Methods and algorithms to learn spatio-temporal changes from longitudinal manifold-valued observations

Jean-Baptiste Schiratti

NNT : 2017SACLX009

Thèse de doctorat de l'Université Paris-Saclay préparée Ecole Polytechnique

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

par

Jean-Baptiste SCHIRATTI

Models and algorithms to learn spatiotemporal changes from longitudinal manifold-valued observations

Thèse présentée et soutenue à l'Institut du Cerveau et de la Moelle épinière (ICM), le 23 janvier 2017.

Composition du Jury :

M.	Marc Lavielle	Directeur de recherche INRIA, Ecole Polytechnique	(Président du jury)
M.	Ian Dryden	Professeur University of Nottingham	(Rapporteur)
M.	Jean-Michel Marin	Professeur Université de Montpellier	(Rapporteur)
M.	Tom Fletcher	Professeur associé University of Utah	(Rapporteur)
M.	Xavier Pennec	Directeur de recherche INRIA	(Examinateur)
M.	Olivier Colliot	Directeur de recherche CNRS, ARAMIS Lab	(Membre invité)
Mme.	Stéphanie Allassonnière	Professeur chargée de cours Ecole Polytechnique	(Directrice de thèse)
M.	Stanley Durrleman	Ph.D., Chercheur INRIA INRIA, ARAMIS Lab	(Directeur de thèse)

Abstract

We propose a generic Bayesian mixed-effects model to estimate the temporal progression of a biological phenomenon from manifold-valued observations obtained at multiple time points for an individual or group of individuals. The progression is modeled by continuous trajectories in the space of measurements, which is assumed to be a Riemannian manifold. The group-average trajectory is defined by the fixed effects of the model. To define the individual trajectories, we introduced the notion of "parallel variations" of a curve on a Riemannian manifold. For each individual, the individual trajectory is constructed by considering a parallel variation of the average trajectory and reparametrizing this parallel in time. The subject-specific spatiotemporal transformations, namely parallel variation and time reparametrization, are defined by the individual random effects and allow to quantify the changes in direction and pace at which the trajectories are followed. The framework of Riemannian geometry allows the model to be used with any kind of measurements with smooth constraints. Particular cases of the model are derived for the analysis of longitudinal normalized scalar measurements, symmetric positive definite matrices or to study the temporal progression of a family of biological features.

A stochastic version of the Expectation-Maximization algorithm, namely the Monte Carlo Markov Chains Stochastic Approximation EM algorithm (MCMC-SAEM), is used to produce produce *maximum a posteriori* estimates of the parameters. The use of the MCMC-SAEM together with a numerical scheme for the approximation of parallel transport is discussed. In addition to this, the method is validated on synthetic data and in high-dimensional settings.

Experimental results illustrate the ability of the model to be used with measurements of varying nature and complexity. They also illustrate the role of the fixed and random effects of the model in the estimation of normative scenarios of progression with its temporal variability of this progression among the population. These results consist in the analysis of neuropsychological test scores from patients with mild cognitive impairments later diagnosed with Alzheimer's disease, percentages of body fat from adolescent girls and simulated evolutions of symmetric positive definite matrices. The data-driven model of the impairment of cognitive functions during the course of the disease provide unique insights into the ordering and timing of the decline of these functions. Results on symmetric positive definite matrices show that the model correctly estimates a significant event in the observed evolution. The analysis of body fat show some limitations of our approach with univariate measurements while giving a satisfying estimation of the impact of adolescent ageing on the evolution of body fat. In these situations, the spatiotemporal transformations allow to put into correspondence the progression of individuals.

Keywords : Riemannian geometry, Longitudinal data, Statistical modeling, Stochastic EM algorithm.

Contents

Première partie

Introduction en langue Française

Sommaire

I.1 Motivation

Étudier l'évolution d'un phénomène biologique au cours du temps est un sujet dont l'intérêt est central dans de nombreux domaines scientifiques. Par exemple, comprendre l'évolution de certaines maladies joue un rôle clef dans le développement de nouveaux traitements. En vision par ordinateur, il peut être question de développer une méthode permettant d'annoter automatiquement des images de visages humains avec une certaine émotion.

Pour un individu ou objet donné, l'évolution du phénomène peut être mesurée à partir de caractéristiques, grandeurs d'intérêt (*features*) qui décrivent l'état de l'individu à un instant donné. Lorsque l'on s'intéresse à l'évolution d'une maladie, ces caractéristiques peuvent être des mesures extraites de bilans sanguins, telles que le nombre de lymphocytes, de globules rouges, la taille, le poids mais également de l'imagerie médicale telle que l'imagerie par résonance magnétique (IRM). En revanche, si l'on considère des images de visages humains, ces caractéristiques peuvent être la position de parties spécifiques du visage, telles que la bouche, le nez ou les joues. Ces mesures se traduisent, à chaque instant, par un nombre réel ou un vecteur de nombres réels. L'ensemble des mesures décrit une partie d'un espace Euclidien où l'évolution d'un individu peut être représentée par une trajectoire continue. Par exemple, des études sur la croissance et le développement de jeunes enfants ont permis d'obtenir des scénarios normatifs de poids et de taille, qui sont régulièrement utilisés par les pédiatres. Ces scénarios normatifs de croissance donnent des trajectoires d'évolution de la taille et du poids en fonction du temps, au cours des premières années de la vie. En particulier, ils donnent une trajectoire de progression moyenne, qui décrit l'évolution du poids et de la taille chez « l'enfant moyen ». De plus, ces scénarios donnent également la variabilité de cette trajectoire moyenne au sein de la population. Cette variabilité est souvent représentée sous forme d'un intervalle de confiance autour de la trajectoire moyenne. Une autre source de variabilité inter-individuelle dans les observations provient des differences de vitesse de progression au sein de la population. En effet, chaque individu progresse avec une allure qui lui est propre, certains évoluant plus rapidement que d'autres. Dans ces scénarios normatifs de taille ou de poids, la variabilité de la vitesse d'évolution n'est pas représentée. En effet, seule la variabilité des mesures à un âge donné est représentée. Pour ce qui est de l'analyse d'images de visages humains pour la détection automatique d'émotions, la variabilité inter-individuelle est d'autant plus importante que la forme du visage, des yeux, de la bouche varie fortement d'un individu à l'autre. Par ailleurs, certains individus vieillissent plus vite que d'autres. Enfin, on peut également remarquer que la dynamique des changements du visage n'est pas la même pour la joie que pour la colère.

Pour pouvoir estimer une trajectoire moyenne de progression ainsi que la variabilité de cette trajectoire au sein de la population, il convient de considérer des *données longitudinales*. Ces données consistent en des observations (du même phénomène biologique) acquises à des instants répétés, pour un groupe d'individus. Les instants auxquels ces

observations sont obtenues ainsi que leur nombre peuvent varier d'un individu à l'autre. De nombreuses études, dont certaines visant à modéliser la progression de maladies neurodégénératives ou l'effet du vieillissement sur le visage humain, ont constitué de larges bases de données longitudinales. La base de données Alzheimer's Disease Neuroimaging Initiative (ADNI), pour la maladie d'Alzheimer, ou la base MORPH, pour les visages humains, en sont des exemples. D'autres exemples incluent la base de donnée Baltimore Longitudinal Study of Ageing, qui permet d'étudier l'effet du vieillissement sur une population d'individus sains. Ces bases de données sont souvent *multimodales*. Les caractéristiques mesurées sont de nature différente. Dans la plupart des études, ces mesures sont représentées par des nombres réels ou des vecteurs de nombres réels. Pour certaines études, l'imagerie médicale - telle que l'Imagerie par Résonance Magnétique (IRM) - joue un rôle important. L'image peut être considérée en tant que telle mais permet également d'extraire des mesures complexes telles que des formes encodées par des maillages. Ces exemples montrent que les observations collectées dans ces bases de données peuvent être hautement *structurées*, comme des images ou des maillages. Dans ce cas, l'espace des mesures est souvent défini par des contraintes lisses et ne peut pas être considéré comme ayant une structure d'espace Euclidien. Les *variétés Riemanniennes* permettent une description mathématique rigoureuse de l'espace des mesures. Le cadre méthodologique offert par les variétés Riemanniennes permet de considérer des observations définies par des contraintes lisses, ainsi que des observations structurées ou non-structurées.

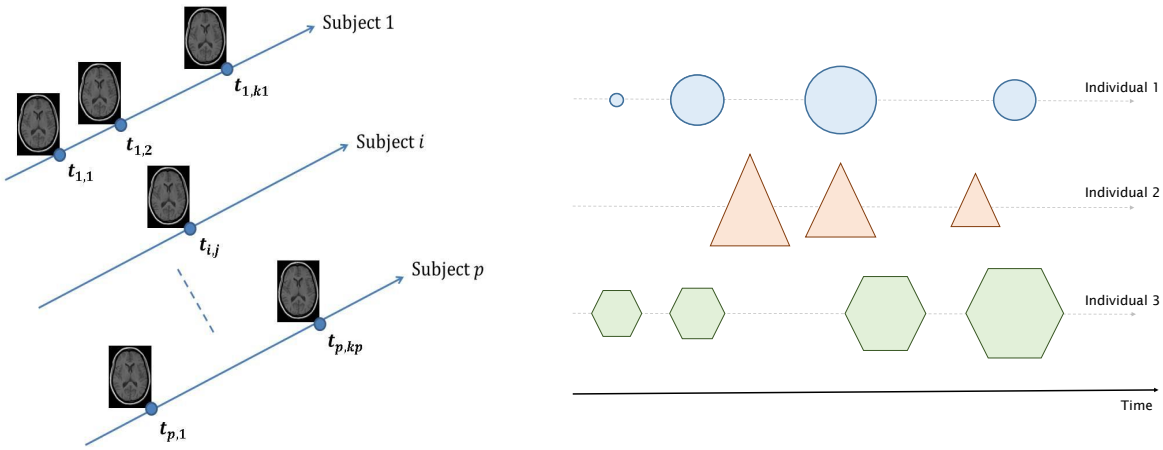


FIGURE 1 – Deux exemples schématiques de données longitudinales.

Ce manuscrit vise à proposer un modèle statistique, pour des données longitudinales issues de l'observation d'un phénomène biologique, satisfaisant aux exigences suivantes :

(i) le modèle est défini dans le cadre méthodologique des variétés Riemanniennes. Cela assure que le modèle peut être utilisé avec des observations définies par des contraintes lisses, comme des observations sans contraintes.

(ii) le modèle permet d'estimer une distribution de trajectoires dans l'espace des mesures. En particulier, une trajectoire moyenne est estimée ainsi que la variabilité de celle-ci au sein de la population. Cela permet de capturer la variabilité inter-individuelle des données longitudinales. De plus, le modèle estime la variabilité de la vitesse de progression ainsi que de l'avance ou du retard de l'évolution du phénomène observé chez les différents individus.

L'analyse statistique des observations collectées dans ces bases de données permet d'apprendre des modèles de l'évolution d'un phénomène biologique. Les *modèles à effets mixtes* [Eisenhart, 1947, Laird and Ware, 1982, Verbeke and Molenberghs, 2009] sont particulièrement populaires pour l'analyse de données longitudinales. Ces modèles statistiques incluent des *effets fixes* et *effets aléatoires* qui leur confèrent une structure hiérarchique. En effet, ces effets permettent de décrire le modèle tant au niveau du groupe qu'au niveau individuel. En adaptant un modèle à effets mixtes aux données, il est alors possible d'apprendre un modèle moyen d'évolution ainsi que des modèles spécifiques à chaque sujet. De plus, les modèles à effets mixtes imposent des conditions sur la loi de probabilité des effets aléatoires. Ainsi, ces effets aléatoires offrent la possibilité d'apprendre une distribution de trajectoires dans l'espace des observations. Les modèles à effets mixtes sont des modèles *génératifs* dont les paramètres peuvent être facilement interprétés. Par ailleurs, ces modèles peuvent gérer des données manquantes.

Les modèles linéaires à effets mixtes (modèles LME) sont les modèles à effets mixtes les plus simples et sont fréquemment utilisés pour l'analyse de données longitudinales. Ces modèles remontent au modèle d'ANOVA à effets mixtes [Scheffé, 1956]. Cependant, ils sont vraiment devenus populaires dans les années 1980 avec le papier fondateur de Laird et Ware [Laird and Ware, 1982]. En partant d'idées introduites dans [Harville, 1977], Laird et Ware ont mis en avant l'intérêt des modèles linéaires à effets mixtes - en particulier dans le domaine des sciences du vivant - et ont proposé une famille de modèles linéaires à effets mixtes flexible qui permet de traiter des observations manquantes. Soit p le nombre d'individus et pour $i \in \{1, \dots, p\}$, soit $\mathbf{y}_i \in \mathbb{R}^{k_i}$ le vecteur des observations du i-ème individu. Le modèle introduit par Laird et Ware suppose que les observations $(\mathbf{y}_i)_{1 \leq i \leq p}$ dérivent du modèle suivant :

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\alpha} + \mathbf{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \qquad \text{[i.1]}$$

Pour chaque individu, les observations $\mathbf{y}_i$ sont modélisées par une fonction linéaire des effets fixes $\boldsymbol{\alpha} \in \mathbb{R}^p$ et des effets aléatoires sujet-spécifiques $\boldsymbol{\beta}_i \in \mathbb{R}^q$. Les matrices $\mathbf{X}_i$ (respectivement $\mathbf{Z}_i$) (appelées *design matrices*) relient les effets fixes (respectivement aléatoires) aux observations. Le modèle LME générique donné en Eq. [i.1] suppose que les effets aléatoires $(\boldsymbol{\beta}_i)_{1 \leq i \leq p}$ sont inpendants et identiquement distribués, de loi normale.

Un cas particulier des modèles LME pour l'analyse de données longitudinales est le modèle avec pente et ordonnée à l'origine aléatoires (*random slope and intercept model*). Ce modèle est fréquemment utilisé pour analyser des données longitudinales

scalaires et s'écrit :

$$\mathbf{y}_{i,j} = (t_{i,j} - t_0)(\overline{\mathbf{A}} + \mathbf{A}_i) + (\overline{\mathbf{B}} + \mathbf{B}_i) + \boldsymbol{\varepsilon}_{i,j} \quad [i.2]$$

où $(t_{i,j})_{1\leq j\leq k_i}$ désignent les instants auxquels les observations du i-ème individu ont été obtenues. Les paramètres de population (ou effets fixes) du modèle sont la pente $\overline{\mathbf{A}}$ et l'ordonnée à l'origine $\overline{\mathbf{B}}$. Les effets aléatoires sujet-spécifiques sont les pentes $(\mathbf{A}_i)_{1\leq i\leq p}$ et les ordonnées à l'origine $(\mathbf{B}_i)_{1\leq i\leq p}$. Ces effets aléatoires sont supposés indépendants entre eux et identiquement distribués, de loi normale. Ce modèle avec pente et ordonnée à l'origine aléatoires permet d'estimer une trajectoire moyenne $\overline{\mathbf{D}}(t) = (t - t_0)\overline{\mathbf{A}} + \overline{\mathbf{B}}$. Les effets aléatoires permetttent d'estimer des trajectoires individuelles $\mathbf{D}_i(t) = (t - t_0)(\overline{\mathbf{A}} + \mathbf{A}_i) + (\overline{\mathbf{B}} + \mathbf{B}_i)$, qui sont obtenues en ajustant la pente et l'ordonnée à l'origine de la trajectoire moyenne. Ce modèle permet essentiellement de régresser les observations par rapport au temps. Le paramètre t_0 du modèle peut être interprété comme un *temps de référence*. Si les données longitudinales proviennent, par exemple, d'études sur le développement et l'élevage de certains animaux, d'études pharmacologiques, le temps de référence t_0 peut être choisi comme la date de naissance d'une portée ou le moment où un médicament a été administré. En revanche, il existe de nombreuses situations pour lesquelles il n'existe pas de temps de référence t_0 auquel les observations peuvent être comparées entre elles. Par exemple, dans les études sur les maladies neurodégénératives, deux individus du même âge peuvent être à des stades très différents de la progression de la maladie. Ainsi, la régression des observations par rapport au temps n'a pas de sens pour ces données. Pour des séquences d'images, il faudrait commencer par trouver l'instant qui correspond au même « événement » ou « état d'émotion » parmi les images de chaque individu. Il s'agit d'une tâche fastidieuse et on souhaiterait qu'un tel alignement entre séquences d'images soit le résultat d'un algorithme et non un prérequis pour analyser ces données. Lorsque le choix d'un temps de référence t_0 n'est pas évident, une solution consisterait à estimer ce paramètre à partir des données avec les autres paramètres du modèle. Cependant, en faisant cela, le modèle avec pente et ordonnée à l'origine aléatoires devient non-identifiable. C'est-à-dire qu'il existe une infinité de triplets $(\overline{A}, \overline{B}, t_0)$ qui maximise la vraisemblance du modèle. Par conséquent, le modèle avec pente et ordonnée à l'origine aléatoires n'est pas adapté pour décrire l'évolution d'un phénomène dont le début et la vitesse de progression varient d'un individu à l'autre.

Dans de nombreuses situations, supposer que les observations dépendent linéairement des effets fixes (ou aléatoires) du modèle pourraient être irréaliste. La famille des *modèles non-linéaires à effets mixtes* (modèles NLME) offre une plus grande flexibilité pour décrire les observations. Ces modèles sont introduits dans les travaux de Sheiner et Beal [Sheiner and Beal, 1980] puis dans [Lindstrom and Bates, 1988]. Ils ont fait l'objet de recherches actives depuis les années 1990. Ils sont maintenant très populaires dans de nombreux domaines tels que la modélisation pharmaco-cinétique, la médecine, etc. Les modèles NLME supposent qu'un jeu de données longitudinales $(\mathbf{y}_{i,j}, t_{i,j})_{1\leq i\leq p,\, 1\leq j\leq k_i}$ avec $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,k_i})$ découlent de :

$$\mathbf{y}_i = f(\boldsymbol{\psi}_i, t_i) + \boldsymbol{\varepsilon}_i \quad [i.3]$$

où f désigne une fonction non-linéaire et $\boldsymbol{\psi}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\boldsymbol{\beta}_i$. Les matrices $(\mathbf{X}_i)_{1\leq i\leq p}$ (respectivement $(\mathbf{Z}_i)_{1\leq i\leq p}$) relient les effets fixes $\boldsymbol{\alpha}$ (respectivement $\boldsymbol{\beta}_i$) à $\boldsymbol{\psi}_i$. Les effets aléatoires $(\boldsymbol{\beta}_i)_{1\leq i\leq p}$ sont supposés indépendants entre eux et identiquement distribués avec : $\boldsymbol{\beta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{D})$. On peut remarquer que les modèles LME sont un cas particulier des modèles NLME. Malgré leur formulation plus générique, les modèles NLME - en général - ne permettent pas non plus de prendre en compte la variabilité de l'âge au début de la maladie et de sa vitesse de progression. Dans [Yang et al., 2011] et [Delor et al., 2013], les auteurs ont cherché à apporter une solution à ce problème en introduisant des *décalages temporels* (*time shifts*). Cependant, les décalages temporels (ainsi que leur variabilité au sein de la population) ne sont pas estimés par l'intermédiaire d'un modèle statistique. Dans [Durrleman et al., 2013], des reparamétrisations temporelles appellées *time warps* (*id est*, des difféomorphismes de la droite réelle) sont considérés pour répondre à ce problème dans le cadre de l'analyse longitudinale de formes. Toutefois, l'estimation des paramètres du modèle statistique est réalisée en minimisant une somme de carrés qui résulte d'une approximation non contrôlée de la vraisemblance. Dans [Hong et al., 2014], les auteurs considèrent des *time warps* paramétriques avec un modèle de régression géodésique pour l'analyse de formes. Cependant, le modèle proposé ne se généralise pas facilement à l'analyse de données longitudinales. Enfin, dans [Lorenzi et al., 2015], les auteurs utilisent des techniques de géométrie Riemannienne pour estimer un modèle de vieillissement normal du cerveau à partir d'images IRM d'individus sains. Le modèle est ensuite utilisé pour calculer un décalage temporel appelé *morphological age shift*, qui correspond à « l'âge anatomique » du sujet, par rapport à un âge moyen estimé pour la population d'individus sains. Toutefois, ces décalages temporels ne sont pas estimés à partir d'un modèle statistique.

Comme mentionné plus haut, une difficulté tient en ce que les observations peuvent être hautement structurées, comme des images ou des maillages, et peuvent être définies par des contraintes lisses. Il s'en suit que l'espace des mesures peut être justement décrit comme une *variété Riemannienne*. Il convient de penser à une variété Riemannienne comme à un espace pouvant être courbe et de grande dimension. De manière similaire aux espaces Euclidiens, il est possible de faire du calcul différentiel sur une variété Riemannienne (définir les notions de fonction lisse, courbe, champ de vecteurs et les « dérivées » de ces quantités), faire des statistiques (définir une moyenne, médiane, variance, etc. d'un ensemble de points, lois de probabilité). Cependant, les calculs dans ces espaces peuvent se révéler complexes, certaines quantités n'ayant pas d'expression explicite, en termes de fonctions mathématiques usuelles. Bien que les variétés Riemanniennes offrent un cadre mathématique flexible et rigoureux pour décrire l'espace des mesures, ils soulèvent également des questions méthodologiques. En effet, les modèles LME ne sont pas définis pour des observations sur une variété Riemannienne. Le modèle de Laird et Ware n'est défini que pour des observations à valeurs dans un espace Euclidien. Des généralisations des modèles à effets mixtes ont toutefois été proposées dans la littérature. Dans [Fletcher, 2011], les auteurs proposent un modèle de régression linéaire sur une variété Riemannienne. Ce modèle, qui apparaît comme une

généralisation des modèles LME aux variété Riemanniennes, s'écrit :

$$\mathbf{y}_i = \mathrm{Exp}\big(\mathrm{Exp}(\mathbf{p}, \mathbf{X}\mathbf{v}), \boldsymbol{\varepsilon}\big) \qquad \text{[i.4]}$$

où $\mathrm{Exp}(\mathbf{p}, \mathbf{v})$ désigne l'*exponentielle Riemannienne* au point $\mathbf{p}$ sur la variété Riemannienne, avec vitesse initiale $\mathbf{v}$. Le modèle de bruit *intrinsèque* considéré ici conduit à une vraisemblance n'ayant pas d'expression explicite. Ainsi, les auteurs proposent d'estimer les paramètres de leur modèle statistique en minimiant un critère des moindres carrés et proposent une expression explicite du gradient de ce critère. Dans [Muralidharan and Fletcher, 2012], le modèle proposé par [Fletcher, 2011] est utilisé pour analyser des données longitudinales sur une variété Riemannienne. Toutefois, aucune loi de probabilité n'est définie sur les effets aléatoires du modèle. Un modèle hiérarchique sur le groupe des difféomorphismes (partageant des propriétés communes avec les variétés Riemanniennes) est proposé dans [Singh et al., 2013, Singh et al., 2014]. Ce modèle hiérarchique permet d'estimer une trajectoire moyenne, que les auteurs supposent être une géodésique. Les trajectoires sujet-spécifiques sont obtenues à partir de la trajectoire moyenne par une portion de géodésique émmanant de cette dernière, à l'instant correspondant à la première observation de chaque sujet. Les paramètres de cette portion de géodésique, qui sont considérés comme étant les effets aléatoires du modèle, dépendent essentiellement de l'instant de la première observation. Changer drastiquement le temps auquel les observations ont été acquises modifie ces effets sujets-spécifiques. Il devient alors difficile de définir une loi de probabilité commune pour ces paramètres, dans un modèle statistique bien posé, tout en s'assurant que le modèle soit robuste à ces changements d'origine temporelle. Il est, par ailleurs, difficile d'inclure la notion de reparamétrisation temporelle dans le modèle.

Ce manuscrit propose un modèle statistique à effets mixtes appelé *modèle générique spatio-temporel*. Le modèle est présenté dans un cadre Bayésien et est défini pour des observations longitudinales sur une variété Riemannienne. Les effets fixes du modèle sont utilisés pour définir une trajectoire moyenne tandis que les effets aléatoires sont utilisés pour définir des trajectoires sujet-spécifiques. Pour pouvoir définir de telles trajectoires individuelles, nous introduisons la notion de « variation parallèle » (*parallel variation*) d'une courbe sur une variété Riemannienne, basée sur la notion de *variation d'une courbe* [Do Carmo Valero, 1992]. Contrairement au modèle présenté dans [Singh et al., 2013], la loi de probabilité des effets aléatoires a la même forme (à une transformation isométrique près) en chaque point le long de la trajectoire moyenne. Cette propriété d'invariance temporelle permet d'inclure dans le modèle des reparamétrisations temporelles, définies à partir des effets aléatoires. Ces reparamétrisations temporelles sujet-spécifiques sont utilisées pour modifier l'allure à laquelle une variation parallèle de la trajectoire moyenne est parcourue. Ainsi, ces reparamétrisations temporelles permettent d'estimer les changements de rythme de progression au sein de la population. Par conséquent, le modèle Bayésien à effets mixtes présenté dans ce manuscrit inclut des transformations spatiales et temporelles grâce auxquelles il est possible de mettre les individus en correspondance et de définir une distribution spatio-temporelle de trajectoires sur une variété Riemannienne.

Le modèle générique spatio-temporel offre un moyen systématique d'obtenir des modèles non-linéaires à effets mixtes adaptés à une large variété d'observations et de variétés Riemanniennes. Nous donnerons la forme du modèle pour des observations scalaires normalisées, vues alors comme des points dans l'intervalle $]0,1[$, ou bien non-bornées. Nous donnerons également la forme du modèle pour la variété des matrices symétriques définies positives, ainsi qu'un produit de ces variétés élémentaires. Ce modèle, intrinsèquement non-linéaire, soulève la question du choix d'un algorithme pour estimer ses paramètres à partir d'observations longitudinales.

