Piven, Guido Gerig, Ross Whitaker, and P Thomas Fletcher. Mixed-effects shape models for estimating longitudinal changes in anatomy. In *International Workshop on Spatio-temporal Image Analysis for Longitudinal and Time-Series Image Data*, pages 76–87. Springer, 2012. [Pages 17 and 44.]

[44] Georges De Rham. *Variétés différentiables : formes, courants, formes harmoniques*, volume 3. Editions Hermann, 1973. [Page 93.]

[45] Vianney Debavelaere, Alexandre Bône, Stanley Durrleman, and Stéphanie Allassonnière. Clustering of longitudinal shape data sets using mixture of separate or branching trajectories. to appear in MICAI 2019, Hal-02103355, 2019. [Pages 25, 52, 167, 174, 177, 180, and 182.]

[46] Deformetrica, software. Learn from shapes. Statistical analysis of images and meshes. Institut du Cerveau et de la Moelle épinière (ICM). Installation instructions, tutorials and examples can be found at `http://www.deformetrica.org`. [Pages 18 and 46.]

[47] I Delor, J-E Charoin, R Gieschke, S Retout, and P Jacqmin. Modeling Alzheimer's disease progression using disease onset time and disease trajectory concepts applied to cdr-sob scores from adni. *CPT : pharmacometrics & systems pharmacology*, 2(10) :1–10, 2013. [Pages 20 and 47.]

[48] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1) :94–128, 1999. [Pages xxi, 8, 9, 27, 36, 54, 61, 70, 129, 130, 133, 136, 137, 138, 139, 140, 141, 142, 146, 148, 149, 150, 151, 167, 169, 176, and 181.]

[49] Arthur 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 (Statistical Methodology)*, 39(1) :1–38, 1977. [Pages 8, 36, 61, 129, and 136.]

[50] Loïc Devilliers, Stéphanie Allassonnière, Alain Trouvé, and Xavier Pennec. Template estimation in computational anatomy : Fréchet means top and quotient spaces are not consistent. *SIAM Journal on Imaging Sciences*, 10(3) :1139–1169, 2017. [Pages 21 and 48.]

[51] Loïc Devilliers, Xavier Pennec, and Stéphanie Allassonnière. Inconsistency of template estimation with the fréchet mean in quotient space. In *International Conference on Information Processing in Medical Imaging*, pages 16–27. Springer, 2017. [Pages 21 and 48.]

[52] Francisco J Diaz, Hung-Wen Yeh, and Jose de Leon. Role of statistical random-effects linear models in personalized medicine. *Current Pharmacogenomics and Personalized Medicine (Formerly Current Pharmacogenomics)*, 10(1) :22–32, 2012. [Pages 7 and 35.]

[53] Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott L Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002. [Pages 6 and 34.]

[54] Manfredo P Do Carmo. *Differential Geometry of Curves and Surfaces*. Courier Dover Publications, 1976. [Page 189.]

[55] 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. [Pages 11 and 39.]

[56] 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, 2011. [Pages 14, 18, 25, 41, 45, and 51.]

[57] Stanley Durrleman. *Statistical models of currents for measuring the variability of anatomical curves, surfaces and their evolution*. PhD thesis, Université de Nice Sophia Antipolis, 2010. [Pages 13, 41, 93, 95, 96, and 97.]

[58] Stanley Durrleman, Xavier Pennec, Guido Gerig, Alain Trouvé, and Nicholas Ayache. Spatiotemporal atlas estimation for developmental delay detection in longitudinal datasets. Technical report, INRIA, 2009. [Pages 19, 20, 21, 22, 46, 47, 48, and 49.]

[59] 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. [Pages 17, 18, 25, 44, 45, and 51.]

[60] Stanley Durrleman, Xavier Pennec, Alain Trouvé, José Braga, Guido Gerig, and Nicholas Ayache. Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *International journal of computer vision*, 103(1) :22–59, 2013. [Pages 20, 21, 47, 48, 173, and 179.]

[61] Churchill Eisenhart. The assumptions underlying the analysis of variance. *Biometrics*, 3 (1) :1–21, 1947. [Pages 5 and 33.]