I.2 Méthodes et algorithmes d'inférence statistique pour les modèles non-linéaires à effets mixtes

Le modèle générique spatio-temporel est un modèle *non-linéaire* à effets mixtes. Plusieurs méthodes et algorithmes ont été proposés dans la littérature pour l'inférence statistique dans les modèles NLME. Ces algorithmes peuvent être regroupés en deux catégories : *algorithmes déterministes* et *algorithmes stochastiques*. Par opposition aux algorithmes stochastiques, les algorithmes déterministes ne nécessitent pas de savoir générer des variables aléatoires.

I.2.1 Algorithmes déterministes

Pour les modèles NLME, la vraisemblance observée $q(\mathbf{y} \mid \boldsymbol{\theta})$ - qui s'exprime comme une intégrale par rapport aux effets aléatoires (ou *variables latentes*) du modèle - est souvent impossible à calculer explicitement :

$$q(\mathbf{y} \mid \boldsymbol{\theta}) = \int_{\boldsymbol{\beta}} q(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) q(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \, d\boldsymbol{\beta} \qquad \text{[i.5]}$$

Les algorithmes déterministes visent à produire un maximum de vraisemblance (ou *maximum a posteriori*) en approximant la vraisemblance obsevée. Depuis les années 1990, de nombreuses contributions méthodologiques ont été faites pour développer ces algorithmes déterministes. Dans [Lindstrom and Bates, 1990], les auteurs ont proposé un algorithme itératif en deux étapes appelé « Linear Mixed-Effects approximation algorithm » (algorithme LME), qui consiste à linéariser le modèle NLME. La première étape de l'algorithme LME est appelée *Penalized Nonlinear Least Squares* (PNLS) et consiste à minimiser une somme de carrés non-linéaire. Cette étape revient à maximiser la loi conditionnelle jointe $q(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{y}, \widetilde{\boldsymbol{\theta}}) \propto q(y \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \widetilde{\boldsymbol{\theta}}) q(\boldsymbol{\beta} \mid \widetilde{\boldsymbol{\theta}})$, où $\boldsymbol{\alpha}$ (respectivement $\boldsymbol{\beta}$) désigne les effets fixes (respectivement aléatoires) du modèle NLME. Dans cette étape, le vecteur $\widetilde{\boldsymbol{\theta}}$ de paramètres de variance-covariance est considéré fixé. Les effets fixes (et aléatoires) $\widehat{\boldsymbol{\alpha}}$ (et $(\widehat{\boldsymbol{\beta}}_i)_{1 \leq i \leq p}$) qui maximisent $q(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{y}, \widetilde{\boldsymbol{\theta}})$ sont appelés *modes*

(de la loi conditionnelle). Ces modes sont utilisés dans la deuxième étape de l'algorithme. Dans cette deuxième étape, un développement de Taylor au premier ordre est utilisé pour linéariser le modèle autour de $\left(\widehat{\boldsymbol{\alpha}}, (\widehat{\boldsymbol{\beta}}_i)_{1\leq i\leq p}\right)$. Ainsi, un modèle LME est obtenu, puis ses paramètres sont estimés et utilisés pour mettre à jour les paramètres de variance-covariance $\widetilde{\boldsymbol{\theta}}$. Toutefois, cet algorithme est basé sur une approximation de la vraisemblance observée pour laquelle il n'a pas de contrôle, ni de garanties théoriques de convergence. De plus, pour pouvoir être utilisé dans un cadre Bayésien, cet algorithme devrait être modifié.

Dans [Davidian and Gallant, 1992], Davidan et al. utilisent une méthode de quadrature Gaussienne adaptative pour approximer la vraisemblance donnée en Eq. [i.5]. Un cas particulier de cette méthode, appelé *Approximation Laplacienne*, est obtenu en considérant cette méthode de quadrature Gaussienne adaptative avec un seul point. Ces deux méthodes, la quadrature Gaussienne adaptative et l'approximation Laplacienne, sont discutées dans [Pinheiro, 1994] et dans le livre [Pinheiro and Bates, 2006]. L'approximation Laplacienne, introduite dans [Tierney and Kadane, 1986], est utilisée pour approximer la vraisemblance observée individuelle $q(\mathbf{y}_i \mid \boldsymbol{\theta})$. De manière similaire à l'algorithme LME, cette approximation repose sur un développement de Taylor au premier ordre. L'approximation est donnée par :

$$q(\mathbf{y}_i \mid \boldsymbol{\theta}) \simeq \frac{\det \boldsymbol{\Delta}}{(\sigma^2 2\pi)^{k_i/2}} \exp\Big(-\frac{1}{2\sigma^2} g_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i)) \Big) \Big(\det \frac{\partial^2 g_{\tilde{\boldsymbol{\theta}}}}{\partial \boldsymbol{\beta}_i^2}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i) \Big)^{-1/2} \qquad \text{[i.6]}$$

où $\boldsymbol{\Delta}$ est la *matrice de précision* telle que $\mathbf{D}^{-1} = \sigma^{-2}\boldsymbol{\Delta}^\top\boldsymbol{\Delta}$ et $g_{\tilde{\boldsymbol{\theta}}}$ est définie par : $g_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_i) = \|\mathbf{y}_i - f(\psi_i, t_i)\|^2 + \|\boldsymbol{\Delta}\boldsymbol{\beta}_i\|^2$ avec $\widehat{\boldsymbol{\beta}}_i(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\theta}}) = \underset{\boldsymbol{\beta}_i}{\operatorname{argmin}}\, g(\boldsymbol{\alpha}, \boldsymbol{\beta}_i, \widetilde{\boldsymbol{\theta}})$. Un défaut de cette approximation est qu'elle requiert le calcul de l'inverse de la matrice Hessienne de $g_{\tilde{\boldsymbol{\theta}}}$ au point $(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i)$. En pratique, pour des modèles complexes, cette matrice Hessienne est trop coûteuse à calculer et son approximation en utilisant des schémas numériques s'avère également coûteuse. Pour palier à cela, Pinheiro a proposé de remplacer la matrice Hessienne dans Eq. [i.6] par une approximation. Avec cette approximation, Eq. [i.6] a une forme plus facile à manipuler et peut être maximisée en utilisant des outils tels que la descente de gradient. Cependant, l'approximation Laplacienne reste très coûteuse en temps de calcul et manque de résultats théoriques concernant sa convergence.

L'algorithme Espérance-Maximisation (EM) [Dempster et al., 1977] est un algorithme populaire permettant d'obtenir un *maximum de vraisemblance* (ou *maximum a posteriori*) pour les modèles LME et NLME. Pour le cas des modèles LME, l'algorithme EM est décrit dans [Laird and Ware, 1982,Laird et al., 1987]. L'algorithme EM a été introduit dans le contexte des modèles statistiques à variables latentes. L'idée clef de cet algorithme est de maximiser, de manière itérative, une borne inférieure sur la vraisemblance observée. Sous des conditions assez génériques, données dans [Dempster et al., 1977] puis corrigées dans [Wu, 1983] et généralisées dans [Delyon et al., 1999], l'algorithme converge vers un maximum (local) de la vraisemblance observée. L'algorithme EM itère, jusqu'à convergence, entre deux étapes : l'« étape E » et l'« étape

M ». Soit $\mathbf{y}$ (respectivement $\mathbf{z}$) les observations (respectivement les variables latentes) du modèle générique spatio-temporel. Soit $k \in \mathbb{N}^*$ et $\boldsymbol{\theta}^{(k)}$ la valeur courrante des paramètres du modèle à la k-ème itération de l'algorithme. Soit $q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$ la loi jointe des observations et des variables latentes conditionnellement aux paramètres $\boldsymbol{\theta}$. L'étape E de l'algorithme consiste à calculer la fonction $\boldsymbol{\theta} \in \Theta \mapsto \mathbf{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ définie par :

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \mathbb{E}_{q(\cdot|\mathbf{y},\boldsymbol{\theta}^{(k)})}\big[\log q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})\big]. \tag{i.7}$$

Dans Eq. [i.7], l'espérance est calculée par rapport à la loi conditionnelle des variables latentes sachant les observations $\mathbf{y}$ et l'état courrant des paramètres du modèle $\boldsymbol{\theta}^{(k)}$. Cette fonction $\mathbf{Q}(\cdot \mid \boldsymbol{\theta}^{(k)})$ est ensuite utilisée dans l'étape M de l'algorithme pour mettre à jour les paramètres comme suit :

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \big(\mathbf{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) + q_{\text{prior}}(\boldsymbol{\theta})\big). \tag{i.8}$$

Cependant, pour de nombreux modèles non-linéaires à effets mixtes, l'espérance apparaissant dans l'étape E de l'algorithme ne peut être calculée explicitement. De plus, la loi conditionnelle $q(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$ des variables latentes $\mathbf{z}$ sachant les observations $\mathbf{y}$ et les paramètres $\boldsymbol{\theta}$ n'est, en général, pas d'une forme connue.

I.2.2 Algorithmes stochastiques

Comme mentioné ci-dessus, l'étape E de l'algorithme EM peut s'avérer impossible à calculer explicitement pour certains modèles NLME. Une façon de résoudre ce problème consiste à considérer une version stochastique de l'algorithme EM : l'algorithme *Monte Carlo Markov Chains - Stochastic Approximation EM* (MCMC-SAEM). Contrairement aux algorithmes déterministes, l'algorithme MCMC-SAEM peut être utilisé sans avoir à calculer de dérivées ou de gradient. En effet, il suffit de savoir évaluer la fonction f du modèle pour utiliser le MCMC-SAEM. Cet algorithme, dont la convergence est prouvée dans [Allassonnière et al., 2010] (basé sur les travaux de [Kuhn and Lavielle, 2004]), est un algorithme itératif en trois étapes : *échantillonnage*, *approximation stochastique* et *maximisation*. Soit $\boldsymbol{\theta}^{(k)}$ l'état courrant des paramères à la k-ème itération de l'algorithme. Dans l'étape d'échantillonnage du MCMC-SAEM, des variables latentes $\mathbf{z}^{(k)}$ sont simulées en utilisant le noyau de transition d'une chaîne de Markov ergodique dont la loi stationnaire est la loi conditionnelle $q(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(k)})$ des variables latentes $\mathbf{z}$ sachant les observations $\mathbf{y}$ et l'état courrant des paramètres $\boldsymbol{\theta}^{(k)}$. Dans l'étape d'approximation stochastique, une fonction $\mathbf{Q}_k(\cdot)$ est définie par :

$$\forall \boldsymbol{\theta} \in \Theta,\ \mathbf{Q}_k(\boldsymbol{\theta}) = \mathbf{Q}_{k-1}(\boldsymbol{\theta}) + \varepsilon_k\big(\log q(\mathbf{y}, \mathbf{z}^{(k)} \mid \boldsymbol{\theta}) - \mathbf{Q}_{k-1}(\boldsymbol{\theta})\big). \tag{i.9}$$

avec $\mathbf{Q}_0 = 0$. Le choix de la suite $(\varepsilon_k)_{k\geq 0}$ est discuté plus bas. L'équation Eq. [i.9] est une approximation stochastique du type Robbins-Monro [Robbins and Monro, 1951], qui converge vers l'espérance $\mathbb{E}_{q(\mathbf{z}|;\mathbf{y},\boldsymbol{\theta}^{(k-1)})}\big[\log q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})\big]$. Ainsi, cette approximation

stochastique est asymptotiquement équivalente à l'étape E de l'algorithme EM. Finalement, l'étape de maximisation consiste à mettre à jour l'état courrant des paramètres en maxmimisant la fonction $\mathbf{Q}_k$ par rapport à $\boldsymbol{\theta} \in \Theta$, où Θ désigne l'espace des paramètres du modèle. On peut remarquer que le MCMC-SAEM permet d'obtenir un mode, c'est-à-dire un maximum local, de la loi a posteriori $q(\boldsymbol{\theta} \mid \mathbf{y})$. On n'apprend pas toute la loi a posteriori $q(\boldsymbol{\theta} \mid \mathbf{y})$.

D'autres algorithmes stochastiques incluent les méthodes « pleinement Bayésiennes ». Il s'agit d'un terme faisant référence à une classe de méthodes Monte Carlo par chaînes de Markov (MCMC) visant à apprendre la loi a posteriori $q(\boldsymbol{\theta} \mid \mathbf{y})$. Pour cela, ces algorithmes construisent une chaîne de Markov ergodique dont la loi stationnaire est la loi $q(\boldsymbol{\theta} \mid \mathbf{y})$. Après un certain nombre d'itérations (ce qui correspond à une période de *burn-in*), les réalisations de cette chaîne de Markov sont approximativement distribuées selon la loi $q(\boldsymbol{\theta} \mid \mathbf{y})$. En considérant un nombre suffisamment large de réalisations de cette chaîne de Markov, il est possible de reconstruire la loi a posteriori en utilisant, par exemple, des méthodes d'estimation de densité à noyaux (*Kernel Density Estimates*). On peut alors obtenir des informations concernant cette loi de probabilité telles que ses modes. Les méthodes de Monte Carlo Hamiltonniennes (HMC) sont des méthodes MCMC qui sont populaires pour l'inférence Bayésienne. Dans [Hoffman and Gelman, 2014], les auteurs proposent une méthode HMC adaptative appelée *No U-Turn Sampler* (NUTS). Cet échantillonneur MCMC est implémenté sous forme d'une librairie R/C++ appelée `STAN`. Dans leur papier, les auteurs mentionnent que l'échantillonneur NUTS offre de meilleures performances en grande dimension qu'un échantillonneur classique de Gibbs ou de Metropolis-Hastings. Toutefois, l'échantillonneur NUTS nécessite d'intégrer un système d'équations Hamiltonniennes et de calculer (numériquement) le gradient de la loi a posteriori $q(\boldsymbol{\theta} \mid \mathbf{y})$ par rapport aux paramètres $\boldsymbol{\theta}$, ce qui peut s'avérer très coûteux.

Dans cette dissertation, les paramètres du modèle générique spatio-temporel seront estimés en utilisant l'algorithme MCMC-SAEM. Nous nous intéresserons également à la validation expérimentale de cet algorithme avec le modèle générique spatio-temporel. En particulier, le MCMC-SAEM est testé sur des observations longitudinales de matrices symétriques définies positives, ce qui est un exemple non-trivial de variété Riemamnnienne de courbure négative. De plus, les résultats numériques et les temps de calcul obtenus avec le MCMC-SAEM sont comparés à ceux obtenus avec d'autres algorithmes classiques pour l'inférence statistique dans les modèles NLME. Enfin, nous nous intéressons à l'utilisation du MCMC-SAEM dans le contexte des variété Riemanniennes et à sa compatibilité avec des schémas numériques pour calculer les transformations spatio-temporelles sujet-spécifiques du modèle. Les résultats expérimentaux de cette dissertation sont obtenus en analysant des jeux de données longitudinales liés à la santé. En particulier, le jeu de données constitué de scores à des tests neuropsychologiques ou de mesures d'épaisseurs corticales permettent d'obtenir des informations pertinentes quant à l'évolution de la maladie d'Alzheimer (AD) chez une population d'individus issue de la cohorte Alzheimer's Disease Neuroimaging Initiative (ADNI). Les résultats

obtenus avec ces données longitudinales montrent l'intérêt des transformations spatio-temporelles. En effet, nous montrons que le modèle générique spatio-temporel permet d'estimer un scénario normatif de progression de la maladie d'Alzheimer, ainsi que son effet sur différentes fonctions cognitives. Le modèle permet également d'estimer l'ordre dans lequel les fonctions cognitives déclinent et l'écart temporel relatif entre le déclin de deux fonctions cognitives. De plus, l'analyse de jeux de données longitudinales de matrices symétriques définies positives et de pourcentages de masse graisseuse montre que le modèle générique spatio-temporel permet d'estimer un modèle moyen de progression et que les transformations spatio-temporelles mettent correctement les individus en correspondance.

I.3 Présentation des chapitres

Le chapitre III présente quelques notions clef de géométrie Riemannienne et de théorie des méthodes de Monte Carlo par chaînes de Markov. Comme précisé ci-dessus, le modèle générique spatio-temporel, introduit plus tard dans cette dissertation, est défini pour des données longitudinales sur une variété Riemannienne et les trajectoires de progression sont des courbes sur une variété Riemannienne. Les notions présentées dans le chapitre III permettent de définir un cadre mathématique rigoureux et flexible dans lequel le modèle générique spatio-temporel sera défini. De plus, des notions sur les méthodes de Monte Carlo par chaînes de Markov sont nécessaires pour introduire la version stochastique de l'algorithme EM que nous utiliserons pour estimer les paramètres du modèle.

Le modèle générique spatio-temporel pour données longitudinales sur une variété Riemannienne est présenté dans le chapitre IV. Ce chapitre commence par définir la notion de « variation parallèle » d'une courbe. Cette définition requiert que des effets aléatoires du modèle, appelés *décalages spatiaux* (*space shifts*) satisfassent une condition d'orthogonalité. Des propriétés et exemples de variations parallèles sur une variété Riemannienne sont donnés au début du chapitre. Ces exemples montrent que cette notion généralise celle de parallélisme aux espaces non-Euclidiens. Enfin, le modèle générique spatio-temporel est présenté en section IV.3. Ce chapitre s'intéresse également à des schémas numériques permettant d'assurer que la condition d'orthogonalité est satisfaite.

Le chapitre V introduit plusieurs cas particuliers du modèle générique spatio-temporel. Comme mentionné ci-dessus, ce modèle permet, étant donné le choix d'une variété Riemannienne et d'une métrique Riemannienne, d'obtenir une large variété de modèles non-linéaires à effets mixtes. Ce chapitre vise à présenter quelques-uns de ces modèles, qui résultent de différent choix de variété Riemannienne. Un modèle appelé « modèle logisitique » (respectivement « modèle de droites ») est proposé pour l'analyse de données longitudinales scalaires normalisées (respectivement non-bornées). Dans les sections suivantes, deux modèles sont proposés pour des observations longitu-

dinales multivariées. Le premier permet d'analyser des jeux de données longitudinales de matrices symétriques définies positives, dont les matrices de covariance sont un cas particulier. Le second, appelé « modèle de propagation » permet de modéliser l'évolution conjointe d'une famille de caractéristiques biologiques. Le modèle permet aussi d'estimer l'écart temoprel relatif entre la progression de deux de ces caractéristiques.

Le chapitre VI concerne l'estimation des paramètres du modèle générique. La première section du chapitre commence avec une revue de la littérature concernant différentes méthodes pour l'inférence statistique dans les modèles non-linéaires à effets mixtes et motive le choix d'utiliser une version stochastique de l'algorithme EM pour estimer les paramètres du modèle. Les sections suivantes de ce chapitre décrivent comment cet algorithme stochastique peut être utilisé dans le contexte des variétés Riemanniennes. Enfin, la dernière section discute la validation expérimentale de l'algorithme et ses aspects computationnels.

Finalement, le chapitre VII présente des résultats expérimentaux obtenus avec différents cas particuliers du modèle générique. Les résutlats présentés dans ce chapitre ont été obtenus en analysant des jeux de données longitudinales en rapport avec la santé. Ces résultats montrent que les transformations spatio-temporelles du modèle permettent de mettre en correspondance des événements similaires le long des trajectoires de progression individuelles. Le modèle générique réussit également à estimer un événement, un changement qui se produit pour chaque individu, au cours de la période d'observation.

I.4 Liste des publications

Ce manuscrit a donné lieu aux publications suivantes.

I.4.1 Articles de conférences avec commité de lecture

- [Schiratti et al., 2015d] **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Alexandre Routier, the Alzheimer's Disease Neuroimaging Initiative (ADNI), Olivier Colliot and Stanley Durrleman. A mixed-effects model for longitudinal univariate manifold-valued data. In : Ourselin, S., Alexander, D. C., Westin, C.-F., Cardoso, M. J. (Eds.), *Information Processing in Medical Imaging. - IPMI 2015.* No. 9123 in Lecture Notes in Computer Science. Springer International Publishing. Poster élu parmi les trois meilleurs posters de la conférence, donnant lieu à une présentation orale.

- [Schiratti et al., 2015a] **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Olivier Colliot and Stanley Durrleman. Learning spatiotemporal trajectories from

manifold-valued longitudinal data. In : *Advances in Neural Information Processing Systems. - NIPS 2015.* pp. 2404-2412. Poster et travel award.

- [Schiratti et al., 2015b] **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Olivier Colliot and Stanley Durrleman. Mixed-effects model for the spatiotemporal analysis of longitudinal manifold-valued data. In : *5th MICCAI Workshop on Mathematical Foundations of Computational Anatomy - MFCA 2015.* Présentation orale

I.4.2 Présentations

- [Schiratti et al., 2015c] **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Alexandre Routier, Olivier Colliot and Stanley Durrleman. Estimating profiles of disease progression using mixed-effects models with time reparametrization. In : *Organization for Human Brain Mapping - OHBM 2015.* Poster

- **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Olivier Colliot and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued measurements. In : *Kickoff Workshop of the European Progression of Neurodegenerative Diseases initiative (EuroPOND) - 2016.* Présentation orale

I.4.3 Brevets

- **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Olivier Colliot and Stanley Durrleman. A method for determining the temporal progression of a biological phenomenon and associated methods and devices. Patent application PCT/IB2016/052699.

Part II

Introduction

Summary

II.1 Motivation

Numerous scientific fields require to study the temporal progression of a biological or natural phenomenon. For instance, the study of progressive diseases plays a crucial role in the development of new treatments. In computer vision, one may be interested in analyzing faces in order to automatically detect whether a face displays an emotion and label the images with the correct emotion.

For a given individual or object, the evolution of the phenomenon can be measured by several characteristics or features, which describe the state of the individual at a given time point. In studies on diseases, the features may be blood samples measurements, like lymphocytes or blood cells count, height, weight, and also medical imaging, as Magnetic Resonance (MR) imaging. For human faces, one could consider the position of specific points of the face such as the nose, mouth or cheeks. Each of these features can be represented, at a given time point, by a real number or vector of real numbers. The collection of these features lies in a subset of the Euclidean space where the evolution of an individual can be represented by a continuous trajectory. For example, developmental and growth studies have provided normative growth scenarios of height and weight, which are often used by pediatricians. These normative growth scenarios give trajectories of weight or height evolution with time, during the first years of life. In particular, they give an average trajectory of progression, which describes the evolution of weight or height in a typical child. These scenarios also provide information on the variability of this average trajectory among the population, which is usually represented by a confidence interval on the average trajectory. Another source of variability in the measurements comes from the differences in pace of progression among the population. Indeed, each individual is progressing at its own pace, with some individuals progressing faster than others. Regarding the analysis of images of faces for the detection of emotions, the inter-individual variability is quite important since the shape of the face, mouth, eyes varies a lot within the population. Also, some individuals may age faster than others. Not only each human has a different face but also, the dynamics of face changes during smiling or anger may vary across individuals. In normative growth scenarios for a single feature, such as height, the variability in pace of growth is not measured. Such scenarios usually measure only the variability of the measurements at a given age.

In order to estimate an average trajectory of progression and the variability of the average trajectory among the population, one usually analyze *longitudinal data*, which consists in observations of the same biological phenomenon at repeated time points, for a group of individuals. The time points and their number may be different for each individual. For several studies, such as modeling the progression of neurodegenerative diseases or assessing the effects of ageing on human faces, large longitudinal databases have been created, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) for Alzheimer's disease, or the MORPH database for human faces. Other examples of longitudinal databases include the Baltimore Longitudinal Study of Ageing, which aims at

studying the effects of ageing on a healthy population and the Beginning Postsecondary Students Longitudinal Study, which collects observations of school and work experience in students starting postsecondary education. Usually, these longitudinal databases are *multimodal*. They consist in repeated observation of features of different nature. In medical studies, the features of interest are usually represented as scalar measurements or vectors of scalar measurements. Still, in some studies, medical imaging, like MR imaging, plays an important role. The image may be considered as a feature of interest in itself. This type of observations may also allow to extract more complex features, such as shapes encoded as meshes. These examples show that the features collected in these databases can be highly *structured*, as are images or meshes. In those cases, the space of measurements is usually defined by smooth constraints and may not behave as a Euclidean space. Indeed, algebraic operations such as addition or scaling do not make sense for images or meshes. When the features are not defined by smooth constraints, the space of measurements is usually the Euclidean space. *Riemannian manifolds* are spaces which provide a rigorous mathematical framework to describe the space of measurements. This framework allows to consider features defined by smooth constraints as well as unconstrained features, and structured or unstructured features.

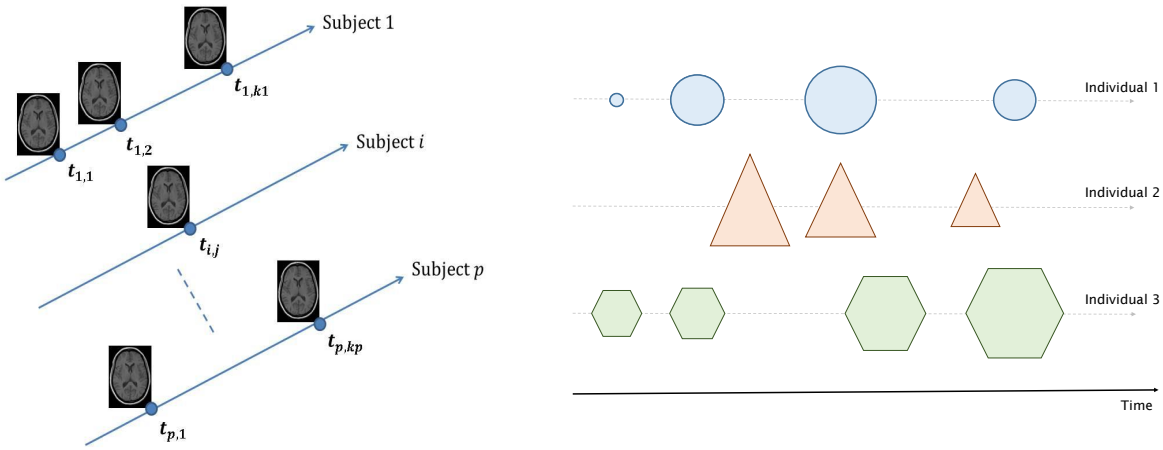


Figure 2 – Two schematic examples of longitudinal datasets.