[62] Bernard Escudier, Camillo Porta, Mélanie Schmidinger, Nathalie Rioux-Leclercq, Axel Bex, Vincent S. Khoo, Viktor Gruenvald, and Alan Horwich. Renal cell carcinoma : ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 27 (suppl 5) :v58–v68, 2016. [Page 107.]

[63] Herbert Federer. *Geometric measure theory*. Springer, 1969. [Pages 93, 94, 95, and 101.]

[64] James Fishbaugh, Stanley Durrleman, Marcel Prastawa, and Guido Gerig. Geodesic shape regression with multiple geometries and sparse parameters. *Medical image analysis*, 39 : 1–17, 2017. [Pages 17, 18, 45, and 46.]

[65] Ronald A Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2) :399–433, 1919. [Pages 5 and 33.]

[66] Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 2006. [Pages 6 and 34.]

[67] Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied longitudinal analysis*, volume 998. John Wiley & Sons, 2012. [Pages 6, 17, 34, and 44.]

[68] Thomas Fletcher. Geodesic regression on Riemannian manifolds. In *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy-Geometrical and Statistical Methods for Modelling Biological Shape Variability*, pages 75–86, 2011. [Pages 15 and 43.]

[69] Gersende Fort and Eric Moulines. Convergence of the monte carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4) :1220–1259, 2003. [Page 139.]

[70] Jean-Louis Foulley. Algorithme em : théorie et application au modèle mixte. *Journal de la société française de statistique*, 143(3-4) :57–109, 2002. [Pages 8 and 36.]

[71] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001. [Pages 13 and 41.]

[72] Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian Geometry*. Universitext. Springer-Verlag Berlin Heidelberg, 3 edition, 2004. [Pages 15, 17, 22, 43, 44, 49, 60, 66, 67, and 189.]

[73] Christophe Giraud. *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2014. [Pages 18, 23, 46, 50, and 71.]

[74] Joan Glaunès. *Transport par difféomorphismes de points, de mesures et de courants pour la comparaison de formes et l'anatomie numérique*. PhD thesis, Université Paris 13, 2005. [Pages 11, 13, 39, 40, 93, 95, and 203.]

[75] Joan Glaunès, Alain Trouvé, and Laurent Younes. Diffeomorphic matching of distributions : A new approach for unlabelled point-sets and sub-manifolds matching. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2004. [Pages 13 and 41.]

[76] Joan Glaunès, Anqi Qiu, Michael I Miller, and Laurent Younes. Large deformation diffeomorphic metric curve mapping. *International journal of computer vision*, 80(3) :317, 2008. [Pages 13 and 41.]

[77] Ulf Grenander. *General pattern theory-A mathematical study of regular structures*. Clarendon Press, 1993. [Pages 10 and 38.]

[78] Maya R Gupta and Yihua Chen. Theory and use of the EM algorithm. *Foundations and Trends® in Signal Processing*, 4(3) :223–296, 2011. [Page 137.]

[79] Peter Hall and Christopher C. Heyde. *Martingale limit theory and its application*. Probability and mathematical statistics. Academic Press, 1980. [Page 151.]

[80] Xavier A Harrison, Lynda Donaldson, Maria Eugenia Correa-Cano, Julian Evans, David N Fisher, Cecily ED Goodwin, Beth S Robinson, David J Hodgson, and Richard Inger. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6 :e4794, 2018. [Pages 7 and 35.]

[81] David A Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358) :320–338, 1977. [Pages 6 and 34.]

[82] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109, 1970. [Pages 9 and 37.]

[83] Yi Hong, Nikhil Singh, Roland Kwitt, and Marc Niethammer. Time-warped geodesic regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 105–112. Springer, 2014. [Pages 21 and 48.]

[84] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004. [Pages 23, 50, and 114.]

[85] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2) :183–233, 1999. [Page 130.]

[86] Sarang C Joshi and Michael I Miller. Landmark matching via large deformation diffeomorphisms. *IEEE transactions on image processing*, 9(8) :1357–1370, 2000. [Pages 14 and 41.]

[87] Jürgen Jost. *Riemannian geometry and geometric analysis*. Universitext. Springer, 3 edition, 2002. [Pages 15, 43, and 189.]

[88] David G Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2) :81–121, 1984. [Pages 11 and 39.]

[89] Hyunwoo J Kim, Nagesh Adluru, Maxwell D Collins, Moo K Chung, Barbara B Bendlin, Sterling C Johnson, Richard J Davidson, and Vikas Singh. Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2705–2712, 2014. [Pages 15 and 43.]

[90] 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 of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2540–2549, 2017. [Pages 24 and 51.]

[91] Eric Klassen, Anuj Srivastava, Washington Mio, and Shantanu Joshi. Analysis of planar shapes using geodesic paths on shape spaces. *IEEE transactions on pattern analysis and machine intelligence*, 26(3) :372–383, 2004. [Page 60.]

[92] Ender Konukoglu, Olivier Clatz, Bjoern H Menze, Bram Stieltjes, Marc-AndrÉ Weber, Emmanuel Mandonnet, Hervé Delingette, and Nicholas Ayache. Image guided personalization of reaction-diffusion type tumor growth models using modified anisotropic eikonal equations. *IEEE transactions on medical imaging*, 29(1) :77–95, 2010. [Page 59.]

[93] Igor Koval, J-B Schiratti, Alexandre Routier, Michael Bacci, Olivier Colliot, Stéphanie Allassonnière, and Stanley Durrleman. Statistical learning of spatiotemporal patterns from longitudinal manifold-valued networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 451–459. Springer, 2017. [Pages 24 and 51.]

[94] Igor Koval, Jean-Baptiste Schiratti, Alexandre Routier, Michael Bacci, Olivier Colliot, Stephanie Allassonniere, and Stanley Durrleman. Spatiotemporal propagation of the cortical atrophy : Population and individual patterns. *Frontiers in Neurology*, 9, 2018. [Pages 24, 51, 69, 174, and 180.]

[95] Estelle Kuhn and Marc Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM : Probability and Statistics*, 8 :115–131, 2004. [Pages 9, 37, and 141.]

[96] Estelle Kuhn and Marc Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4) :1020–1038, 2005. [Pages 70 and 76.]

[97] Jacques Lafontaine. *Introduction aux variétés différentielles (Collection Grenoble sciences)*. EDP Sciences, 1996. [Page 189.]

[98] Nan Laird, Nicholas Lange, and Daniel Stram. Maximum likelihood computations with repeated measures : application of the EM algorithm. *Journal of the American Statistical Association*, 82(397) :97–105, 1987. [Pages 8 and 36.]

[99] Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4) :963–974, 1982. [Pages 6, 8, 15, 34, 36, 43, and 60.]

[100] Kenneth Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society : Series B*, 57(2) :425–437, 1995. [Pages 8, 36, and 138.]

[101] Marc Lavielle. *Mixed Effects Models for the Population Approach : Models, Tasks, Methods and Tools*. CRC Biostatistics Series. Chapman and Hall, 2014. [Pages 7, 35, 61, 129, and 130.]

[102] Marc Lavielle and Leon Aarons. What do we mean by identifiability in mixed effects models ? *Journal of Pharmacokinetics and Pharmacodynamics*, 43(1) :111–122, February 2016. [Page 89.]

[103] 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. [Pages 10, 37, and 130.]

[104] Marc Lavielle and Eric Moulines. A simulated annealing version of the EM algorithm for non-Gaussian deconvolution. *Statistics and Computing*, 7(4) :229–236, 1997. [Pages 129, 130, and 151.]

[105] Mary J Lindstrom and Douglas M Bates. Newton—raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404) :1014–1022, 1988. [Pages 7 and 35.]

[106] Marco Lorenzi, Xavier Pennec, Giovanni B Frisoni, and Nicholas Ayache. Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images. *Neurobiology of aging*, 36 :S42–S52, 2015. [Pages 20 and 47.]