This dissertation aims at proposing a statistical model, for longitudinal observations of a biological or natural phenomenon, which satisfies to the following requirements:

(i) the model is defined in the framework of Riemannian manifolds. This ensures that the model could be used with observations defined by smooth constraints, as well as unconstrained ones.

(ii) The model allows to estimate a distribution of trajectories in the space of measurements. In particular, a group-average trajectory is estimated as well as its

variability among the population. This allows to capture the inter-individual variability in the longitudinal observations. In addition to this, the model also allows to estimate variability in speed and delay of progression among the population.

The statistical analysis of measurements collected longitudinal databases may enable to learn data-driven models of evolution. In the literature, *mixed-effects models* [Eisenhart, 1947, Laird and Ware, 1982, Verbeke and Molenberghs, 2009] appear as a popular method for the analysis of longitudinal data. These statistical models are particularly popular for the analysis of longitudinal data since they include *fixed* and *random effects* which provide these models with a hierarchical structure. Indeed, these effects allow the model to be described at the population (or group) level, as well as the individual level. By fitting a mixed-effects model, one can learn an average model of evolution as well as individual-specific models. Therefore, the information provided by the observations of each individual is averaged and becomes more generalizable to other individuals. Moreover, mixed-effects models enforce conditions on the distribution of the random effects in the model. Thus, the random effects open up the possibility to learn a distribution of trajectories in the space of observations. Mixed-effects are *generative* statistical models whose parameters may be easily interpreted. In addition to this, these models offer the advantage of handling missing data.

Linear Mixed Effects (LME) models are the most simple mixed-effects models and frequently used in longitudinal studies. These models date back to the *mixed-effects ANOVA* [Scheffé, 1956]. However, they really became popular in the early 1980s with the seminal paper of Laid and Ware [Laird and Ware, 1982]. Building upon ideas from [Harville, 1977], Laird and Ware highlighted the usefulness of linear mixed-effects models, especially in the context of life sciences, and proposed a flexible family of linear mixed-effects models which could easily handle missing observations. Let p denote the number of individuals and for $i \in \{1, \dots, p\}$, let $\mathbf{y}_i \in \mathbb{R}^{k_i}$ be the vector of observations for the ith individual. The linear mixed-effect model introduced by Laird and Ware assumes that:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\alpha} + \mathbf{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \qquad \text{[ii.1]}$$

For each individual, the observation $\mathbf{y}_i$ are modeled as a linear function of the fixed effects $\boldsymbol{\alpha} \in \mathbb{R}^p$ and the individual-specific random effects $\boldsymbol{\beta}_i \in \mathbb{R}^q$. The matrices $\mathbf{X}_i$ (respectively $\mathbf{Z}_i$) *design matrices* linking the fixed (respectively random) effects to the observations. The generic LME model Eq. [ii.1] assumes that the random effects $\boldsymbol{\beta}_i$ are normally distributed.

A particular case of the LME models for analyzing longitudinal data is the *random slope and intercept model*. This LME model is usually used to analyze scalar longitudinal observations and writes:

$$\mathbf{y}_{i,j} = (t_{i,j} - t_0)(\overline{\mathbf{A}} + \mathbf{A}_i) + (\overline{\mathbf{B}} + \mathbf{B}_i) + \boldsymbol{\varepsilon}_{i,j} \qquad \text{[ii.2]}$$

where $t_0 \in \mathbb{R}$ and $(t_{i,j})_{1 \leq j \leq k_i}$ denotes the time points at which the observations of the ith individual were obtained. The population parameters (or fixed effects) of the model

are the slope $\overline{\mathbf{A}}$ and the intercept $\overline{\mathbf{B}}$. The random effects are the subject-specific slopes $(\mathbf{A}_i)_{1\leq i\leq p}$ and intercepts $(\mathbf{B}_i)_{1\leq i\leq p}$, which are assumed to be normally distributed and independent of each other. This random slope and intercept model estimates an average trajectory $\overline{\mathbf{D}}(t) = (t-t_0)\overline{\mathbf{A}} + \overline{\mathbf{B}}$. The random effects of the model allow to estimate also individual trajectories $\mathbf{D}_i(t) = (t-t_0)(\overline{\mathbf{A}}+\mathbf{A}_i) + (\overline{\mathbf{B}}+\mathbf{B}_i)$, which are obtained by adjusting the slope and intercept of the average trajectory. This model is essentially built on the idea of regressing the measurements against time. The parameter t_0 can be understood as a *reference time*. If the longitudinal dataset arises from animal breeding studies, developmental studies or pharmacological studies, the reference time t_0 can be chosen to be the date of birth or time at which a drug was administered. However, there are many situations in which there is no obvious reference time t_0 at which observations may be compared. In aging for instance, the different individuals may be at the same age at different stages of aging or disease progression. It therefore does not make sense to regress the measurements against age. In video sequences, one should first find the time-frame which corresponds to the same "event" or "stage of smiling" across the sequences. This task may be difficult and one would like such an alignment to be the output of the algorithm instead of a prerequisite. A way to address this problem would consist in estimating the reference time t_0 along with the other parameters of the model. However, this leads to a non-identifiable model because there are infinitely many triplets $(\overline{A}, \overline{B}, t_0)$ which maximize the likelihood of the model. As a consequence, the random slope and intercept model is inadequate for studies in which the observations describe the evolution of a phenomenon whose onset and pace of progression varies from an individual to another.

In various situations, assuming that the observations depend linearly on the fixed (or random) effects of the model might be unrealistic. The class of nonlinear mixed-effects (NLME) models offer a greater flexibility to describe the observations. These models first appeared in the work of Sheiner and Beal [Sheiner and Beal, 1980] and later in [Lindstrom and Bates, 1988]. They have been a blooming topic of research since 1990. These models are now popular tools in a large variety of areas, such as pharmacokinetic modeling, medicine, etc. NLME models assume that a longitudinal dataset $(\mathbf{y}_{i,j}, t_{i,j})_{1\leq i\leq p, 1\leq j\leq k_i}$, with $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,k_i}) \in \mathbb{R}^{k_i}$, arise from:

$$\mathbf{y}_i = f(\boldsymbol{\psi}_i, t_i) + \boldsymbol{\varepsilon}_i \qquad \text{[ii.3]}$$

where f is a nonlinear mapping and $\boldsymbol{\psi}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\boldsymbol{\beta}_i$. $(\mathbf{X}_i)_i$ and $(\mathbf{Z}_i)_i$ are design matrices linking the fixed (respectively random) effects $\boldsymbol{\alpha}$ (respectively $\boldsymbol{\beta}_i$) to $\boldsymbol{\psi}_i$. The random effects are assumed to be normally distributed with $\boldsymbol{\beta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{D})$ and independent of each other. One can easily note that the LME models appear as a particular case of NLME models. Despite their more generic formulation, NLME models also, in general, do not account for the variability in age at onset and pace of progression. In [Yang et al., 2011] and [Delor et al., 2013], the authors addressed this problem by introducing *time shifts* in their statistical models. However, the time shifts (and their distribution among the population) were not estimated in a statistical framework. In [Durrleman et al., 2013], time reparametrizations called *time warps* (smooth

monotonic transformations of the real line) are considered to address this point in the context of longitudinal shape analysis. Nevertheless, the estimation of the parameters of the statistical model is made by minimizing a sum of squares which results from an uncontrolled likelihood approximation. In [Hong et al., 2014], the authors use parametric time warps with a geodesic regression model for shape analysis. However, the proposed model is not easily extended to longitudinal observations. In [Lorenzi et al., 2015], the authors used Riemannian manifold techniques to estimate a model of the brain's normal ageing from healthy individuals MR images. The model was used to compute a time shift, called *morphological age shift*, which corresponds to the actual anatomical age of the subject with respect to an estimated average age for healthy subjects. However, the subject-specific time shifts were not estimated as parameters of a statistical model.

As mentioned above, the challenge is that the observations may be highly structured data, such as shapes or images, and defined by smooth constraints. As a matter of fact, the space in which observations lie can be better modeled as a *Riemannian manifold*. One should think of a Riemannian manifold as a space which might be curved and high-dimensional. Similarly to Euclidean spaces, one can do differential calculus on a Riemannian manifold (define smooth functions, curves, vector fields, define the "derivative" of such quantities, etc.) and do statistics (define the mean, median, variance of a set of points, probability distributions, etc.). However, the computations in such spaces may be complicated or even intractable in closed-form. Even though Riemannian manifolds offer a very flexible and rigorous framework to describe the space of observations with smooth constraints, they also raise methodological challenges since the LME models are not defined for observations on a Riemannian manifold. Indeed, the Laird and Ware mixed-effect models is defined for observations in the Euclidean space. Generalization of mixed-effects models to Riemannian manifolds have been proposed in the literature. In [Fletcher, 2011], the authors proposes a statistical model of linear regression on a Riemannian manifold. The proposed model, which appears as a generalization of LME models to Riemannian manifolds, writes:

$$\mathbf{y}_i = \mathrm{Exp}\big(\mathrm{Exp}(\mathbf{p}, \mathbf{X}\mathbf{v}), \boldsymbol{\varepsilon}\big) \qquad \text{[ii.4]}$$

where $\mathrm{Exp}(\mathbf{p}, \mathbf{v})$ denotes the *Riemannian exponential* at the point $\mathbf{p}$ on the Riemannian manifold and with initial velocity $\mathbf{v}$. The *intrinsic* noise model considered for this model leads to an intractable computation of the likelihood. Therefore, the authors propose to estimate the parameters of the model by minimizing a least-squares criterion and derive a closed-form expression the gradient of this criterion. In [Muralidharan and Fletcher, 2012], the model proposed in [Fletcher, 2011] is used for the analysis of longitudinal observations on a Riemannian manifold. However, no probability distribution is defined on the individual effects of the model. A hierarchical model on a group of diffeomorphisms (which shares common properties of Riemannian manifold) is proposed in [Singh et al., 2013, Singh et al., 2014]. This hierarchical model estimates an average trajectory, which is modeled as a geodesic on the Riemannian manifold. The subject-specific trajectories are derived from the average trajectory by a portion

of geodesic emanating from the average geodesic at the time point corresponding to the first observation of the subject. The parameters of this portion of geodesic, which are considered as the individual random effects of the model, essentially depend on this first time point, which comes from the design of the study. Changing the time of observations drastically change the value of the subject-specific parameters. It makes difficult therefore to define a common distribution of these parameters in a well-posed mixed-effect model, to make the model reasonably robust to slight changes in the study design and to include the concept of time warps in the model.

This dissertation proposes a Bayesian mixed-effects model, called *generic spatiotemporal model*, defined for longitudinal observations on a Riemannian manifold. The fixed effects of the model are used to define an average trajectory and the random effects are used to define individual-specific trajectories. In order to define such individual trajectories, we introduce the notion of "parallel variations" of a curve on a Riemannian manifold, based on the idea of parallel variations of a curve. In contrast to [Singh et al., 2013], the distribution of the random effects has the same form (up to an isometric transformation) at any time point along the average trajectory. This time-invariance property allows then the inclusion of time reparametrizations defined using temporal random effects of the model. The individual time reparametrizations are used to alter the pace at which a parallel variation of the average trajectory is followed, therefore allowing to account for the possible delay and changes in speed of progression among the population. As a consequence, the generic Bayesian mixed-effects model presented in this dissertation includes spatial and temporal transformations which allow to put into correspondence individual trajectories and therefore define spatiotemporal distributions of trajectories on a Riemannian manifold.

The generic spatiotemporal model gives a systematic way to derive specific nonlinear mixed-effects models for a large variety of observations and Riemannian manifolds. We will give the particular form of the model with bounded and unbounded measurements such as points on $]0, 1[$, symmetric positive definite matrices, as well as product of such elementary manifolds. Such an intrinsically nonlinear mixed-effects model raises the choice of an adapted algorithm to estimate its parameters given a set of longitudinal observations.

II.2 Inference methods and algorithms for nonlinear mixed-effects models

The generic spatiotemporal model is a *nonlinear* mixed-effects model. Several methods and algorithms have been proposed in the literature for the statistical inference in nonlinear mixed-effects models (NLME). These algorithms may be grouped into two categories: *deterministic* and *stochastic* algorithms. As opposed to stochastic algorithms, the deterministic ones do not require to generate random numbers.

II.2.1 Deterministic algorithms

For NLME models, the observed likelihood $q(\mathbf{y} \mid \boldsymbol{\theta})$ is often intractable as it writes an an integral over the random effects (or *latent variables*) of the model:

$$q(\mathbf{y} \mid \boldsymbol{\theta}) = \int_{\boldsymbol{\beta}} q(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) q(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \, d\boldsymbol{\beta} \qquad \text{[ii.5]}$$

Deterministic methods aim at producing maximum likelihood (or maximum a posteriori) by approximating the observed likelihood. Since the 1990's, many methodological contributions have been made to this topic. In [Lindstrom and Bates, 1990], the authors proposed a two-steps algorithm called "Linear Mixed-Effects approximation algorithm" (LME approximation), which consists in linearizing the NLME model. The first step of the LME algorithm is called *Penalized Nonlinear Least-Squares step* (PNLS step) and consists in minimizing a nonlinear sum of squares. This step is equivalent to maximizing the joint conditional distribution $q(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{y}, \widetilde{\boldsymbol{\theta}}) \propto q(y \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \widetilde{\boldsymbol{\theta}}) q(\boldsymbol{\beta} \mid \widetilde{\boldsymbol{\theta}})$, where $\boldsymbol{\alpha}$ (respectively $\boldsymbol{\beta}$) denote the fixed (respectively random) effects of the NLME model. In this step, the vector $\widetilde{\boldsymbol{\theta}}$ of variance-covariance parameters is considered fixed. The fixed (and random) effects $\widehat{\boldsymbol{\alpha}}$ (and $(\widehat{\boldsymbol{\beta}}_i)_{1 \leq i \leq p}$) which maximize $q(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{y}, \widetilde{\boldsymbol{\theta}})$ are called *conditional modes*. These conditional modes are used in the second step of the algorithm. In this step, a first-order Taylor expansion of the model function around the conditional modes is used to linearize the model. Thus, a LME model is obtained and its parameters are estimated and used to update the estimates of the variance-covariance parameters. However, the LME algorithm is based on an approximation of the observed likelihood without any control or theoretical guarantee of the convergence toward a local maximum of the likelihood. Moreover, the method has to be adapted in order to be used within a Bayesian framework.

In [Davidian and Gallant, 1992], Davidian et al. used an adaptive Gaussian quadrature to approximate the likelihood Eq. [ii.5]. A particular case of this method, called the *Laplacian approximation*, is obtained by considering the adaptive Gaussian quadrature method with only one quadrature point. Both the adaptive Gaussian quadrature and the Laplacian approximation are discussed in [Pinheiro, 1994] and in the book [Pinheiro and Bates, 2006]. The Laplacian Approximation, first introduced in [Tierney and Kadane, 1986], is used to approximate the individual observed likelihood $q(\mathbf{y}_i \mid \boldsymbol{\theta})$. Similarly to the LME algorithm, the approximation is based on a first-order Taylor expansion. This approximation writes:

$$q(\mathbf{y}_i \mid \boldsymbol{\theta}) \simeq \frac{\det \boldsymbol{\Delta}}{(\sigma^2 2\pi)^{k_i/2}} \exp\Big(-\frac{1}{2\sigma^2} g_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i) \Big) \Big(\det \frac{\partial^2 g_{\widetilde{\boldsymbol{\theta}}}}{\partial \boldsymbol{\beta}_i^2}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i) \Big)^{-1/2} \qquad \text{[ii.6]}$$

where $\boldsymbol{\Delta}$ is the *precision* matrix such that $\mathbf{D}^{-1} = \sigma^{-2} \boldsymbol{\Delta}^\top \boldsymbol{\Delta}$, $g_{\widetilde{\boldsymbol{\theta}}}$ is defined by: $g_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_i) = \|\mathbf{y}_i - f(\psi_i, t_i)\|^2 + \|\boldsymbol{\Delta} \boldsymbol{\beta}_i\|^2$ and $\widehat{\boldsymbol{\beta}}_i(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\theta}}) = \underset{\boldsymbol{\beta}_i}{\operatorname{argmin}} \, g(\boldsymbol{\alpha}, \boldsymbol{\beta}_i, \widetilde{\boldsymbol{\theta}})$. A notable drawback of this approximation is that it requires the computation of the inverse of the

Hessian matrix of the sum of squares $g_{\widetilde{\boldsymbol{\theta}}}$ at $(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i)$. In practice, for complex models, this matrix cannot be computed in closed-form and its approximation using numerical schemes is very costly. To address this problem, Pinheiro proposed to approximate the Hessian matrix. Using this approximation, Equation Eq. [ii.6] reduces to a more tractable form which can be maximized using gradient descent methods. Still, the Laplacian approximation remains computationally intensive and without control over the approximation and convergence of the algorithm.

The Expectation-Maximization (EM) [Dempster et al., 1977] is a popular algorithm which allows to obtain *maximum likelihood* (or *maximum a posteriori*) estimates of the parameters of a LME or NLME model. For LME models, the EM algorithm is described in [Laird and Ware, 1982, Laird et al., 1987]. The EM algorithm was introduced in the context of statistical models with latent variables. The idea of the EM algorithm is to maximize a *lower bound* on the observed likelihood. Under generic conditions described in [Dempster et al., 1977], corrected in [Wu, 1983] and generalized in [Delyon et al., 1999], the algorithm converges to a local maximum of the observed likelihood. The EM algorithm iterates, until convergence, between two steps: the "E-step" and the "M-step". Let $\mathbf{y}$ (respectively $\mathbf{z}$) denote the observations (respectively latent variables) of the generic spatiotemporal model. Let $k \in \mathbb{N}^*$ and $\boldsymbol{\theta}^{(k)}$ denote the estimate of the parameters of the model at the kth iteration of the algorithm. Let $q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$ denote the distribution of the observations and latent variables conditionally on the parameters $\boldsymbol{\theta}$. The "E-step" consists in computing the function $\boldsymbol{\theta} \in \Theta \mapsto \mathbf{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ defined by:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \mathbb{E}_{q(\cdot \mid \mathbf{y}, \boldsymbol{\theta}^{(k)})}\big[\log q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})\big]. \qquad \text{[ii.7]}$$

In Eq. [ii.7], the expectation is taken with respect to the conditional distribution of the latent variables knowing the observations and current estimate of the parameters. This function $\mathbf{Q}(\cdot \mid \boldsymbol{\theta}^{(k)})$ is then used in the "M-step" to update the estimate of the parameters as follows:

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \big(\mathbf{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) + q_{\text{prior}}(\boldsymbol{\theta})\big). \qquad \text{[ii.8]}$$

However, for most nonlinear mixed-effects models, the "E-step" of the EM algorithm is intractable because the expectation cannot be computed in closed-form and the conditional distribution $q(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta})$ of the latent variables given the observations $\mathbf{y}$ and the parameters $\boldsymbol{\theta}$, in general, does not belong to a known family of probability distributions.

II.2.2 Stochastic algorithms

As mentioned above, the "E-step" of the EM algorithm may be intractable in NLME models. To address this problem, one can consider a stochastic version of this algorithm: the *Monte Carlo Markov Chains - Stochastic Approximation EM* (MCMC-SAEM) algorithm. In contrast with deterministic algorithms, the MCMC-SAEM may

be used without computing derivatives of the model function. Indeed, only evaluations of the model function f are required to use the MCMC-SAEM. The algorithm, which is proved convergent in [Allassonnière et al., 2010] (based on the work of [Kuhn and Lavielle, 2004]), alternates between three steps: *simulation*, *stochastic approximation* and *maximization*. Let $\boldsymbol{\theta}^{(k)}$ denote the estimates of the model parameters at the kth iteration. In the simulation step, a set of latent variables $\mathbf{z}^{(k)}$ is sampled using the transition kernel of an ergodic Markov chain whose stationary distribution is the conditional distribution $q(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}^{(k)})$ of the latent variables $\mathbf{z}$ given the observations $\mathbf{y}$ and the current estimates $\boldsymbol{\theta}^{(k)}$. In the stochastic approximation step, a function $\mathbf{Q}_k(\cdot)$ is defined by:

$$\forall \boldsymbol{\theta} \in \Theta,\ \mathbf{Q}_k(\boldsymbol{\theta}) = \mathbf{Q}_{k-1}(\boldsymbol{\theta}) + \varepsilon_k\big(\log q(\mathbf{y}, \mathbf{z}^{(k)} \mid \boldsymbol{\theta}) - \mathbf{Q}_{k-1}(\boldsymbol{\theta})\big). \qquad \text{[ii.9]}$$

with $\mathbf{Q}_0 = 0$. The choice of the sequence $(\varepsilon_k)_{k\geq 0}$ is discussed later. Equation Eq. [ii.9] is a stochastic approximation of the Robbins-Monro type [Robbins and Monro, 1951] which converges to the expectation $\mathbb{E}_{q(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(k-1)})}\big[\log q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})\big]$. This stochastic approximation step is asymptotically equivalent to the "E-step" of the classical EM algorithm. Finally, the maximization step consists in maximizing the function $\mathbf{Q}_k$ on the parameter space Θ to update the current estimates of the parameters. Note that the MCMC-SAEM only provides "point estimates", as the algorithm converges to the modes, *id est* a local maximum, of the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$.

Other stochastic algorithms include "Fully-Bayesian" methods. This term refers to a class of Monte Carlo Markov Chains (MCMC) algorithms which aim at learning the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$. To achieve this goal, these algorithms create an ergodic Markov chain whose stationary distribution is the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$. After a number of iterations (which corresponds to a "burn-in period"), samples from this Markov chain are approximately distributed as $q(\boldsymbol{\theta} \mid \mathbf{y})$. With a sufficiently large number of samples from this Markov chain, one can reconstruct the posterior distribution using, for examples, Kernel Density Estimates and derive other informations about this posterior, such as its modes. Hamiltonian Monte Carlo (HMC) is a popular MCMC method which can be used for Bayesian inference. In [Hoffman and Gelman, 2014], the authors propose an adaptive HMC sampler called the *No U-Turns Sampler* (NUTS). This sampler is implemented in a R/C++ library called `STAN`. In this paper, the authors mention that the NUTS sampler offers much better performance in high-dimensional settings that classical samplers such as the Gibbs sampler or the Metropolis-Hastings algorithm. However, HMC samplers require to integrate a system of Hamiltonian equations and compute gradients of the posterior with respect to the parameters of the model, which can be intractable.

In this dissertation, the parameters of the generic spatiotemporal model will be estimated using the MCMC-SAEM algorithm. The validation of this algorithm with the generic spatiotemporal model is considered in this dissertation. In particular, the MCMC-SAEM is tested on longitudinal observations of symmetric positive definite matrices, which represents a non-trivial example of Riemannian manifold of non-positive

curvature. Moreover, the results and runtime of the proposed algorithm is compared with other classical algorithms for statistical inference in NLME models. Finally, the use of the MCMC-SAEM in a Riemannian framework and its compatibility with numerical schemes for the computation of the individual spatiotemporal transformations is discussed. Experimental results proposed in this dissertation are obtained by analyzing longitudinal datasets of health data. In particular, the datasets of neuropsychological test scores and cortical thickness measurements provide insightful informations regarding the progression of Alzheimer's Disease (AD) among a population of individuals from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. The results obtained with these longitudinal datasets validate the use of spatiotemporal transformations. Indeed, we show that the generic model allows to estimate a normative scenario of AD progression and its effects on the cognitive functions. The model also estimates the relative timing between these cognitive impairments. Furthermore, the analysis of longitudinal datasets of symmetric positive definite matrices and body fat measurements shows that the generic spatiotemporal model can estimate a data-driven model of progression and the spatiotemporal transformations actually put into correspondence the progression of individuals.

II.3 Overview of the chapters

Chapter III consists in an overview of key notions in Riemannian geometry and Monte Carlo Markov Chain theory. As emphasized in the previous section, the generic Bayesian model introduced later in this dissertation is defined for longitudinal observations on a Riemannian manifold and trajectories of progression are curves on a Riemannian manifold. The notions presented in Chapter III define a coherent and flexible mathematical framework in which the generic model is defined. Notions of Monte Carlo Markov Chain theory are needed to introduce the stochastic version of the EM algorithm which is used to estimate the parameters of the model.

The generic Bayesian model for longitudinal observations on a Riemannian manifold is presented in Chapter IV. This chapter starts by defining the notion of "parallel variations" of a curve. The definition of a parallel variation requires that some random effects of the model, called *space shifts*, satisfy an orthogonality condition. Properties and examples of parallel variations are given at the beginning of the chapter. These examples show that the notion of parallel variations on a Riemannian manifold generalizes the notion of parallelism to non-Euclidean spaces. Eventually, the generic Bayesian model is presented in Section IV.3 and numerical schemes to ensure that the orthogonality conditions is satisfied are proposed.

Chapter V introduces several particular cases of the generic model. As mentioned above, the generic spatiotemporal model enables, given the choice of a Riemannian manifold and Riemannian metric, to obtain a variety of nonlinear mixed-effects models. Chapter V aims at giving nonlinear mixed-effects models which result from different

choices of Riemannian manifold. A model called "the logistic curves model" (respectively "the straight lines model") is proposed for the analysis of bounded (respectively unbounded) scalar measurements. In the following sections, two specific models are proposed for multivariate observations. One allows the analysis of longitudinal datasets of symmetric positive definite matrices, of which covariance matrices are a particular case. Also, a "propagation model" is proposed to model the temporal progression of a family of biological characteristics or family of features. This model also estimates the relative timing between the progression of two features.