[107] Maxime Louis, Alexandre Bône, Benjamin Charlier, and Stanley Durrleman. Parallel transport in shape analysis : a scalable numerical scheme. In *International Conference on Geometric Science of Information*, pages 29–37. Springer, 2017. [Pages 22, 49, 175, and 180.]

[108] Maxime Louis, Benjamin Charlier, Paul Jusselin, Susovan Pal, and Stanley Durrleman. A fanning scheme for the parallel transport along geodesics on Riemannian manifolds. *SIAM Journal on Numerical Analysis*, 56(4) :2563–2584, 2018. [Pages 22 and 49.]

[109] Maxime Louis, Raphaël Couronné, Igor Koval, Benjamin Charlier, and Stanley Durrleman. Riemannian geometry learning for disease progression modelling. Hal-02079820, 2019. [Pages 175 and 180.]

[110] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6) :1167–1180, 2012. [Page 130.]

[111] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007. [Pages 8, 36, and 137.]

[112] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wile, 2000. [Page 153.]

[113] Xiao-Li Meng and David Van Dyk. The EM algorithm – an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society : Series B*, 59(3) :511–567, 1997. [Pages 8 and 36.]

[114] Nicholas Metropolis and Stanislaw Ulam. The Monte Carlo method. *Journal of the American statistical association*, 44(247) :335–341, 1949. [Pages 9 and 37.]

[115] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6) :1087–1092, 1953. [Pages 9 and 37.]

[116] Michael I Miller, Alain Trouvé, and Laurent Younes. Geodesic shooting for computational anatomy. *Journal of mathematical imaging and vision*, 24(2) :209–228, 2006. [Pages 14, 16, 42, 44, and 114.]

[117] J. Kevin Milliken and Steven D. Edland. Mixed effect models of longitudinal Alzheimer's disease data : A cautionary note. *Statistics in Medicine*, 19(11-12) :1617–1629, 2000. [Pages 6, 33, and 60.]

[118] Frank Morgan. *Geometric measure theory : a beginner's guide*. Academic press, 1995. [Pages 94 and 95.]

[119] Eric Moulines, Jean-François Cardoso, and Elisabeth Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Acoustics, Speech, and Signal Processing*, volume 5, pages 3617–3620. IEEE, 1997. [Pages 153 and 160.]

[120] Prasanna Muralidharan and P Thomas Fletcher. Sasaki metrics for analysis of longitudinal data on manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1027–1034. IEEE, 2012. [Pages 15, 16, 18, 19, 24, 43, 44, 46, and 51.]

[121] Juan David Ospina, Pascal Benquet, J Correa, and Oscar Acosta. A mixed-effects spatio-temporal model to assess the joint effects of ageing and Alzheimer's disease in gray matter volume. In *MICCAI-NIBAD'12 : Workshop on Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders*, pages 190–7, 2012. [Pages 6 and 34.]

[122] Umberto Picchini and Adeline Samson. Coupling stochastic EM and approximate Bayesian computation for parameter inference in state-space models. *Computational Statistics*, 33 (1) :179–212, 2018. [Pages 28, 54, 130, 143, 167, 176, and 181.]

[123] Islem Rekik, Stéphanie Allassonnière, Olivier Clatz, Ezequiel Geremia, Erin Stretton, Hervé Delingette, and Nicholas Ayache. Tumor growth parameters estimation and source localization from a unique time point : Application to low-grade gliomas. *Computer Vision and Image Understanding*, 117(3) :238–249, 2013. [Page 59.]

[124] B Ribba, NH Holford, P Magni, I Trocóniz, I Gueorguieva, P Girard, C Sarr, M Elishmereni, C Kloft, and LE Friberg. A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis. *CPT : Pharmacometrics & Systems Pharmacology*, 3(5) :1–10, 2014. [Pages 6, 34, and 60.]

[125] Richard Rios, Renaud De Crevoisier, Juan D Ospina, Frederic Commandeur, Caroline Lafond, Antoine Simon, Pascal Haigron, Jairo Espinosa, and Oscar Acosta. Population model of bladder motion and deformation based on dominant eigenmodes and mixed-effects models in prostate cancer radiotherapy. *Medical image analysis*, 38 :133–149, 2017. [Pages 6 and 34.]

[126] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. [Pages 9 and 36.]

[127] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag New York, 1999. [Pages 9, 37, and 76.]

[128] Gareth O. Roberts and Jeffrey S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2) :458–475, 03 2007. [Pages 118, 177, and 182.]

[129] Gareth O. Roberts and Jeffrey S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2) :349–367, 2009. [Pages 118, 177, and 182.]

[130] Alexis Roche. EM algorithm and variants : An informal tutorial, 2012. [Page 137.]

[131] Pierre Roussillon and Joan Alexis Glaunès. Kernel metrics on normal cycles and application to curve matching. *SIAM Journal on Imaging Sciences*, 9(4) :1991–2038, 2016. [Pages 13 and 41.]

[132] F. Rouvière and A. Debreil. *Initiation à la géométrie de Riemann*. Mathématiques en devenir. Calvage et Mounet, 2016. [Page 189.]

[133] 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. [Page 130.]

[134] Olivier Saut, Jean-Baptiste Lagaert, Thierry Colin, and Hassan M Fathallah-Shaykh. A multilayer grow-or-go model for gbm : effects of invasive cells and anti-angiogenesis on growth. *Bulletin of mathematical biology*, 76(9) :2306–2333, 2014. [Page 59.]

[135] Henry Scheffe. A "mixed model" for the analysis of variance. *The Annals of Mathematical Statistics*, pages 23–36, 1956. [Pages 6 and 34.]

[136] Jean-Baptiste Schiratti. *Models and algorithms to learn spatiotemporal changes from longitudinal manifold-valued observations*. PhD thesis, École Polytechnique, 2016. [Pages 7, 23, 35, and 50.]

[137] Jean-Baptiste Schiratti, Stéphanie Allassonnière, Olivier Colliot, and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *Neural Information Processing Systems*, number 28 in Advances in Neural Information Processing Systems, Montréal, Canada, 2015. [Pages i, iii, 3, 7, 21, 22, 23, 24, 25, 26, 27, 31, 33, 35, 48, 49, 50, 51, 52, 53, 60, 61, 63, 65, 69, 89, 113, 114, 115, 173, 174, 175, 179, and 180.]

[138] 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. *Journal of Machine Learning Research*, 18(133) :1–33, 2017. [Pages i, iii, 3, 7, 21, 22, 23, 24, 25, 26, 27, 31, 33, 35, 48, 49, 50, 51, 52, 53, 60, 61, 63, 65, 69, 89, 173, 174, 175, 179, and 180.]

[139] Seg3D, software. Volumetric Image Segmentation and Visualization. Center for Integrative Biomedical Computing (CIBC), Scientific Computing and Imaging Institute (SCI), Download from : `http://www.seg3d.org`. [Pages 17 and 45.]

[140] 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. [Pages 7 and 35.]

[141] Nikhil Singh, Jacob Hinkle, Sarang Joshi, and P Thomas Fletcher. A hierarchical geodesic model for diffeomorphic longitudinal shape analysis. In *International Conference on Information Processing in Medical Imaging (IPMI)*, pages 560–571. Springer, 2013. [Pages 16, 18, 19, 44, and 46.]

[142] Nikhil Singh, Jacob Hinkle, Sarang Joshi, and P Thomas Fletcher. An efficient parallel algorithm for hierarchical geodesic models in diffeomorphisms. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 341–344. IEEE, 2014. [Pages 16, 18, 19, 44, and 46.]

[143] 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. [Pages 17 and 44.]

[144] Jingyong Su, Sebastian Kurtek, Eric Klassen, and Anuj Srivastava. Statistical analysis of trajectories on Riemannian manifolds : bird migration, hurricane tracking and video surveillance. *The Annals of Applied Statistics*, 8(1) :530–552, 2014. [Pages 21, 48, and 60.]

[145] Jingyong Su, Anuj Srivastava, Fillipe DM de Souza, and Sudeep Sarkar. Rate-invariant analysis of trajectories on Riemannian manifolds with application in visual speech recognition. In *Computer Vision and Pattern Recognition*, pages 620–627, 2014. [Pages 21, 48, and 60.]