Chapter VI deals with the estimation of the parameters of the generic model. The first section of this chapter starts with a review of several methods for the statistical inference in nonlinear mixed-effects models and motivates the choice of a stochastic version of the EM algorithm is chosen for the estimation of the parameters of the model. The following sections of this chapter describe discuss how this stochastic algorithm can be used within a Riemannian framework. The last section discusses the validation of the proposed algorithm and computational aspects.

Eventually, chapter VII presents experimental results obtained with the particular cases of the generic spatiotemporal model. The results presented in this chapter were obtained by analyzing several longitudinal datasets of health data. These results show that the estimated spatiotemporal transformation put in correspondence similar events along the individual trajectories. The generic model also succeeds in estimating a specific event, change in the evolution which occurs for each individual during the observation period.

II.4 List of publications

This thesis led to the following publications.

II.4.1 Peer-reviewed conference articles

- [Schiratti et al., 2015d] **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Alexandre Routier, the Alzheimer's Disease Neuroimaging Initiative (ADNI), Olivier Colliot and Stanley Durrleman. A mixed-effects model for longitudinal univariate manifold-valued data. In: Ourselin, S., Alexander, D. C., Westin, C.-F., Cardoso, M. J. (Eds.), *Information Processing in Medical Imaging. - IPMI 2015*. No. 9123 in Lecture Notes in Computer Science. Springer International Publishing. Poster elected among the three best posters of the conference yielding an elective oral presentation.

- [Schiratti et al., 2015a] **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Olivier Colliot and Stanley Durrleman. Learning spatiotemporal trajectories

from manifold-valued longitudinal data. In: *Advances in Neural Information Processing Systems. - NIPS 2015.* pp. 2404-2412. Poster and travel award.

- [Schiratti et al., 2015b] **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Olivier Colliot and Stanley Durrleman. Mixed-effects model for the spatiotemporal analysis of longitudinal manifold-valued data. In: *5th MICCAI Workshop on Mathematical Foundations of Computational Anatomy - MFCA 2015.* Oral presentation

II.4.2 Presentations

- [Schiratti et al., 2015c] **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Alexandre Routier, Olivier Colliot and Stanley Durrleman. Estimating profiles of disease progression using mixed-effects models with time reparametrization. In: *Organization for Human Brain Mapping - OHBM 2015.* Poster

- **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Olivier Colliot and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued measurements. In: *Kickoff Workshop of the European Progression of Neurodegenerative Diseases initiative (EuroPOND) - 2016.* Oral presentation

II.4.3 Patents

- **Jean-Baptiste Schiratti**, Stéphanie Allassonnière, Olivier Colliot and Stanley Durrleman. A method for determining the temporal progression of a biological phenomenon and associated methods and devices. Patent application PCT/IB2016/052699.

Part III

Mathematical background

Summary

III.1 Notions of Riemannian geometry

Riemannian geometry consists in the study of Riemannian manifolds, a smooth manifold equipped with a Riemannian metric, and their properties. Section III.1.1 starts by giving a formal definition of smooth manifolds and reviews elementary notions of differential geometry on such manifolds. Sections III.1.2, III.1.5 and III.1.4 introduce key concepts of Riemannian geometry which are used to define the generic model in Chapter IV. A comprehensive overview of differential geometry in smooth manifolds and Riemannian geometry can be found in [Lang, 1972, Gallot et al., 1990, Do Carmo Valero, 1992, Lee, 2003, Petersen, 2006].

III.1.1 Smooth manifolds

The general notion of manifold is quite difficult to define precisely. A surface gives the idea of a two-dimensional manifold. If we take for instance a sphere, or a torus, we can decompose this surface into a finite number of parts such that each of them can be bijectively mapped into a simply-connected region of the Euclidean plane.

Elie Cartan - *Leçons sur la Géométrie des espaces de Riemann* (1928)

Let $n \in \mathbb{N}^*$ and $\mathbb{M}$ be a Hausdorff topological space such that every point of $\mathbb{M}$ admits a neighborhood homeomorphic to an open subset of $\mathbb{R}^n$. Such a space is called a **topological manifold**. Around each point, an homeomorphism allows to identify the manifold with the Euclidean space. However, this is not enough to carry the differentiable structure from the Euclidean space to the manifold. The following notions will allow to define *smooth manifolds* and do differential calculus on these spaces.

A **smooth atlas** on $\mathbb{M}$ is a collection $\{\mathcal{U}_\alpha, \phi_\alpha\}_{\alpha \in I}$ where $(\mathcal{U}_\alpha)_{\alpha \in I}$ is an open cover of $\mathbb{M}$ and the maps $\phi_\alpha : \mathcal{U}_\alpha \to \phi_\alpha(\mathcal{U}_\alpha) \subset \mathbb{R}^n$ are homeomorphisms onto subsets of $\mathbb{R}^n$. In addition to this, for any $\alpha, \beta \in I$, the maps $\psi_{\alpha,\beta} = \phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \to \phi_\beta(\mathcal{U}_\alpha \cap \mathcal{U}_\beta)$ are smooth diffeomorphisms, called **transition maps**. A pair $(\mathcal{U}_\alpha, \phi_\alpha)$ is a **local chart** (or **local coordinates chart**) for $\mathbb{M}$. If $\mathcal{A}$ denotes an atlas on $\mathbb{M}$, a local chart $(\mathcal{U}, \phi)$ on $\mathbb{M}$ is **compatible** with $\mathcal{A}$ if $\mathcal{A} \cup \{(\mathcal{U}, \phi)\}$ is a smooth atlas. An atlas $\mathcal{A}$ is **maximal** if it contains all the local charts that are compatible with it. Finally, a maximal atlas on $\mathbb{M}$ is called **smooth differentiable structure**.

Definition III.1. A n-dimensional **smooth manifold** $\mathbb{M}$ is a topological manifold equipped with a (smooth) differentiable structure.

Example 1 (Sphere). The sphere $\mathbb{S}^n \subset \mathbb{R}^{n+1}$ defined by:

$$\mathbb{S}^n = \left\{ \mathbf{x} \in \mathbb{R}^{n+1},\ \mathbf{x}^\top \mathbf{x} = 1 \right\}$$

is a n-dimensional smooth manifold of $\mathbb{R}^n$. Two charts are enough to define an atlas on $\mathbb{S}^n$. These charts are given by the **stereographic projections** (see [Gallot et al., 1990]).

The Euclidean space $\mathbb{R}^n$ is another example of n-dimensional smooth manifold. Indeed, an atlas on $\mathbb{R}^n$ is given by $(\mathbb{R}^n, \mathrm{Id})$. Similarly, every open subset of $\mathbb{R}^n$ is a n-dimensional smooth manifold. Together with the sphere $\mathbb{S}^n$, these are actually examples of *smooth submanifolds of the Euclidean space* $\mathbb{R}^n$. More precisely, if $n, k \in \mathbb{N}^*$ and $\mathbb{M} \subset \mathbb{R}^n$, $\mathbb{M}$ is a k-dimensional **smooth submanifold** of $\mathbb{R}^n$ if every point in $\mathbb{M}$ has an open neighborhood which is diffeomorphic to an open subset of $\mathbb{R}^k$. Every submanifold of $\mathbb{R}^n$ is a manifold according to Definition III.1. However, some manifolds are naturally defined as quotient spaces or embedded in infinite dimensional Hilbert spaces and cannot be considered as *submanifolds* of the Euclidean plane.

Other examples of interest in this dissertation are given below.

Example 2 (Linear group). Let $\mathcal{M}(n, \mathbb{R})$ denote the vector space of $n \times n$ real matrices and $\mathbf{GL}(n, \mathbb{R})$ be the group of *invertible* $n \times n$ real matrices. $\mathrm{GL}(n, \mathbb{R})$ is an open subset of $\mathcal{M}(n, \mathbb{R})$. Therefore, it is a smooth submanifold of $\mathcal{M}(n, \mathbb{R}) \simeq \mathbb{R}^{n^2}$.

Example 3 (Products of manifolds). If $\mathbb{M}_1$ (respectively $\mathbb{M}_2$) is a n_1-dimensional (respectively n_2-dimensional) smooth manifold, the Cartesian product $\mathbb{M}_1 \times \mathbb{M}_2$ can be equipped with a structure of $(n_1 + n_2)$-dimensional smooth manifold.

The notion of *local charts* introduced above allows to define smooth maps between manifolds, by going back to open subsets of the Euclidean space. Indeed, if $\mathbb{M}$ and $\mathbb{V}$ denote two smooth manifolds and $f : \mathbb{M} \to \mathbb{V}$, the map f is **smooth** if and only if it is continuous and its "expression in local charts" $\psi \circ f \circ \phi^{-1} : \phi\big(\mathcal{U} \cap f^{-1}(\mathcal{V})\big) \to \psi(\mathcal{V})$ is smooth for every local chart $(\mathcal{U}, \phi)$ (respectively $(\mathcal{V}, \psi)$) on $\mathbb{M}$ (respectively $\mathbb{V}$). The notion of smooth map between manifolds is independent of the choice of an atlas. This definition allows to differentiate maps which take values in a smooth manifold. This is particularly useful to define the *velocity of a curve* on a smooth manifold and the notion of *tangent space*.

Tangent spaces

Let $\mathbb{M}$ be a n-dimensional smooth manifold and $\mathbf{p} \in \mathbb{M}$. The tangent space at $\mathbf{p}$ is the space of all possible velocities $\dot{c}(0)$ for any curve $c :]-\varepsilon, \varepsilon[\to \mathbb{M}$ such that $c(0) = \mathbf{p}$. We shall see later that the tangent space at $\mathbf{p}$ provides a linear approximation of the manifold around $\mathbf{p}$.

Let $\mathbb{M}$ be a smooth manifold and $\mathbf{p} \in \mathbb{M}$. Let $(\mathcal{U}, \phi)$ be a local chart around $\mathbf{p}$ and $\varepsilon > 0$. Two curves $c_1, c_2 :]-\varepsilon, \varepsilon[\to \mathbb{M}$ with $c_1(0) = c_2(0) = \mathbf{p}$ are **equivalent at** $\mathbf{p}$ if $(\phi \circ c_1)'(0) = (\phi \circ c_2)'(0)$. The relation "equivalent at $\mathbf{p}$" defines an equivalence relation. For a curve $c :]-\varepsilon, \varepsilon[\to \mathbb{M}$ such that $c(0) = \mathbf{p}$, its equivalence class is denoted by $[c]_{\mathbf{p}}$.

Definition III.2. The tangent space to $\mathbb{M}$ at $\mathbf{p}$ is defined by:

$$\mathrm{T}_{\mathbf{p}}\mathbb{M} = \big\{ [c]_{\mathbf{p}},\ c :]-\varepsilon, \varepsilon[\to \mathbb{M} \text{ smooth with } c(0) = \mathbf{p} \big\}. \qquad \text{[iii.1]}$$

Another definition of tangent spaces consists in defining $\mathrm{T}_{\mathbf{p}}\mathbb{M}$ as the set of derivation on $\mathbb{M}$ at $\mathbf{p}$. Let $\mathcal{C}^\infty(\mathbb{M})$ denote the space of smooth functions on $\mathbb{M}$. A **derivation** on $\mathbb{M}$ at $\mathbf{p}$ is a linear map $D : \mathcal{C}^\infty(\mathbb{M}) \to \mathbb{R}$ which satisfies the Leibniz rule: $\forall (f, g) \in \mathcal{C}^\infty(\mathbb{M})$, $D(fg) = f(\mathbf{p})D(g) + D(f)g(\mathbf{p})$. Indeed, if $\mathbf{v} \in \mathrm{T}_{\mathbf{p}}\mathbb{M}$, the corresponding derivation is the map $\mathbf{v}[\cdot] : \mathcal{C}^\infty(\mathbb{M}) \to \mathbb{R}$ such that: $\mathbf{v}[f] = \frac{d}{dt}\big(f \circ c\big)\big|_{t=0}$, where c is any smooth curve on $\mathbb{M}$ such that $c(0) = \mathbf{p}$ and $\dot{c}(0) = \mathbf{v}$. The map $\mathbf{v}[\cdot]$ is a derivation on $\mathbb{M}$ at $\mathbf{p}$ and the map $\mathbf{v} \mapsto \mathbf{v}[\cdot]$ is a bijective correspondence. Given a coordinate chart (U, ϕ) around $\mathbf{p}$, we can construct a basis of $\mathrm{T}_{\mathbf{p}}\mathbb{M}$ as follows. Define $\partial/\partial x_i$ as the derivation: $\partial/\partial x_i\, f = \frac{d}{dt}\big(f \circ \phi^{-1}\big)\big|_{t=\phi(\mathbf{p})}$. Intuitively, $\partial/\partial x_i$ is the ith derivative in the coordinates given by ϕ. One can show that $(\partial/\partial x_1, \ldots, \partial/\partial x_n)$ forms a basis of $\mathrm{T}_{\mathbf{p}}\mathbb{M}$.

Note that if $\mathbb{M}$ is a smooth submanifold of $\mathbb{R}^n$ and $\mathbf{p} \in \mathbb{M}$, then the tangent space to $\mathbb{M}$ at $\mathbf{p}$ is *exactly* the set of derivatives $\dot{c}(0)$ of smooth curves $c :]-\varepsilon, \varepsilon[\to \mathbb{M}$ with $c(0) = \mathbf{p}$. Theorems 1.22 and 1.23 of [Gallot et al., 1990] can be used to compute the tangent space, at a given point, to a submanifold of $\mathbb{R}^n$. The following examples appear as consequences of these theorems.

Example 4. If $U \subset \mathbb{R}^n$ is an open subset of $\mathbb{R}^n$ and $\mathbf{p} \in U$, then $\mathrm{T}_{\mathbf{p}}U = U$.

Example 5 (Sphere). Let $\mathbf{p} \in \mathbb{S}^n \subset \mathbb{R}^{n+1}$. The tangent space at $\mathbf{p}$ to $\mathbb{S}^n$ is given by:

$$\mathrm{T}_{\mathbf{p}}\mathbf{S}^n = \{\mathbf{p}\}^\perp = \left\{\mathbf{x} \in \mathbb{R}^{n+1},\ \mathbf{x}^\top \mathbf{p} = 0\right\}.$$

Remark. It is important to note that, contrary to a Euclidean space, one should think of the vector in $\mathrm{T}_{\mathbf{p}}\mathbb{M}$ as *attached to the point* $\mathbf{p}$. A direct and important consequence is that if $\mathbf{p}$ and $\mathbf{q}$ are two neighbouring points on $\mathbb{M}$, the tangent spaces $\mathrm{T}_{\mathbf{p}}\mathbb{M}$ and $\mathrm{T}_{\mathbf{q}}\mathbb{M}$ are, in general, different spaces. This remark will play a key role in Section III.1.4.

Having introduced the notions of smooth maps and tangent space, one can define the *differential of a smooth map*.

Definition III.3. Let $\mathbb{M}$,$\mathbb{V}$ be two smooth manifolds and $f : \mathbb{M} \to \mathbb{V}$ a smooth map. Given $\mathbf{p} \in \mathbb{M}$, the **differential** of f at $\mathbf{p}$ is the map $\mathrm{D}_{\mathbf{p}}f : \mathrm{T}_{\mathbf{p}}\mathbb{M} \to \mathrm{T}_{f(\mathbf{p})}\mathbb{V}$ such that: $\forall [c]_{\mathbf{p}} \in \mathrm{T}_{\mathbf{p}}\mathbb{M}$, $\mathrm{D}_{\mathbf{p}}f \cdot [c]_{f(\mathbf{p})} = [f \circ c]_{\mathbf{p}}$, where $c :]-\varepsilon, \varepsilon[\to \mathbb{M}$ is a smooth curve such that $c(0) = \mathbf{p}$.

Note that in the previous definition, the differential $\mathrm{D}_{\mathbf{p}}f \cdot \mathbf{v}$ of f at $\mathbf{p}$ in the direction $\mathbf{v} \in \mathrm{T}_{\mathbf{p}}\mathbb{M}$ does not depend on the choice of the curve c. The previous definition allows to give the following example.

Example 6 (Tangent space to a product of smooth manifolds). Let $n_1, n_2, k_1, k_2 \in \mathbb{N}^*$ and $\mathbb{M}_1 \subset \mathbb{R}^{n_1}$ (respectively $\mathbb{M}_2 \subset \mathbb{R}^{n_2}$) be a k_1-dimensional (respectively k_2-dimensional) smooth manifold. Let $\mathbb{M} = \mathbb{M}_1 \times \mathbb{M}_2$ and $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2) \in \mathbb{M}$. Let π_1 :

$\mathbb{M} \to \mathbb{M}_1$ (respectively $\pi_2 : \mathbb{M} \to \mathbb{M}_2$) be the canonical projection on $\mathbb{M}_1$ (respectively $\mathbb{M}_2$). Note that, by definition, the map

$$\alpha\colon \begin{cases} \mathrm{T}_{\mathbf{p}}\mathbb{M} \longrightarrow \mathrm{T}_{\mathbf{p}_1}\mathbb{M}_1 \times \mathrm{T}_{\mathbf{p}_2}\mathbb{M}_2 \\ \mathbf{v} \quad \longmapsto \left(\mathrm{D}_{\mathbf{p}}\pi_1 \cdot \mathbf{v}, \mathrm{D}_{\mathbf{p}}\pi_2 \cdot \mathbf{v}\right) \end{cases}$$

provides a linear isomorphism between $\mathrm{T}_{\mathbf{p}}\mathbb{M}$ and $\mathrm{T}_{\mathbf{p}_1}\mathbb{M}_1 \times \mathrm{T}_{\mathbf{p}_2}\mathbb{M}_2$. Indeed, introduce the canonical inclusions $\iota_1 : \mathbf{p}_1 \in \mathbb{M}_1 \to (\mathbf{p}_1, \mathbf{p}_2) \in \mathbb{M}$ (respectively $\iota_2 : \mathbf{p}_2 \in \mathbb{M}_2 \to (\mathbf{p}_1, \mathbf{p}_2) \in \mathbb{M}$) and note that the map defined by

$$\beta\colon \begin{cases} \mathrm{T}_{\mathbf{p}_1}\mathbb{M}_1 \times \mathrm{T}_{\mathbf{p}_2}\mathbb{M}_2 \longrightarrow \mathrm{T}_{\mathbf{p}}\mathbb{M} \\ \qquad (\mathbf{x}, \mathbf{y}) \qquad \longmapsto \mathrm{D}_{\mathbf{p}_1}\iota_1 \cdot \mathbf{x} + \mathrm{D}_{\mathbf{p}_2}\iota_2 \cdot \mathbf{y} \end{cases}$$

satisfies $\alpha \circ \beta = \mathrm{Id}$. As a consequence, $\mathrm{T}_{\mathbf{p}}\mathbb{M} \simeq \mathrm{T}_{\mathbf{p}_1}\mathbb{M}_1 \oplus \mathrm{T}_{\mathbf{p}_2}\mathbb{M}_2$.

The next definition introduces the notion of **tangent bundle**. Intuitively, the tangent bundle of $\mathbb{M}$ is the space obtained by "gluing" together all the tangent spaces $(\mathrm{T}_{\mathbf{p}}\mathbb{M})_{\mathbf{p}\in\mathbb{M}}$. This space appears naturally in problems dealing with position and velocity.

Definition III.4. Let $n \in \mathbb{N}^*$ and $\mathbb{M}$ be a n-dimensional smooth manifold. The set

$$\mathrm{T}\mathbb{M} = \bigcup_{\mathbf{p}\in\mathbb{M}} \{\mathbf{p}\} \times \mathrm{T}_{\mathbf{p}}\mathbb{M} = \{(\mathbf{p}, \mathbf{v}),\ \mathbf{p} \in \mathbb{M},\ \mathbf{v} \in \mathrm{T}_{\mathbf{p}}\mathbb{M}\}$$

is the **tangent bundle** of $\mathbb{M}$.

In [Gallot et al., 1990], it is proven that the tangent bundle $\mathrm{T}\mathbb{M}$ can be equipped with a structure of $2n$-dimensional smooth manifold.

Vector fields and affine connections

Definition III.5. A smooth **vector field** on $\mathbb{M}$ is a smooth map V which associates to each point $\mathbf{p} \in \mathbb{M}$ a tangent vector $V(\mathbf{p}) \in \mathrm{T}_{\mathbf{p}}\mathbb{M}$. Equivalently, V is a smooth map from $\mathbb{M}$ to $\mathrm{T}\mathbb{M}$.

The space of vector fields on $\mathbb{M}$ is usually denoted by $\chi(\mathbb{M})$. If (U, ϕ) is a local chart around $\mathbf{p} \in \mathbb{M}$, $(\partial/\partial x_1, \ldots, \partial/\partial x_n)$ is the basis of $\mathrm{T}_{\mathbf{p}}\mathbb{M}$ defined above and $V \in \chi(\mathbb{M})$, then $V(\mathbf{p})$ writes $V(\mathbf{p}) = \sum_i V_i(\mathbf{p}) \frac{\partial}{\partial x_i}$ where $V_1, \ldots, V_n$ are smooth function $U \to \mathbb{R}$. Vector fields "act" on smooth functions in the following way: if f is real-valued smooth function on $\mathbb{M}$, define $(Xf)(\mathbf{p}) = \sum_i V_i(\mathbf{p}) \frac{\partial f}{\partial x_i}(\mathbf{p})$. For every $f \in \mathcal{C}^\infty(\mathbb{M}, \mathbb{R})$, Xf is also a real-valued smooth function on $\mathbb{M}$. Having defined the action of vector fields on $\mathcal{C}^\infty(\mathbb{M}, \mathbb{R})$, if X and Y are two vector fields on $\mathbb{M}$ and $f \in \mathcal{C}^\infty(\mathbb{M}, \mathbb{R})$, then $X(Yf)$ and $Y(Xf)$ make sense. One can show that there exist a unique vector field on $\mathbb{M}$, denoted

by $[X,Y]$ such that: for all $f \in \mathcal{C}^\infty(\mathbb{M},\mathbb{R})$, $[X,Y]f = X(Yf) - Y(Xf)$. This vector field $[X,Y]$ is called the **Lie bracket** of X and Y.

These definitions related to vector fields allow to introduce the notion of *affine connection*, which plays a crucial role in the definition of *geodesics* and *parallel transport* on a Riemannian manifold.

Definition III.6. An **affine connection** on a smooth manifold $\mathbb{M}$ is a mapping $\nabla : (X,Y) \in \chi(\mathbb{M}) \times \chi(\mathbb{M}) \mapsto \nabla_X Y \in \chi(\mathbb{M})$ such that:

(i) ∇ is bilinear,

(ii) ∇ is $\mathcal{C}^\infty(\mathbb{M},\mathbb{R})$-linear in the first variable,

(iii) For all $f \in \mathcal{C}^\infty(\mathbb{M},\mathbb{R})$, $\nabla_X(fY) = f\nabla_X Y + X(f)Y$.

Intuitively, if X and Y are smooth vector fields on $\mathbb{M}$ and $\mathbf{p} \in \mathbb{M}$, $\nabla_X Y(\mathbf{p})$ should be thought as the derivative of Y at $\mathbf{p}$, in the direction of $X(\mathbf{p})$. If $\mathbb{M} \subset \mathbb{R}^n$, a (smooth) vector field on $\mathbb{M}$ is a smooth map $V : \mathbb{M} \to \mathbb{R}^n$. In that case, $\nabla_X Y := \pi_{\mathrm{T}_\mathbf{p}\mathbb{M}}\big(\mathrm{D}_\mathbf{p} Y \cdot X(\mathbf{p})\big)$, where $\pi_{\mathrm{T}_\mathbf{p}\mathbb{M}}$ denotes the orthogonal projection on $\mathrm{T}_\mathbf{p}\mathbb{M}$, defines an affine connection on $\mathbb{M}$.

III.1.2 Riemannian metrics

A **Riemannian metric** is a smooth family of inner product which allows to measure the length of tangent vectors. Therefore, it also allows to measure the length of curves on a manifold. This lead to the introduction of a new family of curves called **geodesics** (see Section III.1.5). More generally, a Riemannian metric allows to do geometry on smooth manifolds.

Definition III.7. A **Riemannian metric** $g^\mathbb{M}$ on $\mathbb{M}$ is a smooth map which associates to each point $\mathbf{p} \in \mathbb{M}$ an inner product $\langle \cdot, \cdot \rangle_\mathbf{p}$ on the linear space $\mathrm{T}_\mathbf{p}\mathbb{M}$. A pair $(\mathbb{M}, g^\mathbb{M})$ is a **Riemannian manifold**.

Let $\mathbf{p} \in \mathbb{M}$ and $\mathbf{x} = (x_1, \ldots, x_n)$ be a coordinate system around $\mathbf{p}$. As discussed above, the tangent vectors $\frac{\partial}{\partial x_1}(\mathbf{p}), \ldots, \frac{\partial}{\partial x_n}(\mathbf{p})$ for a basis of the tangent space $\mathrm{T}_\mathbf{p}\mathbb{M}$. For $(i,j) \in \{1,\ldots,n\}^2$, define: $g_{i,j}(\mathbf{p}) = \left\langle \frac{\partial}{\partial x_i}(\mathbf{p}), \frac{\partial}{\partial x_j}(\mathbf{p}) \right\rangle_\mathbf{p}$. The functions $(g_{i,j})_{1 \leq i,j \leq n}$ are smooth and characterize the Riemannian metric on $\mathbb{M}$.