[146] Monolix, software. MOdèles NOn LInéaires à effets miXtes. Advanced and simple solution for non-linear mixed effects modeling for pharmacometrics. http://lixoft.com/products/monolix/. [Pages 9, 37, 141, 152, 167, 177, and 182.]

[147] 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, and Steve G. Gwyther. New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, 92(3) :205–216, 2000. [Page 107.]

[148] Alain Trouvé and François-Xavier Vialard. Shape splines and stochastic shape evolutions : A second order point of view. *Quarterly of Applied Mathematics*, pages 219–251, 2012. [Pages 17 and 44.]

[149] Alain Trouvé and Laurent Younes. *Shape Spaces*, pages 1759–1817. Springer New York, New York, USA, 2015. [Pages 10, 38, and 60.]

[150] Marc Vaillant and Joan Glaunès. Surface matching via currents. In *Information Processing in Medical Imaging*, volume 3565, Glenwood Springs, USA, 2005. [Pages 13, 27, 41, 53, 54, 60, 91, 105, and 113.]

[151] Adrianus Willem van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000. [Pages 73, 86, 174, and 180.]

[152] Geert Verbeke, Steffen Fieuws, Geert Molenberghs, and Marie Davidian. The analysis of multivariate longitudinal data : A review. *Statistical methods in medical research*, 23(1) : 42–59, 2014. [Pages 6 and 34.]

[153] Claude Viterbo. Géométrie différentielle. Notes de cours, L3, ENS Ulm, 2013. [Page 203.]

[154] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2) :1–305, 2008. [Page 130.]

[155] Greg C. 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. [Pages 8, 36, 129, 138, and 139.]

[156] Yung-Chow Wong. Differential geometry of Grassmann manifolds. *Proceedings of the National Academy of Sciences of the United States of America*, 57(3) :589, 1967. [Page 99.]

[157] Jeff CF. Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, 11(1) :95–103, 1983. [Pages 8, 36, and 136.]

[158] Eric Yang, Michael Farnum, Victor Lobanov, Tim Schultz, Nandini Raghavan, Mahesh N Samtani, Gerald Novak, Vaibhav Narayan, and Allitia DiBernardo. Quantifying the pathophysiological timeline of Alzheimer's disease. *Journal of Alzheimer's Disease*, 26(4) : 745–753, 2011. [Pages 20 and 47.]

[159] Laurent Younes. *Shapes and diffeomorphisms*, volume 171. Springer Science & Business Media, 2010. [Pages 11, 12, 39, and 40.]

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Modèles statistiques et algorithmes stochastiques pour l'analyse de données longitudinales à dynamiques multiples et à valeurs sur des variétés riemanniennes.

Mots Clefs : Géométrie riemannienne, Données longitudinales, Optimisation stochastique, Modèles non-linéaires à effets-mixtes, Algorithmes de type EM, Analyse spatio-temporelle.