Theorem III.1 ([Gallot et al., 1990], Theorem 2.2)**.** *There exist at least one Riemannian metric on a smooth manifold.*

Example 7. If $\mathbb{M}$ is a smooth submanifold of $\mathbb{R}^n$, the **induced metric** on $\mathbb{M}$ is the Riemannian metric obtained by restricting the Euclidean metric to each tangent space.

The sphere $\mathbf{S}^n \subset \mathbb{R}^{n+1}$ is usually equipped with the induced metric. More generally, a smooth Riemannian metric on an open subset $U \subset \mathbb{R}^n$ is of the form $g : \mathbf{p} \in U \mapsto g_{\mathbf{p}}$ such that, for all $(\mathbf{u}, \mathbf{v}) \in \mathrm{T}_{\mathbf{p}}U \simeq \mathbb{R}^n$, $g_{\mathbf{p}}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top F(\mathbf{p})\mathbf{v}$, where F is a positive smooth function on U.

Example 8 (Product of Riemannian manifolds). Let $(\mathbb{M}_1, g^{\mathbb{M}_1})$ and $(\mathbb{M}_2, g^{\mathbb{M}_2})$ be two Riemannian manifolds. The product manifold $\mathbb{M} = \mathbb{M}_1 \times \mathbb{M}_2$ can be equipped with a Riemannian metric $g^{\mathbb{M}}$ called **product metric**. Let $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2) \in \mathbb{M}$. With the identification $\mathrm{T}_{\mathbf{p}}\mathbb{M} \simeq \mathrm{T}_{\mathbf{p}_1}\mathbb{M}_1 \oplus \mathrm{T}_{\mathbf{p}_2}\mathbb{M}_2$ (Example 6), $g^{\mathbb{M}}$ is defined as follows:

$$\forall \mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2, \mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 \in \mathrm{T}_{\mathbf{p}}\mathbb{M},\ g^{\mathbb{M}}(\mathbf{u}, \mathbf{v}) = g^{\mathbb{M}_1}_{\mathbf{p}_1}(\mathbf{u}_1, \mathbf{v}_1) + g^{\mathbb{M}_2}_{\mathbf{p}_2}(\mathbf{u}_2, \mathbf{v}_2).$$

III.1.2.1 Push-forward

A way of obtaining Riemannian metrics on a smooth manifold is through the use of a diffeomorphism. Indeed, a diffeomorphism between two smooth manifolds can carry a Riemannian metric from one manifold to another.

Definition III.8. Let $(\mathbb{M}, g^{\mathbb{M}})$ be a Riemannian manifold, $\mathbb{V}$ a smooth manifold and $f : \mathbb{M} \to \mathbb{V}$ a diffeomorphism. The **push-forward** of the metric $g^{\mathbb{M}}$ on $\mathbb{V}$ is the Riemannian metric $f_* g$ defined, for all $\mathbf{p} \in \mathbb{V}$ by:

$$\forall (\mathbf{u}, \mathbf{v}) \in \mathrm{T}_{\mathbf{p}}\mathbb{V},\ (f_* g^{\mathbb{M}})_{\mathbf{p}}(\mathbf{u}, \mathbf{v}) = g^{\mathbb{M}}_{f^{-1}(\mathbf{p})}\big(\mathrm{D}_{\mathbf{p}}(f^{-1}) \cdot \mathbf{u}, \mathrm{D}_{\mathbf{p}}(f^{-1}) \cdot \mathbf{v}\big).$$

The notion of push-forward will be used in Section IV.1 to derive Riemannian metrics on several smooth manifolds.

III.1.2.2 Isometries

As before, $(\mathbb{M}, g)$ denotes a Riemannian manifold equipped with a Riemannian metric g. Let $f : \mathbb{M} \to \mathbb{M}$ be a smooth map. The notion of *isometry* of a Riemannian manifold is defined as follows.

Definition III.9. A smooth map $f : \mathbb{M} \to \mathbb{M}$ is a **local isometry** of $\mathbb{M}$ if:

$$\forall \mathbf{p} \in \mathbb{M},\ \forall (\mathbf{u}, \mathbf{v}) \in \mathrm{T}_{\mathbf{p}}\mathbb{M},\ g_{f(\mathbf{p})}\big(\mathrm{D}_{\mathbf{p}} f \cdot \mathbf{u}, \mathrm{D}_{\mathbf{p}} f \cdot \mathbf{v}\big) = g_{\mathbf{p}}(\mathbf{u}, \mathbf{v}). \qquad \text{[iii.2]}$$

A local isometry is called **isometry** if it is a global diffeomorphism.

The set of isometries of $\mathbb{M}$ is a group denoted by $\mathrm{Isom}(\mathbb{M})$.

III.1.2.3 Gradient

Let $f : \mathbb{M} \to \mathbb{R}$ be a smooth function defined on a Riemannian manifold $(\mathbb{M}, g)$. Let $\mathbf{p} \in \mathbb{M}$. Since the Riemannian metric g defines an inner product on the tangent space $\mathrm{T}_{\mathbf{p}}\mathbb{M}$, we can define the *gradient* of f at $\mathbf{p}$ following the same ideas as in the Euclidean space $\mathbb{R}^n$.

Definition III.10. The gradient of f at $\mathbf{p}$ is the unique tangent vector $\mathrm{grad}f(\mathbf{p}) \in \mathrm{T}_{\mathbf{p}}\mathbb{M}$ such that:

$$\forall \mathbf{v} \in \mathrm{T}_{\mathbf{p}}\mathbb{M},\ g_{\mathbf{p}}\big(\mathrm{grad}f(\mathbf{p}), \mathbf{v}\big) = \mathrm{D}_{\mathbf{p}}f \cdot \mathbf{v} \qquad \text{[iii.3]}$$

where $\mathrm{D}_{\mathbf{p}}f : \mathrm{T}_{\mathbf{p}}\mathbb{M} \to \mathrm{T}_{f(\mathbf{p})}\mathbb{R} \simeq \mathbb{R}$ is the differential of f at $\mathbf{p}$.

Let $\mathbf{p} \in \mathbb{M}$ and $\mathbf{x} = (x_1, \ldots, x_n)$ be a coordinate system around $\mathbf{p}$. Let $(\partial/\partial x_1, \ldots, \partial/\partial x_n)$ denote the basis of the tangent space $\mathrm{T}_{\mathbf{p}}\mathbb{M}$ defined above. Since the gradient $\mathrm{grad}f(\mathbf{p})$ is a tangent vector in $\mathrm{T}_{\mathbf{p}}\mathbb{M}$, it writes: $\mathrm{grad}f(\mathbf{p}) = \sum_{i=1}^{n} a_i \frac{\partial}{\partial x_i}$. Applying Definition III.10 with $\mathbf{v} = \partial/\partial x_j$ $(1 \leq j \leq n)$, we obtain:

$$\forall 1 \leq j \leq n,\ \frac{\partial f}{\partial x_j}(\mathbf{p}) = \sum_{i=1}^{n} a_i g_{\mathbf{p}}\Big(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\Big) = \sum_{i=1}^{n} a_i g_{i,j}(\mathbf{p}). \qquad \text{[iii.4]}$$

As a consequence:

$$\forall 1 \leq i \leq n,\ a_i(\mathbf{p}) = \sum_{j=1}^{n} g^{i,j}(\mathbf{p}) \frac{\partial f}{\partial x_j}(\mathbf{p}) \qquad \text{[iii.5]}$$

where $g^{i,j}(\mathbf{p})$ denotes the inverse of $g_{i,j}(\mathbf{p})$. This last equation gives the expression of the gradient of f at $\mathbf{p}$ in local coordinates.

III.1.3 Affine connections

As mentioned above, on a smooth manifold of positive dimension, there are infinitely many affine connections (or ways to differentiate vector fields). However, on a Riemannian manifold, the Riemannian metric $g^{\mathbb{M}}$ ensures the existence of a natural affine connection called the **Levi-Civita connection**. In order to introduce the Levi-Civita connection, a few definitions are required.

If $(\mathbb{M}, g^{\mathbb{M}})$ is a Riemannian manifold and ∇ an affine connection on $\mathbb{M}$, the connection is said to be **compatible with the metric** $g^{\mathbb{M}}$ if $\nabla_X g^{\mathbb{M}}(Y, Z) = g^{\mathbb{M}}(\nabla_X Y, Z) + g^{\mathbb{M}}(Y, \nabla_X Z)$ for any triplet of vector fields (X, Y, Z) on $\mathbb{M}$. In addition to this, the connection ∇ is **symmetric** if: $\nabla_X Y - \nabla_Y X = [X, Y]$ for any pair (X, Y) of vector fields on $\mathbb{M}$.

Theorem III.2 ([Do Carmo Valero, 1992], Theorem 3.36)**.** *On a Riemannian manifold $(\mathbb{M}, g^{\mathbb{M}})$, there exist a unique affine connection which is symmetric and compatible with the metric $g^{\mathbb{M}}$. This affine connection is the* ***Levi-Civita connection****.*

Example 9. If $\mathbb{M} \subset \mathbb{R}^n$ is a smooth submanifold of $\mathbb{R}^n$ (equipped with the induced metric), a (smooth) vector field on $\mathbb{M}$ is a map $X : M \to \mathbb{R}^n$ such that: $\forall \mathbf{p} \in \mathbb{M}$, $X(\mathbf{p}) \in \mathrm{T}_{\mathbf{p}}\mathbb{M}$. If the vector field is regarded only as a map $X : \mathbb{M} \to \mathbb{R}^n$, its differential at $\mathbf{p} \in \mathbb{M}$ provides a mapping $\mathrm{D}_{\mathbf{p}}X : \mathrm{T}_{\mathbf{p}}\mathbb{M} \to \mathbb{R}^n$. If $Y : \mathbb{M} \to \mathbb{R}^n$ is another vector field on $\mathbb{M}$, the derivative of Y at $\mathbf{p} \in \mathbb{M}$, in the direction of $X(\mathbf{p})$, is given by $\mathrm{D}_{\mathbf{p}}Y \cdot X(\mathbf{p})$. However, projecting this quantity orthogonally on the tangent space $\mathrm{T}_{\mathbf{p}}\mathbb{M}$, we obtain a tangent vector and define an affine connection on $\mathbb{M}$, which is precisely the Levi-Civita connection induced by the metric from $\mathbb{R}^n$.

III.1.4 Parallel transport

In general, if $\mathbf{p}$ and $\mathbf{q}$ are two neighbouring points on a smooth manifold $\mathbb{M}$, there is no natural correspondence between the tangent spaces $\mathrm{T}_{\mathbf{p}}\mathbb{M}$ and $\mathrm{T}_{\mathbf{q}}\mathbb{M}$. **Parallel transport** provides a way of comparing tangent vectors which belong to different tangent spaces. The notion of parallel transport will play a key role in the definition of individual trajectories of progression in Chapter IV. Let $(\mathbb{M}, g^{\mathbb{M}})$ denote a Riemannian manifold and ∇ its Levi-Civita connection.

Definition III.11. Let $c : I \subset \mathbb{R} \to \mathbb{M}$ be a differentiable curve on $\mathbb{M}$ and X a vector field along c. The vector field X is **parallel along** c if $\frac{DX}{dt} = 0$.

Proposition III.1 ([Do Carmo Valero, 1992], Proposition 2.6)**.** *Let $c : [0,1] \to \mathbb{M}$ be a smooth curve on $\mathbb{M}$. Let $\mathbf{w}_0 \in \mathrm{T}_{c(0)}\mathbb{M}$. There exist a unique vector field $t \in [0,1] \mapsto \mathbf{w}(t)$ parallel along c such that $\mathbf{w}(0) = \mathbf{w}_0$. This vector field is denoted by $\mathrm{P}_{c,0,t}(\mathbf{w}_0)$.*

This last proposition shows that parallel transport along a curve $c : [0,1] \to \mathbb{M}$ defines a mapping $(t, \mathbf{w}_0) \in [0,1] \times \mathrm{T}_{c(0)}\mathbb{M} \to \mathrm{P}_{c,0,t}(\mathbf{w}_0) = \mathbf{w}_t \in \mathrm{T}_{c(t)}\mathbb{M}$. The following property of parallel transport will be very useful.

Proposition III.2 ([Do Carmo Valero, 1992])**.** *For all $t \in [0,1]$, the mapping $\mathrm{P}_{c,0,t} : \mathrm{T}_{c(0)}\mathbb{M} \to \mathrm{T}_{c(t)}\mathbb{M}$ is an isometry.*

III.1.5 Geodesics

Intuitively, one can think of geodesics as a generalization of straight lines to Riemannian manifolds. More formally, a geodesic is a smooth curve with zero acceleration. The notion of geodesics allow to introduce the notion of **Riemannian exponential**: a mapping which "generates" geodesics and provides, locally, a parametrization of the manifold by the tangent space. Let $(\mathbb{M}, g^{\mathbb{M}})$ be a Riemannian manifold and ∇ its Levi-Civita connection.

Definition III.12. Let $\gamma : I \subset \mathbb{R} \to \mathbb{M}$ a smooth curve. γ is a **geodesic** of $\mathbb{M}$ if $\nabla_{\dot\gamma}\dot\gamma = 0$. Equivalently, γ is a geodesic if and only if the velocity field $\dot\gamma$ is parallel along γ.

Geodesics locally satisfy to a second order system of differential equations. In several situations, this system of differential equations can be solved, explicitly (or numerically), to compute the geodesics of a Riemannian manifold $\mathbb{M}$. Let $\gamma : I \subset \mathbb{R} \to \mathbb{M}$ a smooth curve on $\mathbb{M}$ and $t \in I$. Let (U, x) be a system of coordinates around $\gamma(t)$. In U, the curve γ writes: $\gamma = (\gamma_1, \ldots, \gamma_n)$ and is a geodesic if and only if:

$$\frac{d^2\gamma_k}{dt^2} + \sum_{1 \leq i,j \leq n} \Gamma^k_{i,j}\big(\gamma(t)\big) \frac{d\gamma_i}{dt} \frac{d\gamma_j}{dt} = 0 \qquad \text{[iii.6]}$$

for all $k \in \{1, \ldots, n\}$. In this system of differential equations, $(\Gamma^k_{i,j})_{1 \leq i,j,k \leq n}$ are the **Christoffel symbols** (of the metric $g^{\mathbb{M}}$), defined by:

$$\Gamma^k_{i,j} = \frac{1}{2} \sum_{l=1}^{n} g^{k,l} \Big(\frac{\partial g_{j,l}}{\partial x_i} + \frac{\partial g_{i,l}}{\partial x_j} - \frac{\partial g_{i,j}}{\partial x_l} \Big). \qquad \text{[iii.7]}$$

Here, $(g_{i,j})_{1 \leq i,j \leq n}$ (respectively, $(g^{i,j})_{1 \leq i,j \leq n}$) are the coefficients of the metric (see III.1.2) (respectively, their inverse).

Example 10 (Euclidean space). In the Euclidean space $\mathbb{R}^n$, the geodesics are of the form $t \in \mathbb{R} \mapsto t\mathbf{A} + \mathbf{B}$ where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^n$. This result follows directly from the form of the Levi-Civita on $\mathbb{R}^n$ (equipped with its canonical metric).

Example 11 (Product of Riemannian manifolds). Let $(\mathbb{M}_1, g^{\mathbb{M}_1})$ and $(\mathbb{M}_2, g^{\mathbb{M}_2})$ be two Riemannian manifolds. Let $\nabla^{\mathbb{M}_1}$ (respectively $\nabla^{\mathbb{M}_2}$) denote the Levi-Civita connection of $\mathbb{M}_1$ (respectively $\mathbb{M}_2$) and let $\mathbb{M} = \mathbb{M}_1 \times \mathbb{M}_2$. We assume that $\mathbb{M}$ is equipped with the product metric (see Example 8). One can show ([Do Carmo Valero, 1992], Chapter 6) that the Levi-Civita connection $\nabla^{\mathbb{M}}$ of $\mathbb{M}$ is characterized by:

$$\forall (X, Z) \in \chi(\mathbb{M}_1)^2,\ \forall (Y, T) \in \chi(\mathbb{M}_2)^2,\ \nabla^{\mathbb{M}}_{X+Y}(Z + T) = \nabla^{\mathbb{M}_1}_X Z + \nabla^{\mathbb{M}_2}_Y T.$$

As a consequence the geodesics of $\mathbb{M}$ are of the form $t \mapsto (\boldsymbol{\gamma}_1(t), \boldsymbol{\gamma}_2(t))$ where $\boldsymbol{\gamma}_1$ (respectively $\boldsymbol{\gamma}_2$) is a geodesic of $\mathbb{M}_1$ (respectively $\mathbb{M}_2$).

Exponential map

As a consequence of Proposition 2.7 ([Do Carmo Valero, 1992]), for each point $\mathbf{p} \in \mathbb{M}$ and time $t_0 \in \mathbb{R}$, there exist an open set $\mathcal{U}$ in $\mathrm{T}_{\mathbf{p}}\mathbb{M}$ such that, for each $\mathbf{v} \in \mathcal{U}$, there exist a unique geodesic $\boldsymbol{\gamma} : I \to \mathbb{M}$, defined on an open neighborhood $I \subset \mathbb{R}$ of t_0 and such that $\boldsymbol{\gamma}(t_0) = \mathbf{p}$, $\dot{\boldsymbol{\gamma}}(t_0) = \mathbf{v}$.

Definition III.13. Let $\mathbf{p} \in \mathbb{M}$, $t_0 \in \mathbb{R}$ and $\mathbf{v} \in U \subset \mathrm{T}_{\mathbf{p}}\mathbb{M}$. The mapping

$$\mathrm{Exp}_{\mathbf{p},t_0}(\mathbf{v})(\cdot) : \begin{cases} I \longrightarrow \mathbb{M} \\ t \longmapsto \mathrm{Exp}_{\mathbf{p},t_0}(\mathbf{v})(t) \end{cases}$$

is the unique geodesic of $\mathbb{M}$ which goes through the point $\mathbf{p} \in \mathbb{M}$ at time t_0, with velocity $\mathbf{v} \in \mathrm{T}_{\mathbf{p}}\mathbb{M}$.

The Riemannian exponential at $\mathbf{p}$, defined in [Do Carmo Valero, 1992, Gallot et al., 1990], appears as a particular case of the previous definition, where the time t_0 equals 0. We have the following definition:

Definition III.14. The **Exponential map** at $\mathbf{p} \in \mathbb{M}$ is the mapping

$$\mathrm{Exp}_{\mathbf{p}} : \begin{cases} U \subset \mathrm{T}_{\mathbf{p}}\mathbb{M} \longrightarrow \mathbb{M} \\ \mathbf{v} \qquad \longmapsto \mathrm{Exp}_{\mathbf{p}}(\mathbf{v}) := \mathrm{Exp}_{\mathbf{p},0}(\mathbf{v})(1) \end{cases}$$

which associates to each $\mathbf{v} \in U$ the value at time $t = 1$ of the unique geodesic $\boldsymbol{\gamma}$ satisfying $\boldsymbol{\gamma}(0) = \mathbf{p}$ and $\dot{\boldsymbol{\gamma}}(0) = \mathbf{v}$.

From Proposition 2.9 of [Do Carmo Valero, 1992], we know that the Riemannian exponential $\mathrm{Exp}_{\mathbf{p}}$ at $\mathbf{p} \in \mathbb{M}$ defines a diffeomorphism from an open ball $B(\mathbf{0}, \varepsilon) \subset \mathrm{T}_{\mathbf{p}}\mathbb{M}$ ($\varepsilon > 0$) onto an open subset of $\mathbb{M}$. The inverse of the Riemannian exponential is denoted by $\mathrm{Log}_{\mathbf{p}}$ and called **Riemannian logarithm**. If the Riemannian exponential $\mathrm{Exp}_{\mathbf{p}}$ is defined on the entire tangent space $\mathrm{T}_{\mathbf{p}}\mathbb{M}$, the Riemannian manifold $\mathbb{M}$ is **geodesically complete**.

III.2 Notions of Markov chains theory

III.2.1 Markov chains, transition kernels and stationary distribution

This section consists in a review of fundamental concepts of Markov chains theory, among which, the transition kernel of a Markov chain and stationary distributions. These notions will play a central role in the following section on Monte Carlo Markov Chains (MCMC) methods. A comprehensive overview of Markov chains theory can be found in [Meyn and Tweedie, 2012, Billingsley, 2013]. More about MCMC methods and their implementation can be found in [Robert and Casella, 2009, Liang et al., 2011, Robert and Casella, 2013].

Markov chains and kernels

Before giving the definition of a Markov chain, we start by introducing the notions of *stochastic process* and *filtration*. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X, \mathcal{G})$ be a measurable space. A **stochastic process** on X is a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables with values in $(X, \mathcal{G})$. A **filtration** on $(\Omega, \mathcal{F})$ is an increasing sequence $(\mathcal{F}_n)_{n\in\mathbb{N}}$ of sub-σ-fields of $\mathcal{F}$ and a **filtered probability space** is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration. If $Y : (\Omega, \mathcal{F}) \to (X, \mathcal{G})$ is measurable, $\sigma(Y)$ is the smallest σ-algebra of $\mathcal{F}$ such that Y is measurable. In particular, if $(X_n)_{n\in\mathbb{N}}$ is

a stochastic process, the **natural filtration** of this process if the filtration $(\mathcal{F}_n^X)_{n\in\mathbb{N}}$ defined by: $\forall n \in \mathbb{N}$, $\mathcal{F}_n^X = \sigma(X_k,\ 0 \leq k \leq n)$. A stochastic process $(X_n)_{n\in\mathbb{N}}$ is **adapted to the filtration** $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if, for each $n \in \mathbb{N}$, X_n is $\mathcal{F}_n$-measurable.

In the following, $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space equipped with a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$.

Definition III.15. An adapted stochastic process $(X_n)_{n\in\mathbb{N}}$ with values in $(X, \mathcal{G})$ is a **Markov chain** if, for all $n \in \mathbb{N}$ and all $G \in \mathcal{G}$,

$$\mathbb{P}(X_{n+1} \in G \mid \mathcal{F}_n) = \mathbb{P}(X_{n+1} \in G \mid X_n) \quad \mathbb{P} - \text{a.e.} \qquad \text{[iii.8]}$$

The measurable space $(X, \mathcal{G})$ is often called **state space**.

With $(\mathcal{F}_n)_{n\in\mathbb{N}} = (\mathcal{F}_n^X)_{n\in\mathbb{N}}$, Definition III.15 states that the future of the chain is conditionally independent of its past, given the present state. An adapted stochastic process $(X_n)_{n\in\mathbb{N}}$ is a Markov chain if and only if, for all positive measurable function f on $(X, \mathcal{G})$, $\mathbb{E}[f(X_{n+1}) \mid \mathcal{F}_n] = \mathbb{E}[f(X_{n+1}) \mid X_n]$ $\mathbb{P} - $ a.e. Examples of Markov chains are given below.

Example 12 (Random walk). Let $(W_n)_{n\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables taking values in $\mathbb{Z}^d$, with distribution μ. Let W_0 be a random variable in $\mathbb{Z}^d$ independent of $(W_n)_{n\in\mathbb{N}^*}$. A *random walk* on $\mathbb{Z}^d$, with distribution μ, is a stochastic process $(X_n)_{n\in\mathbb{N}}$ defined by: $X_0 = W_0$ and, for all $n \in \mathbb{N}^*$, $X_{n+1} = X_n + W_{n+1}$. A random walk on $\mathbb{Z}^d$ is a simple example of Markov chain on a discrete state space.

Example 13 (Auto-regressive process). Auto-regressive processes are a classical example of Markov chains on a continuous state space. The *AR(1)* process $(X_n)_{n\in\mathbb{N}}$ is defined as follows: $\forall n \in \mathbb{N}^*$, $X_n = a + bX_{n-1} + W_n$, where $(W_n)_{n\in\mathbb{N}^*}$ is a sequence of i.i.d. real-valued random variables which are independent of X_0. The chain $(X_n)_{n\in\mathbb{N}}$ is a Markov chain on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

The following definitions introduce the notion of *kernels*. This allows to define later the *transition kernel* of a Markov chain.

Definition III.16. Let $(X, \mathcal{G})$ and $(E, \mathcal{E})$ be two measurable spaces. A **kernel** P on the product $X \times \mathcal{E}$ is a mapping $P : X \times \mathcal{E} \to [0, +\infty]$ such that:

(i) for every $x \in X$, the mapping $P(x, \cdot) : A \in \mathcal{E} \to P(x, A)$ is a measure on $\mathcal{E}$,

(ii) for every $A \in \mathcal{E}$, the mapping $P(\cdot, A) : x \in X \to P(x, A)$ is a positive measurable function on $(X, \mathcal{G})$. The kernel P is a **Markov kernel** if, for all $x \in X$, $P(x, \cdot)$ is a *probability* measure on $(E, \mathcal{E})$.

In particular, if X and E are *countable* sets, $\mathcal{E}$ is the set of $\mathcal{P}(E)$ of parts of E and a kernel on $X \times \mathcal{P}(E)$ can be identified with a *matrix* (possibly infinite) $P = \big(P(x, y),\ x \in X,\ y \in E\big)$. The matrix P is Markovian if, for all $x \in X$, $\sum_{y\in E} P(x, y) = 1$.

Kernels can "act" on positive measurable functions, on measures and can be composed. These operations, which shall be used later, are defined as follows.

Definition III.17. Let P be a kernel on $X \times \mathcal{E}$ and $f : E \to [0, +\infty[$ be a measurable function. The function $Pf : X \to [0, +\infty[$ is defined by:

$$\forall x \in X,\ Pf(x) = \int_X P(x, dy) f(y). \tag{iii.9}$$

If μ denotes a positive measure (respectively probability measure) on $(X, \mathcal{G})$, the positive measure (respectively probability measure) μP on $(E, \mathcal{E})$ is defined by:

$$\forall A \in \mathcal{E},\ \mu P(A) = \int_X \mu(dx) P(x, A). \tag{iii.10}$$

Let $(Z, \mathcal{H})$ be another measurable space and P (respectively Q) a kernel on $X \times \mathcal{E}$ (respectively on $E \times \mathcal{H}$). The composition PQ, defined by:

$$\forall x \in X,\ \forall A \in \mathcal{H},\ PQ(x, A) = \int_E P(x, dy) Q(y, A) \tag{iii.11}$$

is a kernel on $X \times \mathcal{H}$. For a positive measurable function f on $(Z, \mathcal{H})$, for all $x \in X$, $(PQ)f(x) = P(Qf)(x)$.

The previous definitions and properties of kernels are aimed at defining the *transition kernel* of an *homogeneous Markov chain*. Let $(X, \mathcal{G})$ be a measured space, P a kernel and μ a probability measure on $(X, \mathcal{G})$. An adapted stochastic process $(X_n)_{n\in\mathbb{N}}$ is an **homogeneous Markov chain** with **transition kernel** P and **initial distribution** μ if, for all $n \in \mathbb{N}$ and all $A \in \mathcal{G}$, $\mathbb{P}(X_0 \in A) = \mu(A)$ and $\mathbb{P}(X_{n+1} \in A \mid \mathcal{F}_n) = P(X_n, A)$. This last condition is equivalent to: $\mathbb{E}[f(X_{n+1}) \mid \mathcal{F}_n] = Pf(X_n)$ $\mathbb{P}$-almost everywhere and for every measurable positive function on $(X, \mathcal{G})$.

Example 14. In Example 12 of the random walk, $(X_n)_{n\in\mathbb{N}}$ is a Markov chain on $\mathbb{Z}^d$ with transition kernel P defined on $\mathbb{Z}^d \times \mathbb{Z}^d$ by: $\forall (x, y) \in \mathbb{Z}^d,\ P(x, y) = \mu(y - x)$. Regarding Example 13 of the auto-regressive process, under the condition that $\mathbb{E}[|W_1|] < +\infty$ and $\mathbb{E}[W_1] = 0$, $(X_n)_{n\in\mathbb{N}}$ is a Markov chain with transition kernel P defined on $\mathbb{R} \times \mathcal{B}(\mathbb{R})$ by: $\forall x \in \mathbb{R}, \forall A \in \mathcal{B}(\mathbb{R}),\ P(x, A) = \mathbb{P}(W_1 + bx + a \in A)$.

Invariant measure, stationarity and ergodicity

We now introduce the definition of *invariant measure* and the notion of *stationarity*.

Definition III.18. Let P be a Markov kernel on $X \times G$. A positive σ-finite measure π is **invariant** with respect to P if it satisfies $\pi P = P$.

A stochastic process $(X_n)_{n\in\mathbb{N}}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is **stationary** if, for all integers $n, p \in \mathbb{N}$, the distribution of $(X_n, \ldots, X_{n+p})$ does not depend on n. Now, let $(X_n)_{n\in\mathbb{N}}$ is a Markov chain with initial distribution μ and transition kernel P. If

$(X_n)_{n\in\mathbb{N}}$ is stationary, then $\mu P = \mu$. Conversely, if the initial distribution μ satisfies to $\mu P = \mu$, then it can be shown that $(X_n)_{n\in\mathbb{N}}$ is stationary. A probability distribution π which satisfies to $\pi P = \pi$ is called **stationary distribution** of the chain $(X_n)_{n\in\mathbb{N}}$.

If there exist a probability distribution π on $(X, \mathcal{G})$ such that $\lim_{n\to+\infty} \sup_{A\in\mathcal{G}} |\pi P^n(A) - \pi(A)| = 0$, one can show that π in invariant with respect to P. If the limit distribution π is independent of the initial distribution μ, the chain $(X_n)_{n\in\mathbb{N}}$ is **ergodic**. Intuitively, an ergodic Markov chain asymptotically forgets its initial distribution. Moreover, its distribution μP^n converges (as above) to π, which is invariant with respect to P. The notion of ergodicity will play an important role in Section III.2.2 on MCMC algorithms.

A probability measure π on $(X, \mathcal{G})$ is said to be **reversible** with respect to P if, for all $(A, B) \in \mathcal{G}$,

$$\int_A \pi(dx)P(x, B) = \int_B \pi(dy)P(y, A) \qquad \text{[iii.12]}$$

Eq. [iii.12] is called **the detailed balance condition**. It provides a sufficient condition for a Markov kernel to have an invariant distribution. In fact, if π is reversible with respect to P, then π is invariant with respect to P.

III.2.2 Monte Carlo Markov Chains methods

Numerous situations require to compute integrals of the form

$$\mathbb{E}_\pi[f(X)] = \int f(x)\pi(x)dx \qquad \text{[iii.13]}$$

where π is a probability density on $\mathbb{R}^d$. The density π is often quite complex, which avoids direct sampling from π using methods such as the Inverse Transform method or Acceptance-Rejection method. Moreover, in general, only an *unnormalized* version $\tilde{\pi}$ of π is available. Here, $\tilde{\pi}$ is a positive function such that $\tilde{\pi} = C_\pi \pi$ with $C_\pi > 0$. The *normalizing constant* C_π cannot, in general, be computed in closed-form. To address this problem, MCMC methods consist in generating an ergodic Markov chain which admits π as (unique) invariant distribution. Since the chain is ergodic, after some time, the distribution of the terms of the chain should be close to π. However, the terms of the chain are not independent, as opposed to direct sampling methods which produce i.i.d. draws from π. If $(X_n)_{n\in\mathbb{N}}$ is an ergodic Markov chain whose invariant distribution is π, the **ergodic theorem** ensures that:

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \underset{n\to+\infty}{\longrightarrow} \mathbb{E}_\pi[f(X)]. \qquad \text{[iii.14]}$$

Such Markov chains are constructed using **MCMC samplers**. Among the large number of MCMC samplers, two are considered below: the **Metropolis-Hastings sampler** (MH) and the **Gibbs sampler**.

III.2.2.1 Metropolis-Hastings algorithm

Let λ denote a measure on $(X, \mathcal{G})$ and π a positive measurable function on X such that $\int_X \pi(x)\lambda(dx) < +\infty$. Also let Q denote a Markov kernel on $(X, \mathcal{G})$ which has a density q with respect to λ. In other words, Q satisfies: $\forall x \in X$, $\forall A \in \mathcal{G}$, $Q(x, A) = \int_A q(x, y)\lambda(dy)$. The MH algorithm works as follows. Initialized with a value Z_0, the algorithm builds a stochastic process $(Z_n)_{n\in\mathbb{N}}$ of random variables as follows:

Algorithm 1 Metropolis-Hastings algorithm: kth step

1: Given Z_k $(k \geq 0)$:
2: Sample $Z_k^* \sim Q(Z_k, \cdot)$
3: Compute $\alpha(Z_k, Z_k^*)$ defined by:

$$\alpha(Z_k, Z_k^*) = \frac{\pi(Z_k^*)q(Z_k^*, Z_k)}{\pi(Z_k)q(Z_k, Z_k^*)} \wedge 1.$$

4: Sample $U \sim \text{Uniform}([0, 1])$
5: Define Z_{k+1}:

$$Z_{k+1} = \begin{cases} Z_k & \text{if } U \leq \alpha(Z_k, Z_k^*) \\ Z_k^* & \text{otherwise.} \end{cases}$$

6: **Return** Z_{k+1}.

The ratio $\alpha(Z_k, Z_k^*)$ in Algorithm 1 is called **acceptance ratio** and the density q is the **proposal density**. Note that this quantity only depends on the target density π through the ratio $\pi(Z_k^*)/\pi(Z_k)$. As a consequence, the MH algorithm can be used whenever π is known only up to a normalizing constant. This property of the acceptance ratio is particularly desirable in Bayesian inference. The stochastic process $(Z_n)_{n\in\mathbb{N}}$ produced by the algorithm is a Markov chain and its transition kernel is given as follows.

Proposition III.3 ([Liang et al., 2011], Chapter 3)**.** *The transition kernel P of the Markov chain $(Z_n)_{n\in\mathbb{N}}$ is given by:*

$$P(x, A) = \int_A \alpha(x, y)q(x, y)\lambda(dy) + \delta_x(A)\Big(\int_X \big(1 - \alpha(x, y)\big)q(x, y)\lambda(dy)\Big). \qquad \text{[iii.15]}$$

for all $x \in X$ and all $A \in \mathcal{G}$.

Moreover, the following result holds:

Proposition III.4 ([Liang et al., 2011], Chapter 3)**.** *The distribution π is reversible with respect to the transition kernel P.*

III.2.2.2 Gibbs sampler

Consider a product space $X_1 \times \dots X_n$ equipped with the σ field $\mathcal{B}(X_1) \otimes \dots \otimes \mathcal{B}(X_n)$ and π a density on $X_1 \times \dots \times X_n$. In the following, the notation $\mathbf{x}_{-i}$ denotes the vector $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ with $i \in \{1, \dots, n\}$. We assume that, for all i, π can be written: $\pi(dx_1, \dots, dx_n) = \tilde{\pi}_i(d\mathbf{x}_{-i}) R_i(\mathbf{x}_{-i}, dx_i)$ where R_i is a Markov transition kernel on $\prod_{j \neq i} X_j \times \mathcal{B}(X_i)$ given by:

$$R(\mathbf{x}_{-i}, A) = \frac{\displaystyle\int_A \pi(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n)\, dy}{\displaystyle\int_{X_i} \pi(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n)\, dy}. \qquad \text{[iii.16]}$$

for all $\mathbf{x}_{-i} \in \prod_{j \neq i} X_j$ and all $A \in \mathcal{B}(X_i)$. Initialized with $\mathbf{Z}^0 = (Z_1^0, \dots, Z_n^0)$, the **Gibbs sampler** (GS) constructs a stochastic process $(\mathbf{Z}^n)_{n \in \mathbb{N}}$ as follows:

Algorithm 2 Gibbs sampler: kth step

1: Given $\mathbf{Z}^k = (Z_1^k, \dots, Z_n^k)$ $(k \geq 0)$:
2: Sample $Z_1^{k+1} \sim R_1\big((Z_2^k, \dots, Z_n^k), \cdot\big)$
3: Sample $Z_2^{k+1} \sim R_2\big((Z_1^{k+1}, Z_3^k, \dots, Z_n^k), \cdot\big)$
4: $\dots$
5: Sample $Z_n^{k+1} \sim R_n\big((Z_1^{k+1}, \dots, Z_{n-1}^{k+1}), \cdot\big)$.
6: **Return**: $\mathbf{Z}^{k+1}$.

The stochastic sequence $(\mathbf{Z}^n)_{n \in \mathbb{N}}$ is a Markov chain.

Proposition III.5. *The transition kernel of the Markov chain $(\mathbf{Z}^n)_{n \in \mathbb{N}}$ is $R_1 \dots R_n$.*

In addition to this:

Proposition III.6. *The density π is reversible with respect to the transition kernel $R_1 \dots R_n$ of the Markov chain $(\mathbf{Z}^n)_{n \in \mathbb{N}}$.*

Note that using the GS requires to be able to sample from the densities

$$\pi_i(y \mid \mathbf{x}_{-i}) = \frac{\pi(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n)}{\displaystyle\int_{X_i} \pi(x_1, \dots, x_{i-1}, s, x_{i+1}, \dots, x_n)\, ds}. \qquad \text{[iii.17]}$$

These densities are usually called **full conditionals** of π. However, if the target density π is a complex distribution, the normalizing constant of the full conditionals will be intractable. In this situation, a way to address this problem is to use a MH algorithm for the sampling steps of the GS. Then, this leads to a different algorithm called **hybrid Gibbs sampler** or **Metropolis-Hastings-within-Gibbs sampler** (MHwGS).

Part IV

A Bayesian mixed-effects model for longitudinal observations on a Riemannian manifold

Summary

The purpose of this chapter is to introduce a Bayesian mixed-effects model, called *generic spatiotemporal model*, to learn trajectories of progression from longitudinal *manifold-valued* observations. As emphasized in the introduction, the generic spatiotemporal model assumes that each observation is a random perturbation of a point on a Riemannian manifold. In addition to this, the model estimates an average trajectory of progression, which is assumed to be a geodesic on a Riemannian manifold. Individual trajectories result from spatiotemporal transformations, namely *parallel transport* and *time reparametrization*, of the average trajectory. The notions of Riemannian geometry introduced in Chapter III are used in this chapter to define the generic model. Section IV.1 gives examples of Riemannian manifolds in which the geodesics and parallel transport can be computed in closed-form. The Riemannian manifolds discussed in this section will be considered in Chapter VII. Section IV.2 introduces the concept of *parallel variation* of a curve on a Riemannian manifold. This notion is used to define the individual trajectories. It also enforces an orthogonality constraint on some random effects of the model called *space shifts*. Methods to include this orthogonality constraint into the model are discussed in this section. Finally, Section IV.3 presents the generic spatiotemporal model.

IV.1 Geodesics and parallel transport in some classical manifolds

IV.1.1 One-dimensional Riemannian manifolds

IV.1.1.1 Geodesics of a one-dimensional Riemannian manifold

Proposition IV.1 (Geodesics of one-dimensional Riemannian manifolds)**.** *Let $\mathbb{M} \subset \mathbb{R}$ be an open interval of $\mathbb{R}$ and g a Riemannian metric on $\mathbb{M}$. The geodesics of the one-dimensional Riemannian manifold $(\mathbb{M}, g)$ are of the form $t \mapsto \phi(at + b)$ with $a, b \in \mathbb{R}$, $\phi : \mathbb{M} \to \phi(\mathbb{M}) \subset]0, +\infty[$ an increasing $\mathcal{C}^1$ diffeomorphism.*

Proof. Note that the Riemannian metric g is of the form $p \in \mathbb{M} \mapsto g_p$ with: $\forall (u, v) \in \mathrm{T}_p\mathbb{M} \simeq \mathbb{R}$, $g_p(u, v) = uf(p)v$ and $f : \mathbb{M} \to]0, +\infty[$ a smooth function.

It follows from Eq. [iii.7] that the Riemannian metric g is characterized by a single Christoffel symbol $\Gamma^1_{1,1}$. This symbol is defined by: $\forall p \in \mathbb{M}, \Gamma^1_{1,1}(p) = (1/2)(f'(p)/f(p))$. Therefore, γ is a geodesic if and only if it satisfies the differential equation

$$\ddot{\gamma}(t) + \frac{1}{2}\frac{f'\big(\gamma(t)\big)}{f\big(\gamma(t)\big)}\big(\dot{\gamma}(t)\big)^2 = 0. \qquad \text{[iv.1]}$$

Equivalently, γ is a geodesic if and only if:

$$\frac{d}{dt}\Big(\dot{\gamma}(t)\sqrt{(f\circ\gamma)(t)}\Big) = 0. \tag{iv.2}$$

As a consequence, γ is a geodesic if and only if there exist $a \in \mathbb{R}$ such that:

$$\forall t \in \mathbb{R},\ \dot{\gamma}(t)\sqrt{(f\circ\gamma)(t)} = a. \tag{iv.3}$$

Let $p \in \mathbb{M}$ and $F : u \in \mathbb{M} \subset \mathbb{R} \mapsto \int_p^u \sqrt{f}(t)\,dt$. F is an increasing $\mathcal{C}^1$ diffeomorphism from $\mathbb{M}$ to its image $F(\mathbb{M}) \subset \mathbb{R}$. Since Eq. [iv.2] writes $\big(F\circ\gamma\big)'(t) = a$, γ is a geodesic if and only if: $\gamma(t) = F^{-1}(at+b)$. □

IV.1.1.2 The case $\mathbb{M} = \mathbb{R}$

If $\mathbb{M} = \mathbb{R}$ is equipped with the canonical metric (defined by: $\forall p \in \mathbb{R}$, $\forall (u,v) \in \mathrm{T}_p\mathbb{R} \simeq \mathbb{R}$, $g_p^{\mathrm{eucl}}(u,v) = uv$), the geodesics are straight lines of the form $t \in \mathbb{R} \mapsto p + tv$. In particular, if $p_0 \in \mathbb{R}$, $t_0 \in \mathbb{R}$ and $v_0 \in \mathrm{T}_{p_0}\mathbb{R} \simeq \mathbb{R}$, the geodesic $\gamma_0(\cdot) = \mathrm{Exp}_{p_0,t_0}(v_0)(\cdot)$ is defined by:

$$\gamma_0(t) = p_0 + v_0(t - t_0). \tag{iv.4}$$

It follows that $(\mathbb{R}, g^{\mathbb{R}})$ is geodesically complete.

IV.1.1.3 The case $\mathbb{M} =]0,1[$

When $\mathbb{M} =]0,1[$, one could equip $\mathbb{M}$ with the induced metric from $\mathbb{R}$. However, the geodesics for the induced metric are straight lines, which may "go out of $\mathbb{M}$" in finite time. Therefore, $\mathbb{M}$ would not be geodesically complete for the induced metric. To address this problem, consider that $\mathbb{M}$ is equipped with the Riemannian metric $g = (g_p)_{p\in]0,1[}$ defined by:

$$\forall p \in \mathbb{M},\ \forall (u,v) \in T_p\mathbb{M},\ g_p(u,v) = uG(p)v \quad \text{with} \quad G(p) = \frac{1}{p^2(1-p)^2}. \tag{iv.5}$$

This Riemannian metric corresponds to the canonical metric on $\mathbb{R}$, modified by a *conformal factor* G. In the literature, such Riemannian metrics are usually referred to as *conformal metrics*. Proposition IV.1 ensures that the geodesics of $\mathbb{M}$ are of the form $t \in \mathbb{R} \mapsto F^{-1}(at+b)$ with $a \in \mathbb{R}$, $b \in \mathbb{R}$ and F given by: $\forall u \in]0,1[$, $F(u) = \int_{1/2}^u \frac{1}{u(1-u)}\,du = \ln\big(\frac{u}{1-u}\big)$. The inverse mapping of F is given by: $\forall s \in \mathbb{R}$, $F^{-1}(s) = \frac{1}{1+e^{-s}}$. Therefore, the geodesics of $\mathbb{M}$ are of the form: $t \in \mathbb{R} \mapsto \frac{1}{1+e^{-(at+b)}}$ with $a \in \mathbb{R}$ and $b \in \mathbb{R}$. In particular, if $p_0 \in \mathbb{M}$, $t_0 \in \mathbb{R}$ and $v_0 \in \mathrm{T}_{p_0}M$, the geodesic $\gamma_{p_0,t_0,v_0}(\cdot) = \mathrm{Exp}_{p_0,t_0}(v_0)(\cdot)$ of $\mathbb{M}$ is:

$$\forall t \in \mathbb{R},\ \gamma_{p_0,t_0,v_0}(t) = \left(1 + \left(\frac{1}{p_0} - 1\right)\exp\left(\frac{-v_0(t-t_0)}{p_0(1-p_0)}\right)\right)^{-1}. \tag{iv.6}$$

With this metric, the open interval $]0,1[$ is a geodesically complete Riemannian manifold. The geodesics of this metric are usually called *logistic curves* (or *sigmoid curves*).

The Riemannian metric on $]0,1[$ and the logit transform

The open interval $]0,1[$ naturally appears when dealing with *normalized* scalar observations. Usually, in the literature, the common practice when dealing with observations in $]0,1[$ consists in taking the *logit* transform of these observations to map these to the real line. Then, analyzes are performed on these observations using, usually, linear models such as LME models. Recall that the **logit transform** is the map

$$\operatorname{logit}\colon \begin{cases}]0,1[& \longrightarrow \mathbb{R} \\ p & \longmapsto \ln\left(\dfrac{p}{1-p}\right) \end{cases}. \tag{iv.7}$$

Its inverse is the **sigmoid** function

$$S = \operatorname{logit}^{-1}\colon \begin{cases} \mathbb{R} & \longrightarrow]0,1[\\ t & \longmapsto \dfrac{1}{1+e^{-t}} \end{cases}. \tag{iv.8}$$

Consider, as above, $\mathbb{R}$ as a Riemannian manifold equipped with the metric g^{eucl}. Noting that S is a *diffeomorphism* from $\mathbb{R}$ to $]0,1[$, one can use this sigmoid function to *push-forward* (see Definition III.8) the Riemannian metric g^{eucl} onto $]0,1[$. Indeed, for $p \in]0,1[$, the differential of the logit transform is given by:

$$\forall u \in \mathrm{T}_p]0,1[\simeq \mathbb{R},\ \mathrm{D}_p \operatorname{logit} \cdot u = \frac{u}{p-p^2}. \tag{iv.9}$$

Then, it follows from the definition of the push-forward that $\operatorname{logit}_* g^{\text{eucl}}$ defines a Riemannian metric on $]0,1[$, which is given by:

$$\forall p \in]0,1[,\ \forall (u,v) \in \mathrm{T}_p]0,1[\simeq \mathbb{R},\ \left(\operatorname{logit}_* g^{\text{eucl}}\right)_p (u,v) = \frac{uv}{p^2(1-p)^2}. \tag{iv.10}$$

Remark. Also note that this Riemannian metric on the open interval $]0,1[$ could be used to define a Riemannian metric on $\mathbb{S}^1 \smallsetminus \{1\}$, where $\mathbb{S}^1 = \{z \in \mathbb{C},\ |z| = 1\}$. Indeed, the mapping $t \in]0,1[\mapsto e^{2i\pi t}$ is a diffeomorphism from $]0,1[$ onto $\mathbb{S}^1 \smallsetminus \{1\}$ which could be used, again, to push-forward the Riemannian metric on $]0,1[$ onto $\mathbb{S}^1 \smallsetminus \{1\}$.

IV.1.1.4 The case $\mathbb{M} =]0,+\infty[$

Similarly to the case of $\mathbb{M} =]0,1[$, the open interval $]0,+\infty[$ can be equipped with a Riemannian metric. This Riemannian metric is obtained as the push-forward of the canonical metric g^{eucl} on $\mathbb{R}$ by the exponential. We have:

$$\forall p \in]0,+\infty[,\ \forall (u,v) \in \mathrm{T}_p]0,+\infty[\simeq \mathbb{R},\ \left(\exp_* g^{\text{eucl}}\right)_p (u,v) = \frac{uv}{p^2}. \tag{iv.11}$$

As a result of Proposition IV.1, the geodesics of the one-dimensional Riemannian manifold $\mathbb{M} =]0, +\infty[$ are of the form: $\forall t \in \mathbb{R}$, $\gamma(t) = \exp(at + b)$ with $a, b \in \mathbb{R}$. In particular, if $p_0 \in]0, +\infty[$, $v_0 \in \mathrm{T}_{p_0}\mathbb{M} \simeq \mathbb{R}$ and $t_0 \in \mathbb{R}$ then the geodesic $\gamma_0(\cdot) = \mathrm{Exp}_{p_0,t_0}(v_0)(\cdot)$ is given by:

$$\forall t \in \mathbb{R},\ \gamma_0(t) = p_0 \exp\Big(\frac{v_0}{p_0}(t - t_0)\Big). \tag{iv.12}$$

IV.1.2 Product of one-dimensional Riemannian manifolds

Let $M \subset \mathbb{R}$ be an open interval of $\mathbb{R}$ equipped with a Riemannian metric g, such that (M, g) is geodesically complete. Let $N \in \mathbb{N}^*$. The product manifold $\mathbb{M} = M^N$ is equipped with the product metric (see Example 8). As discussed in Example 11, the geodesics of $\mathbb{M}$ are of the form:

$$t \in \mathbb{R} \mapsto \big(\gamma_1(t), \ldots, \gamma_N(t)\big) \tag{iv.13}$$

where $\gamma_1, \ldots, \gamma_N$ are geodesics of the one-dimensional manifold M.

The following proposition allows to characterize the parallel transport on this product manifold.

Proposition IV.2. *Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)$ be a geodesic of $\mathbb{M}$ and $t_0 \in \mathbb{R}$. Let $\mathbf{w} \in \mathrm{T}_{\boldsymbol{\gamma}(t_0)}\mathbb{M}$ be a tangent vector with $\mathbf{w} = (w_1, \ldots, w_N) \in \mathrm{T}_{\boldsymbol{\gamma}(t_0)}\mathbb{M}$. The parallel transport $\mathrm{P}_{\boldsymbol{\gamma},t_0,t}(\mathbf{w})$ is given by:*

$$\forall t \in \mathbb{R},\ \mathrm{P}_{\boldsymbol{\gamma},t_0,t}(\mathbf{w}) = \Big(\frac{w_1}{\dot\gamma_1(t_0)}\dot\gamma_1(t), \ldots, \frac{w_N}{\dot\gamma_N(t_0)}\dot\gamma_N(t)\Big). \tag{iv.14}$$

Proof. Since the Riemannian metric on $\mathbb{M}$ is the *product metric*, the computation the parallel transport boils down to the computation of $\mathrm{P}_{\gamma_i,t_0,t}(w_i)$). Indeed:

$$\mathrm{P}_{\boldsymbol{\gamma},t_0,t}(\mathbf{w}) = \Big(\mathrm{P}_{\gamma_1,t_0,t}(w_1), \ldots, \mathrm{P}_{\gamma_N,t_0,t}(w_N)\Big). \tag{iv.15}$$

Let $i \in \{1, \ldots, N\}$. The parallel transport $\mathrm{P}_{\gamma_i,t_0,t}(w_i)$ is computed as follows. As noted in the proof of Proposition IV.1, the Riemannian metric g of M is necessarily of the form $p \in M \mapsto g_p$ with: $\forall (u, v) \in \mathrm{T}_pM$, $g_p(u, v) = uvf(p)$ where $f : M \to]0, +\infty[$ is a smooth function.