Résumé : Par delà les études transversales, étudier l'évolution temporelle de phénomènes connait un intérêt croissant. En effet, pour comprendre un phénomène, il semble plus adapté de comparer l'évolution des marqueurs de celui-ci au cours du temps plutôt que ceux-ci à un stade donné. C'est par exemple le cas pour le suivi de chimiothérapie qui repose de plus en plus sur la compréhension de la progression globale de la maladie que sur que l'état de santé ponctuels des patients.
Les modèles à effets mixtes ont prouvé leur efficacité dans l'étude des données longitudinales, notamment à dans le cadre d'applications médicales. Des travaux récents ont notamment permis l'étude de données complexes, telles que des données anatomiques. L'idée sous-jacente est de modéliser la progression temporelle d'un phénomène par des trajectoires continues dans un espace de mesures, que l'on suppose être une variété riemannienne. Sont alors estimées conjointement une trajectoire moyenne représentative de l'évolution globale de la population, à l'échelle macroscopique, et la variabilité inter-individuelle. Cependant, ces travaux supposent une progression unidirectionnelle et échouent à décrire des situations telles que la sclérose en plaques ou le suivi de chimiothérapie. En effet, pour ces pathologies, vont se succéder des phases de progression, de stabilisation et de remision de la maladie, induisant un changement de la dynamique d'évolution globale.
Le but de cette thèse est de développer des outils méthodologiques et algorithmiques pour l'analyse de données longitudinales, dans le cas de phénomènes dont la dynamique d'évolution est multiple et d'appliquer ces nouveaux outils pour le suivi de chimiothérapie. Nous proposons un modèle non-linéaire à effets mixtes dans lequel les trajectoires d'évolution individuelles sont vues comme des déformations spatio-temporelles d'une trajectoire géodésique par morceaux et représentative de l'évolution de la population. Nous présentons ce modèle sous des hypothèses très génériques afin d'englober une grande classe de modèles plus spécifiques.
L'estimation des paramètres du modèle géométrique est réalisée par un estimateur du MAP dont nous démontrons l'existence et la consistance sous des hypothèses standards. Numériquement, du fait de la non-linéarité de notre modèle, l'estimation est réalisée par une version stochastique de l'algorithme EM, à savoir l'algorithme MCMC-SAEM. La convergence du SAEM vers les maxima locaux de la vraisemblance observée ainsi que son efficacité numérique ont été démontrées. En dépit de cette performance, l'algorithme SAEM est très sensible à ses conditions initiales. Afin de palier ce problème, nous proposons une nouvelle classe d'algorithmes SAEM dont nous démontrons la convergence vers des minima locaux. Cette classe repose sur la simulation par une loi approchée de la vraie loi conditionnelle dans l'étape de simulation. Enfin, en se basant sur des techniques de recuit simulé, nous proposons une version tempérée de l'algorithme SAEM afin de favoriser sa convergence vers des minima globaux.

Title: Statistical models and stochastic algorithms for the analysis of longitudinal Riemanian manifold valued data with multiple dynamic

Keywords: Riemannian geometry, Longitudinal data, Stochastic optimization, Non-linear mixed-effect models, EM-like algorithms, Spatio-temporal analysis.

Abstract: Beyond transversal studies, temporal evolution of phenomena is a field of growing interest. For the purpose of understanding a phenomenon, it appears more suitable to compare the evolution of its markers over time than to do so at a given stage. For instance, efforts in chemotherapy monitoring rely more and more on the understanding of the global disease progression and not only on punctual states of health.
Mixed effects models have proved their efficiency in the study of longitudinal data, especially for medical purposes. Recent works allowed the study of complex data, such as anatomical data. The underlying idea is to model the temporal progression of a phenomenon by continuous trajectories in a space of measurements, which is assumed to be a Riemannian manifold. Then, both a group-representative trajectory and inter-individual variability are estimated. However, these works assume an unidirectional dynamic and fail to encompass situations like multiple sclerosis or chemotherapy monitoring. Indeed, such diseases follow a chronic courses, with phases of worsening, stabilization and improvement, inducing changes in the global dynamic.
This thesis is devoted to the development of methodological tools and algorithms suited for the analysis of longitudinal data arising from phenomena that undergo multiple dynamics and to apply them to chemotherapy monitoring. We propose a nonlinear mixed effects model in which individual trajectories are seen as spatio-temporal deformation of representative piecewise-geodesic trajectory of the global progression. We present this model under very generic assumptions in order to encompass a wide range of more specific models.
The estimation is formulated as a well-defined MAP problem which we prove to be consistent under mild assumptions. Numerically, due to the non-linearity of the proposed model, the estimation of the parameters is performed through a stochastic version of the EM algorithm, namely the MCMC-SAEM algorithm. The convergence of the SAEM algorithm toward local maxima of the observed likelihood has been proved and its numerical efficiency has been demonstrated. However, despite appealing features, the limit position of this algorithm can strongly depend on its starting position. To cope with this issue, we propose a new version of the SAEM in which we do not sample from the exact distribution in the expectation phase of the procedure. We first prove the convergence of this algorithm toward local maxima. Then, with the thought of the simulated annealing, we propose an instantiation of this general procedure to favor convergence toward global maxima: the tempering-SAEM.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France