It follows from the definition of parallel transport along the curve $t \mapsto \gamma_i$ that: $\forall t$, $\mathrm{P}_{\gamma_i,t_0,t}(w_i) \in \mathrm{T}_{\gamma_i(t)}M$. Since, for all t, $\mathrm{T}_{\gamma_i(t)}M$ is a one-dimensional vector space, the tangent vector $\dot\gamma_i(t) \neq 0$ spans this space. As a consequence, there exist a smooth

function $\xi_i : \mathbb{R} \to \mathbb{R}$ such that: $\forall t \in \mathbb{R}$, $\mathrm{P}_{\gamma_i,t_0,t}(w_i) = \xi_i(t)\dot{\gamma}_i(t)$. Because the parallel transport is an isometry and because γ_i is a geodesic, we have:

$$\forall t \in \mathbb{R},\ g_{\gamma_i(t)}\big(\mathrm{P}_{\gamma_i,t_0,t}(w_i), \dot{\gamma}_i(t)\big) = g_{\gamma_i(t_0)}(w_i, \dot{\gamma}_i(t_0)). \tag{iv.16}$$

The bilinearity of $g_{\gamma_i(t)}$ gives:

$$\begin{aligned} g_{\gamma_i(t)}\big(\mathrm{P}_{\gamma_i,t_0,t}(w_i), \dot{\gamma}_i(t)\big) &= g_{\gamma_i(t)}\big(\xi_i(t)\dot{\gamma}_i(t), \dot{\gamma}_i(t)\big) \\ &= \xi_i(t) g_{\gamma_i(t)}\big(\dot{\gamma}_i(t), \dot{\gamma}_i(t)\big). \end{aligned} \tag{iv.17}$$

Using that $\dot{\gamma}_i$ is parallel along γ_i, we have:

$$\forall t \in \mathbb{R},\ g_{\gamma_i(t)}\big(\dot{\gamma}_i(t), \dot{\gamma}_i(t)\big) = g_{\gamma_i(t_0)}\big(\dot{\gamma}_i(t_0), \dot{\gamma}_i(t_0)\big). \tag{iv.18}$$

As a consequence, Eq. [iv.16], Eq. [iv.17] and Eq. [iv.18] give:

$$\forall t \in \mathbb{R},\ \xi_i(t) g_{\gamma_i(t_0)}\big(\dot{\gamma}_i(t_0), \dot{\gamma}_i(t_0)\big) = g_{\gamma_i(t_0)}\big(w_i, \dot{\gamma}_i(t_0)\big). \tag{iv.19}$$

Using the form of the metric on M, Eq. [iv.19] writes:

$$\forall t \in \mathbb{R},\ \xi_i(t)\big(\dot{\gamma}_i(t_0)\big)^2 f\big(\gamma_i(t_0)\big) = w_i \dot{\gamma}_i(t_0) f\big(\gamma_i(t_0)\big). \tag{iv.20}$$

This last equation gives: $\forall t,\ \xi_i(t) = w_i/\dot{\gamma_i}(t_0)$. Finally,

$$\forall i \in \{1, \ldots, N\},\ \forall t \in \mathbb{R},\ \mathrm{P}_{\gamma_i,t_0,t}(w_i) = \frac{w_i}{\dot{\gamma}_i(t_0)} \dot{\gamma}_i(t). \tag{iv.21}$$

This last equation completes the proof of the proposition. □

IV.1.3 The 2-sphere

The 2-sphere $\mathbb{S}^2 = \{\mathbf{x} \in \mathbb{R}^3,\ \|\mathbf{x}\|_2 = 1\} \subset \mathbb{R}^3$ is a smooth 2-dimensional submanifold of $\mathbb{R}^3$. For each $\mathbf{p} \in \mathbb{S}^2$, $\mathrm{T}_{\mathbf{p}}\mathbb{S}^2 = \{\mathbf{p}\}^\perp$. If $\mathbb{S}^2$ is equipped with the induced metric from $\mathbb{R}^3$, $\mathbb{S}^2$ is a geodesically complete Riemannian manifold, as proved by the following proposition.

Proposition IV.3. *The geodesics of the sphere $\mathbb{S}^2$ are of the form:*

$$t \in \mathbb{R} \mapsto \cos(t\|\mathbf{v}\|)\mathbf{p} + \frac{\mathbf{v}}{\|\mathbf{v}\|} \sin(t\|v\|) \tag{iv.22}$$

where $\mathbf{p} \in \mathbb{S}^2$ and $\mathbf{v} \in \mathrm{T}_{\mathbf{p}}\mathbb{S}^2 = \{\mathbf{p}\}^\perp$.

Proof. Let $\mathbf{p} \in \mathbb{S}^2$ and $\mathbf{v} \in \mathrm{T}_{\mathbf{p}}\mathbb{S}^2 = \{\mathbf{p}\}^\perp$. Let $\mathcal{V} = \mathrm{Span}\big(\mathbf{p}, \frac{\mathbf{v}}{\|\mathbf{v}\|}\big)$ and R be the reflection with respect to $\mathcal{V}$, *i.e.* $R = \mathrm{Id}$ on $\mathcal{V}$ and $R = -\mathrm{Id}$ on $\mathcal{V}^\perp$. It follows from Pythagoras' theorem that $R : \mathbb{S}^2 \to \mathbb{S}^2$ is an isometry of $\mathbb{S}^2$. Let $\boldsymbol{\gamma}$ denote the geodesic of $\mathbb{S}^2$ such

that $\boldsymbol{\gamma}(0) = \mathbf{p}$ and $\dot{\boldsymbol{\gamma}}(0) = \mathbf{v}$. Because R is an isometry, $R \circ \boldsymbol{\gamma}$ is a geodesic of $\mathbb{S}^2$. By unicity of a geodesic given its starting point and initial velocity, $R \circ \boldsymbol{\gamma} = \boldsymbol{\gamma}$. Hence, $\boldsymbol{\gamma}$ is fixed under the reflection R. It follows that there exist real-valued functions α and β such that:

$$\forall t \in \mathbb{R},\ \boldsymbol{\gamma}(t) = \alpha(t)\mathbf{p} + \beta(t)\frac{\mathbf{v}}{\|\mathbf{v}\|}. \qquad \text{[iv.23]}$$

One can easily see that: $\forall t \in \mathbb{R}$, $\alpha(t) = \cos(t\|\mathbf{v}\|)$ and $\beta(t) = \sin(t\|\mathbf{v}\|)$. Finally, the geodesics of $\mathbb{S}^2$ are of the desired form. □

A direct consequence of this proposition is that if $\mathbf{p}_0 \in \mathbf{S}^2$, $t_0 \in \mathbb{R}$ and $\mathbf{v}_0 \in \mathrm{T}_{\mathbf{p}_0}\mathbb{S}^2$, the geodesic $\boldsymbol{\gamma}_0(\cdot) = \mathrm{Exp}_{\mathbf{p}_0,t_0}(\mathbf{v}_0)(\cdot)$ is given by:

$$\forall t \in \mathbb{R},\ \boldsymbol{\gamma}_0(t) = \cos\big((t-t_0)\|\mathbf{v}_0\|\big)\mathbf{p}_0 + \frac{\mathbf{v}_0}{\|\mathbf{v}_0\|}\sin\big((t-t_0)\|\mathbf{v}_0\|\big). \qquad \text{[iv.24]}$$

The geodesics of $\mathbb{S}^2$ are *great circles*, *i.e.* the intersection of the sphere with a plane going through the origin of the sphere.

Let $\boldsymbol{\gamma}$ denote a geodesic of $\mathbb{S}^2$ such that $\boldsymbol{\gamma}(0) = \mathbf{p}$ and $\dot{\boldsymbol{\gamma}}(0) = \mathbf{v}$. Parallel transport on the 2-sphere is given by the following proposition.

Proposition IV.4. *Let* $\mathbf{w} \in \mathrm{T}_{\mathbf{p}}\mathbb{S}^2$. *Let* $\mathbf{e}_1(t) = \dfrac{\dot{\boldsymbol{\gamma}}(t)}{\|\mathbf{v}\|}$ *and* $\mathbf{e}_2(t) = \boldsymbol{\gamma}(t) \wedge \mathbf{e}_1(t)$. *Then:*

$$\forall t \in \mathbb{R},\ \mathrm{P}_{\boldsymbol{\gamma},0,t}(\mathbf{w}) = \frac{\mathbf{w}^\top\mathbf{v}}{\|\mathbf{v}\|}\,\mathbf{e}_1(t) + \varepsilon\sqrt{\|\mathbf{w}\|^2 - \frac{(\mathbf{w}^\top\mathbf{v})^2}{\|\mathbf{v}\|^2}}\,\mathbf{e}_2(t) \qquad \text{[iv.25]}$$

where $\varepsilon = \mathrm{sign}\big(\mathbf{w}^\top(\mathbf{p}\wedge\mathbf{v})\big)$.

Proof. For all $t \in \mathbb{R}$, the set $\{\mathbf{e}_1(t), \mathbf{e}_2(t)\}$ forms an orthonormal basis of $\mathrm{T}_{\boldsymbol{\gamma}(t)}\mathbb{S}^2$. By definition of the parallel transport (along $\boldsymbol{\gamma}$), there exist continuous real-valued functions α and β such that:

$$\forall t \in \mathbb{R},\ \mathrm{P}_{\boldsymbol{\gamma},0,t}(\mathbf{w}) = \alpha(t)\mathbf{e}_1(t) + \beta(t)\mathbf{e}_2(t). \qquad \text{[iv.26]}$$

Since the parallel transport is an isometry, $\big(\mathrm{P}_{\boldsymbol{\gamma},0,t}(\mathbf{w})\big)^\top\dot{\boldsymbol{\gamma}}(t) = \mathbf{w}^\top\mathbf{v}$ for all t. Similarly, $\|\mathrm{P}_{\boldsymbol{\gamma},0,t}(\mathbf{w})\|^2 = \|\mathbf{w}\|^2$ for all t. These two conditions give:

$$\forall t \in \mathbb{R},\ \alpha(t) = \frac{\mathbf{w}^\top\mathbf{v}}{\|\mathbf{v}\|} \text{ and } \beta(t) = \varepsilon\sqrt{\|\mathbf{w}\|^2 - \frac{(\mathbf{w}^\top\mathbf{v})^2}{\|\mathbf{v}\|^2}} \qquad \text{[iv.27]}$$

with $\varepsilon = \mathrm{sign}\big(\mathbf{w}^\top(\mathbf{p}\wedge\mathbf{v})\big)$. □

IV.1.4 The space $\mathrm{Spd}(n)$ of symmetric positive definite matrices

If $\mathbf{\Sigma} \in \mathrm{SDP}(3)$, the tangent space $\mathrm{T}_{\mathbf{\Sigma}}\mathbb{M}$ at $\mathbf{\Sigma}$ can be identified with $\mathrm{Sym}(3)$. With the **affine-invariant metric**, this tangent space is equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathbf{\Sigma}}$ defined by:

$$\forall (\mathbf{W}_1, \mathbf{W}_2) \in \mathrm{T}_{\mathbf{\Sigma}}\mathbb{M},\ \langle \mathbf{W}_1, \mathbf{W}_2 \rangle_{\mathbf{\Sigma}} = \mathrm{tr}\big(\mathbf{\Sigma}^{-1/2}\mathbf{W}_1^{\top}\mathbf{\Sigma}^{-1}\mathbf{W}_2\mathbf{\Sigma}^{-1/2}\big). \qquad \text{[iv.28]}$$

This Riemannian metric was first introduced by Siegel in [Siegel, 1964], in the context of symplectic geometry, and has been used ever since in several contributions ([Förstner and Moonen, 2003, Pennec et al., 2006, Lenglet et al., 2006, Dryden et al., 2009, Su et al., 2011]). In [Lenglet et al., 2006] and [Moakher and Zéraï, 2011], the authors study the manifold $\mathrm{Spd}(n)$ equipped with the affine-invariant metric. In this section, using Eq. [iv.29] given in [Lenglet et al., 2006], we derive closed-form expression of the geodesics and parallel transport for the affine-invariant metric.

First, some notations need to be introduced. Let $m = n(n+1)/2$ denote the dimension of the linear space $\mathrm{Sym}(n)$, $(\mathbf{E}_i)_{1 \leq i \leq m}$ be the canonical basis of $\mathrm{Sym}(n)$ and $(\mathbf{E}_i^*)_{1 \leq i \leq m}$ its *dual basis*. The matrices are indexed by a single index, which corresponds to an enumeration of the pairs of integers $\{(k,l),\ 1 \leq k,l \leq n,\ k \leq l\}$. A matrix $\mathbf{V}$ in $\mathrm{Sym}(n)$ will be identified to the vector $(v_1, \ldots, v_m)$ of its coefficients from the upper triangular part. Using the expression of the Christoffel symbols in terms of the canonical basis of $\mathrm{Sym}(n)$ and its dual basis, Lenglet and collaborators prove that if $t \mapsto \mathbf{\Sigma}(t) = (\sigma_1(t), \ldots, \sigma_m(t))$ is a smooth curve in $\mathrm{Spd}(n)$ and $t \mapsto \mathbf{V}(t) = (v_1(t), \ldots, v_m(t))$ a vector field along Σ, the *covariant derivative* of $\mathbf{V}$ along $\mathbf{\Sigma}$ is given by the expression:

$$\frac{D\mathbf{V}(t)}{dt} = \sum_i^m \frac{dv_i(t)}{dt}\mathbf{E}_i + \sum_{i,j=1}^m v_i(t)\frac{d\sigma_j(t)}{dt}\nabla_{\mathbf{E}_i}\mathbf{E}_j. \qquad \text{[iv.29]}$$

Applying $\mathbf{E}_k^*$ $(1 \leq k \leq m)$ to Eq. [iv.29] together with the expression of the Christoffel symbols (see [Lenglet et al., 2006], Equations (3) and (4)), one gets that the vector field $\mathbf{V}$ is parallel along the curve $\mathbf{\Sigma}$ if and only if:

$$\frac{d\mathbf{V}(t)}{dt} - \frac{1}{2}\mathbf{V}(t)\mathbf{\Sigma}(t)^{-1}\frac{d\mathbf{\Sigma}(t)}{dt} - \frac{1}{2}\frac{d\mathbf{\Sigma}(t)}{dt}\mathbf{\Sigma}(t)^{-1}\mathbf{V}(t) = 0. \qquad \text{[iv.30]}$$

Lenglet and collaborators also note that one can obtain the geodesics of $\mathrm{Spd}(n)$ by solving Eq. [iv.30] with $\mathbf{V}(t) = d/dt\mathbf{\Sigma}(t)$. One can also obtain the expression of the geodesics by noting that the geodesic starting at I_n with velocity $\mathbf{V} \in \mathrm{Sym}(n)$ is given by $\exp(t\mathbf{V})$ and use the invariance of the affine-invariant metric under congruent transformations. For $\mathbf{P}_0 \in \mathrm{Spd}(n)$, $t_0 \in \mathbb{R}$ and $\mathbf{V}_0 \in \mathrm{T}_{\mathbf{P}_0}\mathrm{Spd}(n) \simeq \mathrm{Sym}(n)$, the geodesic $\boldsymbol{\gamma}_0(t) = \mathrm{Exp}_{\mathbf{P}_0,t_0}(\mathbf{V}_0)(t)$ is given by:

$$\forall t \in \mathbb{R},\ \boldsymbol{\gamma}_0(t) = \mathbf{P}_0^{1/2}\exp\big(t\mathbf{P}_0^{-1/2}\mathbf{V}_0\mathbf{P}_0^{-1/2}\big)\mathbf{P}_0^{1/2}. \qquad \text{[iv.31]}$$

where $\mathbf{P}_0^{1/2}$ (respectively $\mathbf{P}_0^{-1/2}$) denotes the unique symmetric positive definite square root of $\mathbf{P}_0$ (respectively its inverse). The expression of the parallel transport is given by the following proposition.

Proposition IV.5. *Let* $\mathbf{P}_0 \in \mathrm{Spd}(n)$, $t_0 \in \mathbb{R}$ *and* $\mathbf{V}_0 \in \mathrm{T}_{\mathbf{P}_0}\mathrm{Spd}(n) \simeq \mathrm{Sym}(n)$. *Let* $\boldsymbol{\gamma}_0$ *be the geodesic defined as above. If* $\mathbf{W}$ *is a tangent vector in* $\mathrm{T}_{\mathbf{P}_0}\mathrm{Spd}(n)$, *the parallel transport* $\mathrm{P}_{\boldsymbol{\gamma}_0,t_0,t}(\mathbf{W})$ *is given by:*

$$\forall t \in \mathbb{R},\ \mathrm{P}_{\boldsymbol{\gamma}_0,t_0,t}(\mathbf{W}) = \exp\Big(\frac{t-t_0}{2}\mathbf{V}_0\mathbf{P}_0^{-1}\Big)\mathbf{W}\exp\Big(\frac{t-t_0}{2}\mathbf{P}_0^{-1}\mathbf{V}_0\Big). \tag{iv.32}$$

Proof. With the definition Eq. [iv.31] of $\boldsymbol{\gamma}_0$, one can easily see that:

$$\forall t \in \mathbb{R},\ \frac{d\boldsymbol{\gamma}_0(t)}{dt}\boldsymbol{\gamma}_0^{-1}(t) = \mathbf{V}_0\mathbf{P}_0^{-1}. \tag{iv.33}$$

It follows that Eq. [iv.30] is equivalent to:

$$\frac{d\mathbf{V}(t)}{dt} = \frac{1}{2}\mathbf{V}(t)\mathbf{P}_0^{-1}\mathbf{V}_0 - \frac{1}{2}\mathbf{V}_0\mathbf{P}_0^{-1}\mathbf{V}(t). \tag{iv.34}$$

Eq. [iv.34] is a *differential Lyapunov equation*. It can be solved by considering a matrix-valued function of the form $t \mapsto \exp(-t\mathbf{M}^\top)\mathbf{R}(t)\exp(-t\mathbf{M})$. With $M = -(1/2)\mathbf{P}_0^{-1}\mathbf{V}_0$, one has that the parallel transport in $\mathrm{Spd}(n)$ is given by:

$$\mathbf{V}(t) = \exp\Big(\frac{t-t_0}{2}\mathbf{V}_0\mathbf{P}_0^{-1}\Big)\mathbf{W}\exp\Big(\frac{t-t_0}{2}\mathbf{P}_0^{-1}\mathbf{V}_0\Big). \tag{iv.35}$$

□

Another Riemannian metric on $\mathrm{Spd}(n)$

The affine-invariant metric is not the only metric which can be considered on $\mathrm{Spd}(n)$. $\mathrm{Sym}(n) \simeq \mathbb{R}^{n(n+1)/2}$ can be considered as a smooth manifold equipped with the metric $g^{\mathrm{Sym}(n)}$ defined by: $\forall \mathbf{U},\mathbf{V} \in \mathrm{T}_{\mathbf{M}}\mathrm{Sym}(n)$, $g^{\mathrm{Sym}(n)}_{\mathbf{M}}(\mathbf{U},\mathbf{V}) = \mathrm{tr}\big(\mathbf{U}^\top\mathbf{V}\big)$. Using the fact that the matrix exponential $\exp : \mathrm{Sym}(n) \to \mathrm{Spd}(n)$ is a diffeomorphism, the push-forward, with exp of the metric on $\mathrm{Sym}(n)$ defines a Riemannian metric $\exp_* g^{\mathrm{Sym}(n)}$ on $\mathrm{Spd}(n)$ by:

$$\forall \mathbf{S} \in \mathrm{Spd}(n),\ \forall (\mathbf{U},\mathbf{V}) \in \mathrm{T}_{\mathbf{S}}\mathrm{Spd}(n),\ \big(\exp_* g^{\mathrm{Sym}(n)}\big)_{\mathbf{S}}(\mathbf{U},\mathbf{V}) = \mathrm{tr}\big((\mathrm{D}_{\mathbf{S}}\log\cdot\mathbf{U})^\top \mathrm{D}_{\mathbf{S}}\log\cdot\mathbf{V}\big).$$

This Riemannian metric is called the **Log-Euclidean metric** on $\mathrm{Spd}(n)$.

In [Arsigny et al., 2006], the authors consider the space $\mathrm{Spd}(n)$ equipped with the log-Euclidean metric. This metric provides the space of symmetric positive definite matrices with a structure of Riemannian manifold. Unlike with the affine-invariant metric, the space $\mathrm{Spd}(n)$ endowed with the Log-Euclidean metric is a *flat* Riemannian manifold, meaning that its sectional curvature is null everywhere. By contrast, the space $\mathrm{Spd}(n)$ equipped with the affine-invariant metric is a Riemannian manifold of non-positive curvature [Skovgaard, 1984, Moakher and Zéraï, 2011] with no cut-locus. Within the Log-Euclidean framework, the geodesics are of the form: $\exp(\mathbf{V}_1 + t\mathbf{V}_2)$ with $\mathbf{V}_1, \mathbf{V}_2 \in \mathrm{Sym}(n)$. As expected, the geodesics are the image of a straight line in $\mathrm{Sym}(n)$ by the matrix exponential map.

IV.2 The concept of "parallel variations" on a Riemannian manifold

IV.2.1 Definition and properties

This section introduces the notion of "parallel variations" of a curve on a Riemannian manifold $(\mathbb{M}, g^{\mathbb{M}})$. The notion of "variation of a differentiable curve" on a manifold is defined in [Do Carmo Valero, 1992], Chapter 9. This notions allows to define *neighbouring* curves to a given curve c. In the next section, this construction will be used to define individual trajectories. Let $(\mathbb{M}, g^{\mathbb{M}})$ denotes a *geodesically complete* Riemannian manifold equipped with its Levi-Civita connection $\nabla^{\mathbb{M}}$.

Definition IV.1. Let $c : I \subset \mathbb{R} \to \mathbb{M}$ a differentiable curve on $\mathbb{M}$, $t_0 \in I$ and $\mathbf{w} \in \mathrm{T}_{c(t_0)}\mathbb{M}$ a tangent vector to $\mathbb{M}$ at $c(t_0)$. A **parallel variation of c in the direction of $\mathbf{w}$** is a curve $\boldsymbol{\eta}^{\mathbf{w}}(c, \cdot) : I \to \mathbb{M}$ defined by:

$$\forall t \in I,\ \boldsymbol{\eta}^{\mathbf{w}}(c, t) = \mathrm{Exp}^{\mathbb{M}}_{c(t)}\big(\mathrm{P}_{c,t_0,t}(\mathbf{w})\big). \tag{iv.36}$$

This construction is illustrated in Fig. 3. Given $t \in I$, parallel transport carries the tangent vector $\mathbf{w}$ from $\mathrm{T}_{c(t_0)}\mathbb{M}$ to $\mathrm{T}_{c(t)}\mathbb{M}$ along the curve c. At the point $c(t)$, a new point on $\mathbb{M}$ is obtained by taking the Riemannian exponential of the tangent vector $\mathrm{P}_{c,t_0,t}(\mathbf{w})$. This new point is denoted by $\boldsymbol{\eta}^{\mathbf{w}}(c, t)$. As t varies, one describes a curve $\boldsymbol{\eta}^{\mathbf{w}}(c, \cdot)$ on $\mathbb{M}$, which can be understood as a "parallel" to the curve c.

Remark. If c is a geodesic, $\boldsymbol{\eta}^{\mathbf{w}}(c, \cdot)$ is not, in general, a geodesic. The case of the 2-sphere $\mathbf{S}^2 \subset \mathbb{R}^3$ provides a counter-example. Indeed, the geodesics of $\mathbf{S}^2$ are the *great circles* (the intersection of the sphere with a plane which passes through the origin of $\mathbb{R}^3$). In general, a parallel variation of a great circle is not a great circle.

A coordinate system

A notable property of this construction is that it defines a coordinate chart called **Fermi coordinates** (or **Fermi charts**) [Michor, 2008]. As above, let $c : t_0 \in I \subset \mathbb{R} \to \mathbb{M}$ be a differentiable curve on $\mathbb{M}$ with no self intersection. Consider the mapping $\boldsymbol{\eta}$ defined by

$$\boldsymbol{\eta}\colon \begin{cases} I \times \{\dot{c}(t_0)\}^{\perp} \longrightarrow M \\ \quad (t, \mathbf{w}) \qquad \longmapsto \boldsymbol{\eta}^{\mathbf{w}}(c, t) = \mathrm{Exp}_{c(t)}\big(\mathrm{P}_{c,t_0,t}(\mathbf{w})\big). \end{cases} \tag{iv.37}$$

where $\{\dot{c}(t_0)\}^{\perp} = \{\mathbf{x} \in \mathrm{T}_{c(t_0)}\mathbb{M},\ g^{\mathbb{M}}_{c(t_0)}(\mathbf{x}, \dot{c}(t_0)) = 0\}$ is the orthogonal complement of $\mathrm{Span}\big(\dot{c}(t_0)\big)$ in $\mathrm{T}_{c(t_0)}\mathbb{M}$. Let $t \in I$. The tangent map of $\boldsymbol{\eta}$ at $(t, \mathbf{0})$ is given by:

$$T_{(t,\mathbf{0})}\boldsymbol{\eta} : (s, \mathbf{y}) \in \mathbb{R} \times \{\dot{c}(t_0)\}^{\perp} \mapsto s\dot{c}(t) + \mathrm{P}_{c,t_0,t}(\mathbf{y}). \tag{iv.38}$$

Since this map is a linear isomorphism, the inverse functions theorem ensures that there exist an open neighborhood V of $I \times \{\mathbf{0}\}$ such that $\boldsymbol{\eta}$ is a diffeomorphism of V onto its image. The pair $(V, \boldsymbol{\eta}|_V)$ is called **Fermi chart along** c.

IV.2.2 Examples of "parallel variations"

Computing in closed-form a parallel variation of a geodesic $\boldsymbol{\gamma}$, in the direction of a tangent vector $\mathbf{w}$, is not always possible because the Riemannian exponential and parallel transport are not always available in closed-form. Below, examples of geodesically complete Riemannian manifolds for which parallel variations can be computed in closed-form are considered. These examples include the space of 3×3 symmetric positive definite matrices, discussed in IV.1.4. Other examples include the case of a product of one-dimensional geodesically complete Riemannian manifold, in Section IV.1.2, and the 2-sphere $\mathbf{S}^2 \subset \mathbb{R}^3$ in Section IV.1.3. The results presented will be used extensively in Chapter V.

IV.2.2.1 The 2-sphere

To illustrate the notion of parallel variations on the 2-sphere, let:

$$\mathbf{p}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \; t_0 = 0, \; \mathbf{v}_0 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \text{ and } \mathbf{w} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \qquad \text{[iv.39]}$$

Note that the vector $\mathbf{w}$ belongs to $\mathrm{T}_{\mathbf{p}_0}\mathbb{S}^2$ and is orthogonal to $\mathbf{v}_0$. Figure 4 gives two examples of parallel variations of the geodesic $\boldsymbol{\gamma}_0(\cdot) = \mathrm{Exp}_{\mathbf{p}_0,t_0}(\mathbf{v}_0)(\cdot)$ in the direction of the tangent vectors $0.5\mathbf{w}$ and $\mathbf{w}$. One can observe that these parallel variations of a geodesic are no longer geodesics (the red curve is not a great circle of $\mathbb{S}^2$).

IV.2.2.2 Products of one-dimensional manifolds

In Section IV.1.2, closed-form expression for the geodesics and the parallel transport in a product of one-dimensional manifolds are given. This section on examples of parallel variations in such a product manifold starts with a proposition which gives a closed-form expression of a parallel variation. Then, the result of this proposition is illustrated with examples in $\mathbb{R}^N$ and $]0,1[^2$.

Proposition IV.6. *Let $\boldsymbol{\gamma}$ be a geodesic of $\mathbb{M}$ and $t_0 \in \mathbb{R}$. If $\eta^{\mathbf{w}}(\boldsymbol{\gamma}, \cdot)$ denotes a parallel variation of $\boldsymbol{\gamma}$ in the direction of $\mathbf{w}$, with $\mathbf{w} = (w_1, \ldots, w_N) \in \mathrm{T}_{\boldsymbol{\gamma}(t_0)}\mathbb{M}$ and $\boldsymbol{\gamma}(t) = (\gamma_1(t), \ldots, \gamma_N(t))$, we have:*

$$\forall s \in \mathbb{R}, \; \eta^{\mathbf{w}}(\boldsymbol{\gamma}, s) = \left(\gamma_1\Big(\frac{w_1}{\dot{\gamma}(t_0)} + s\Big), \ldots, \gamma_N\Big(\frac{w_N}{\dot{\gamma}(t_0)} + s\Big) \right). \qquad \text{[iv.40]}$$

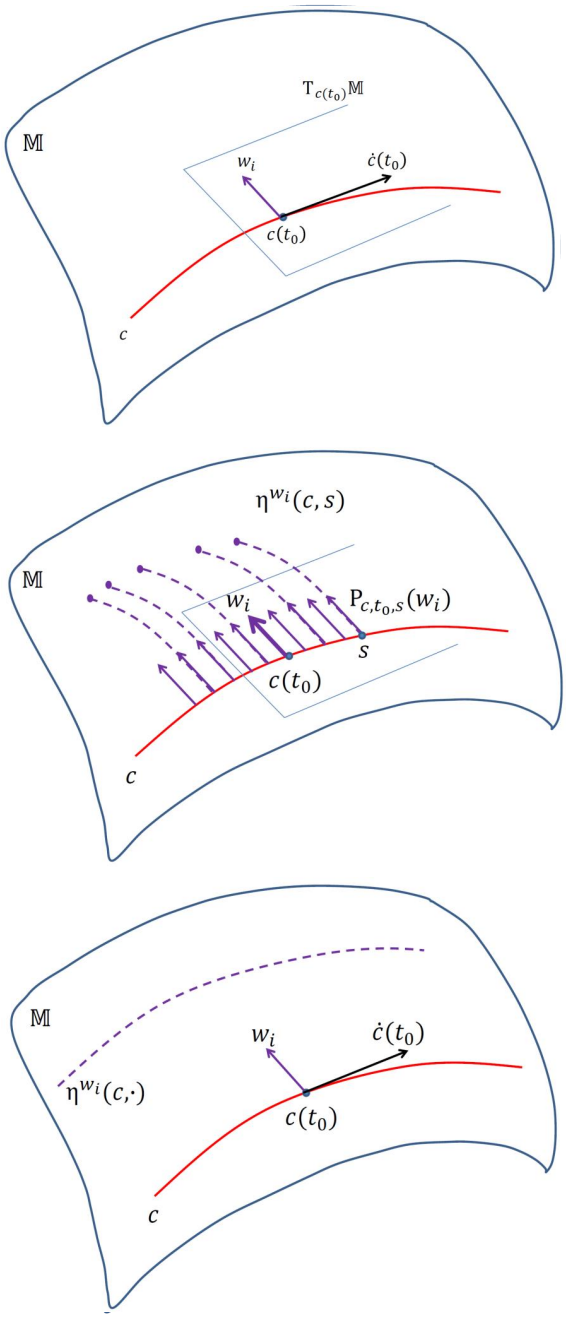


Figure 3 – Model description on a schematic manifold. **Top**: a non-zero vector $\mathbf{w}_i$ is chosen in $\mathrm{T}_{c(t_0)}\mathbb{M}$. **Middle**: the tangent vector $\mathbf{w}_i$ is transported continuously along the curve c. Then, a point $\eta^{\mathbf{w}_i}(c,s)$ is constructed at time s by use of the Riemannian exponential. **Bottom**: The curve $\eta^{\mathbf{w}_i}(c,\cdot)$ is the parallel resulting from the construction.

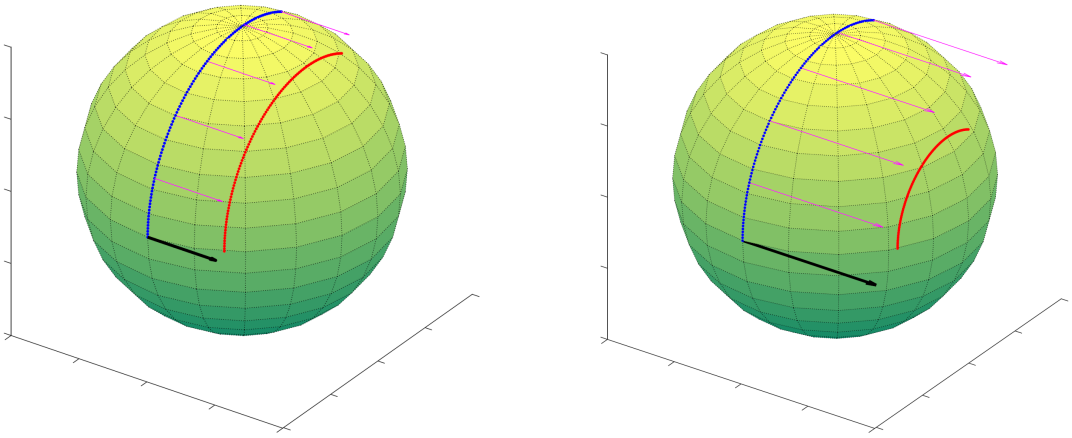


Figure 4 – For both, the black arrow represents the initial tangent vector to be transported. The magenta arrow represent the parallel transport of the black arrow at different time points. The geodesic $\boldsymbol{\gamma}_0$ is plotted in blue, whereas its parallel variation in the direction of the tangent vector is in red. **Left**: a parallel variation of $\boldsymbol{\gamma}_0(\cdot)$ in the direction of the tangent vector $0.5\mathbf{w}$. **Right**: a parallel variation of $\boldsymbol{\gamma}_0(\cdot)$ in the direction of the tangent vector $\mathbf{w}$.

Proof. The result of Proposition IV.2 writes:

$$\mathrm{P}_{\boldsymbol{\gamma},t_0,t}(\mathbf{w}) = \Big(\frac{w_1}{\dot{\gamma}_1(t_0)}\dot{\gamma}_1(t), \ldots, \frac{w_N}{\dot{\gamma}_N(t_0)}\dot{\gamma}_N(t)\Big). \tag{iv.41}$$

Since the Riemannian metric on $\mathbb{M}$ is the *product metric*, the computation of the parallel variation boils down to computing $\mathrm{Exp}_{\gamma_i(t)}\big(\mathrm{P}_{\gamma_i,t_0,t}(w_i)\big)$ with $t \in \mathbb{R}$ fixed. Indeed, for any point $\mathbf{p} = (p_{0,1}, \ldots, p_{0,N}) \in \mathbb{M}$ and any tangent vector $\mathbf{v} = (v_{0,1}, \ldots, v_{0,N}) \in \mathrm{T}_{\mathbf{p}}\mathbb{M}$:

$$\mathrm{Exp}_{\mathbf{p}}(\mathbf{v}) = \Big(\mathrm{Exp}_{p_{0,1}}(v_1), \ldots, \mathrm{Exp}_{p_{0,N}}(v_{0,N}). \tag{iv.42}$$

Introduce the curves

$$c \,:\, s \in [0,1] \;\mapsto\; \mathrm{Exp}_{\gamma_i(t)}\big(s\mathrm{P}_{\gamma_i,t_0,t}(w_i)\big)$$

and

$$\widetilde{c}\,: s \in [0,1] \mapsto \gamma_i\Big(t + s\frac{w_i}{\dot{\gamma}_i(t_0)}\Big)$$

Both curves c and $\widetilde{c}$ are geodesics of M which satisfy to: $c(0) = \widetilde{c}(0) = \gamma_i(t)$ and $\dot{c}(0) = \dot{\widetilde{c}}(0) = \frac{w_i}{\dot{\gamma}_i(t_0)}\dot{\gamma}_i(t)$. By unicity, the two curves are equal. As a consequence, for all $i \in \{1, \ldots, N\}$ and all $t \in \mathbb{R}$,

$$\mathrm{Exp}_{\gamma_i(t)}\big(\mathrm{P}_{\gamma_i,t_0,t}(w_i)\big) = \gamma_i\Big(t + \frac{w_i}{\dot{\gamma}_i(t_0)}\Big) \tag{iv.43}$$

which completes the proof. □

The examples below illustrate the notion of parallel variations in a product of one-dimensional manifolds.

Example 15. When $M = \mathbb{R}$, the geodesic equation (see Eq. [iii.6]) is easily solved and one obtains that the geodesics of the real line are straight lines of the form $t \in \mathbb{R} \mapsto p + tv$, where $p, v \in \mathbb{R}^2$. In particular, if $p_0, t_0, v_0 \in \mathbb{R}$, the geodesic $\gamma_0(\cdot) = \mathrm{Exp}_{p_0,t_0}(v_0)(\cdot)$ is the mapping $t \in \mathbb{R} \mapsto p_0 + v_0(t - t_0)$. As a consequence, a geodesic of $\mathbb{M} = \mathbb{R}^N$ is simply a straight line in $\mathbb{R}^N$ and a parallel variation of this straight line (in the direction of an orthogonal tangent vector $\mathbf{w}$) is the translation of the straight line by the vector $\mathbf{w}$.

Example 16. As discussed above, the geodesics of the Riemannian manifold $M =]0, 1[$, equipped with the Riemannian metric defined in Eq. [iv.5], are logistic curves. To illustrate the notion of parallel variations, consider the product manifold $\mathbb{M} = M^2 =]0, 1[^2$. Let $\mathbf{p}_0 = (0.7, 0.3)$, $t_0 = 50$, $\mathbf{v}_0 = (0.06, 0.03)$. The vector $\mathbf{w} = \big(-v_{0,2}G(p_{0,2}), v_{0,1}G(p_{0,1})\big) \simeq (-0.6803, 1.3605)$ is orthogonal to $\mathbf{v}_0$ for the product metric. In Figure 5, two examples of parallel variations of $\boldsymbol{\gamma}_0$ in the directions $\mathbf{w}_1 = 0.5\mathbf{w}$ and $\mathbf{w}_2 = -0.3\mathbf{w}$ are represented.

IV.2.2.3 The space Spd(3)

To illustrate the notion of parallel variation in the space of 3×3 symmetric positive definite matrices, consider the following matrices:

$$\mathbf{P}_0 = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix} \in \mathrm{Spd}(3), \ \mathbf{V}_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.05 \end{bmatrix} \in \mathrm{Sym}(3). \qquad \text{[iv.44]}$$

With $t_0 = 40$, the parameters $(\mathbf{P}_0, t_0, \mathbf{V}_0)$ define the geodesic $\boldsymbol{\gamma}_0(\cdot) = \mathrm{Exp}_{\mathbf{P}_0,t_0}(\mathbf{V}_0)(\cdot)$ defined in Eq. [iv.31]. This geodesic goes through the sphere $10\mathrm{I}_3$ at time $t_0 = 40$ and progresses to an elongated ellipsoid at t increases. The matrix $\mathbf{W}$ defined by:

$$\mathbf{W} = \begin{bmatrix} -4.8717 & 7.0711 & 7.0711 \\ 7.0711 & 8.7445 & 7.0711 \\ 7.0711 & 7.0711 & 9.9900 \end{bmatrix} \qquad \text{[iv.45]}$$

is orthogonal to $\mathbf{V}_0$ for the inner product on $\mathrm{T}_{\mathbf{P}_0}\mathrm{Spd}(3)$. In Figure 7, examples of parallel variations of $\boldsymbol{\gamma}_0$ in the direction of $\pm 0.8\mathbf{W}$ are given. In this figure, each matrix is colored according to its *fractional anisotropy* (FA), a measure in $[0, 1]$ which characterizes the anisotropy of a symmetric positive definite matrix. Let $\mathbf{M} \in \mathrm{Spd}(3)$ with eigenvalues $\lambda_1, \lambda_2, \lambda_3$. The FA of $\mathbb{M}$ is defined by:

$$\mathrm{FA}(\mathbf{M}) = \sqrt{\frac{1}{2}} \frac{\sqrt{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}. \qquad \text{[iv.46]}$$

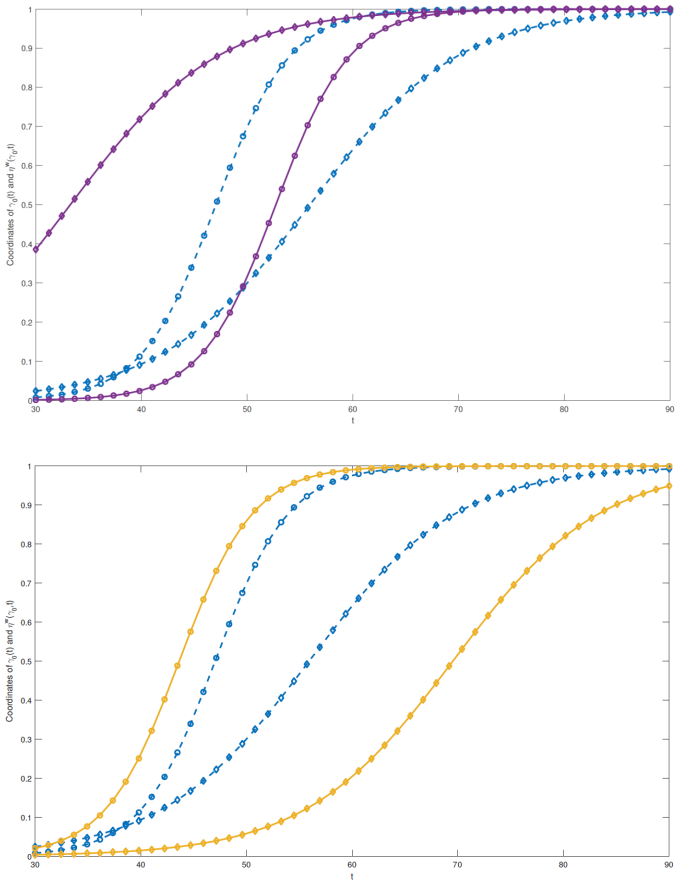


Figure 5 – **Top**: plot of the two coordinates of $\boldsymbol{\gamma}_0(\cdot)$ in dotted line and $\boldsymbol{\eta}^{\mathbf{w}_1}(\boldsymbol{\gamma}_0, \cdot)$ in solid line, with $\mathbf{w}_1 = 0.5\mathbf{w}$. **Bottom**: plot of the two coordinates of $\boldsymbol{\gamma}_0(\cdot)$ in dotted line and $\boldsymbol{\eta}^{\mathbf{w}_2}(\boldsymbol{\gamma}_0, \cdot)$ in solid line, with $\mathbf{w}_2 = -0.3$. For both figures, the line with the round (respectively square) marker corresponds to the first (respectively second) coordinate of the curve.

For $\mathbf{M} \in \mathrm{Spd}(3)$, if $\mathrm{FA}(\mathbf{M}) = 0$, then the ellipsoid defined by $\mathbf{M}$ is a sphere. On the contrary, if $\mathrm{FA}(\mathbf{M})$ is close to 1, the ellipsoid defined by $\mathbf{M}$ is elongated in a direction.

IV.2.3 Discussion

In the sections above, the Riemannian exponential and parallel transport can be computed in closed-form. However, such closed-form expressions cannot be obtained for every geodesically complete Riemannian manifold. When the Riemannian exponential or parallel transport cannot be written explicitly, one can use numerical schemes to approximate these quantities. As mentioned in Section III.1.5, the Riemannian expo-

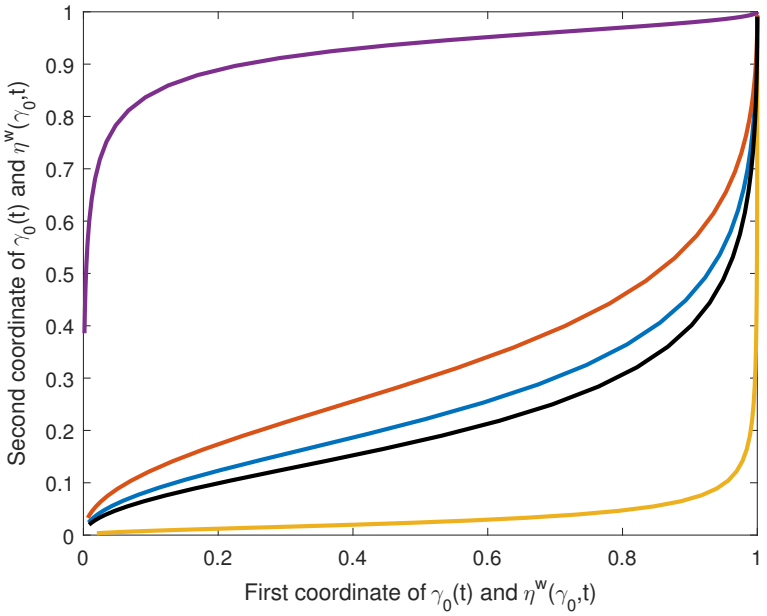


Figure 6 – Plots of the first coordinate of $\boldsymbol{\eta}^{\mathbf{w}_i}(\boldsymbol{\gamma}_0, \cdot)$ against its second coordinate. Blue: $\mathbf{w}_i = 0$. Violet: $\mathbf{w}_i = 0.5\mathbf{w}$. Red: $\mathbf{w}_i = 0.05\mathbf{w}$. Black: $\mathbf{w}_i = -0.03\mathbf{w}$. Yellow: $\mathbf{w}_i = -0.3\mathbf{w}$.

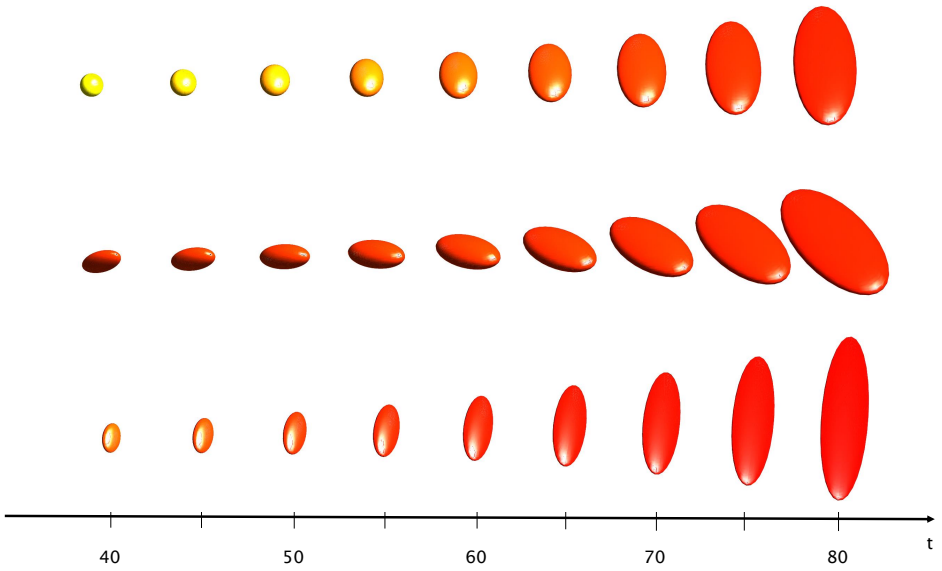


Figure 7 – First row: samples along the geodesic $\boldsymbol{\gamma}_{\mathbf{P}_0,t_0,\mathbf{V}_0}(\cdot)$. Second (respectively third) row: samples along the parallel variation $\boldsymbol{\eta}^{\mathbf{W}_i}(\boldsymbol{\gamma}_0, \cdot)$ with $\mathbf{W}_i = 0.8\mathbf{W}$ (respectively $\mathbf{W}_i = -0.8\mathbf{W}$.) The tangent vector $\mathbf{W}$ used here is given in Eq. [iv.45].

nential can be obtained by solving a set of second-order nonlinear differential equations (see Eq. [iii.6]). Solving these differential equations can be done with a large variety of numerical schemes, such as *Runge-Kutta method* or *Heun's method*. Parallel transport can also be approximated using a numerical scheme called *Schild's ladder* (see [Lorenzi et al., 2011] and Chapter VI). Situations which require to numerically approximate the Riemannian exponential are not considered in this dissertation.

IV.3 A generic model for longitudinal manifold-valued data

In the following sections, $\mathbb{M}$ denotes a convex open subset of the Euclidean space $\mathbb{R}^N$ ($N \geq 1$). This N-dimensional smooth manifold is equipped with a Riemannian metric $g^{\mathbb{M}}$ for which it is geodesically complete.

The observed data consists in repeated measurements for a group of p individuals. Let $i \in \{1, \ldots, p\}$. The number of observations for the ith individual is denoted by $k_i \in \mathbb{N} \setminus \{0, 1\}$. These k_i observations were obtained at the time points $t_{i,1} < \ldots < t_{i,k_i}$. The number of time points can vary from one subject to another. Let $j \in \{1, \ldots, k_i\}$. The jth observation of the ith individual is denoted by $\mathbf{y}_{i,j}$.

IV.3.1 Hierarchical structure and spatiotemporal transformations

The *generic spatiotemporal* model is a nonlinear mixed-effects model. As emphasized in the introduction of this dissertation, mixed-effects models include fixed and random effects. The *fixed-effects* are parameters which are shared by all the individuals and allow to describe the model at the population level. *Random effects* are individual-specific random variable which describe the model at the individual level. These two types of effects provide the model with a hierarchical structure. The generic spatiotemporal model is constructed as follows. To begin with, a group-average trajectory $\boldsymbol{\gamma}_0$ is defined on the manifold $\mathbb{M}$. Given the average trajectory, subject-specific trajectories are obtained by spatiotemporal transformations, which consist in *parallel variations* of the average trajectory $\boldsymbol{\gamma}$ and *time reparametrization*. The data points $\mathbf{y}_{i,j}$ are seen as samples along these individual trajectories. If $\boldsymbol{\gamma}_i$ denotes the trajectory of the ith individual, the model writes: $\mathbf{y}_{i,j} = \boldsymbol{\gamma}_i(t_{i,j}) + \varepsilon_{i,j}$, where $\varepsilon_{i,j}$ is a Gaussian noise. The observation $\mathbf{y}_{i,j}$ is therefore considered as a small perturbation of a quantity which lies in a Riemannian manifold. This hierarchical modeling is summarized in Figure 8.

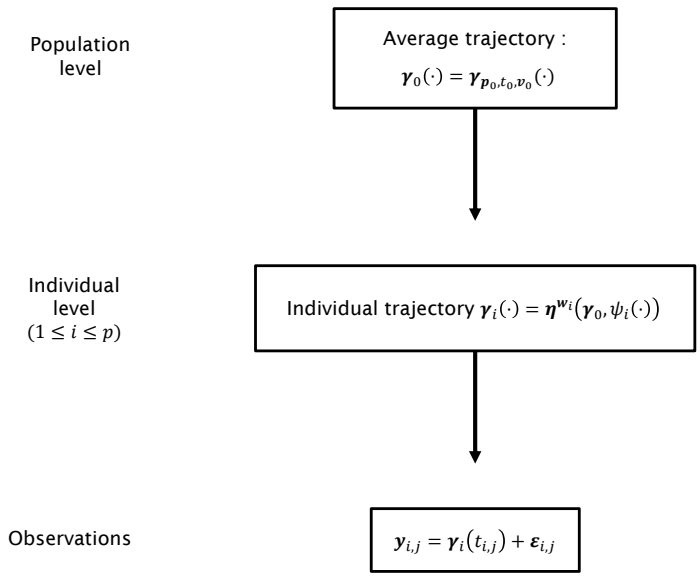


Figure 8 – The hierarchical structure of the *generic spatiotemporal model*. The fixed effects define an average trajectory on $\mathbb{M}$. This trajectory is used together with spatial and temporal transformations to define individual-specific trajectories. The observations $(\mathbf{y}_{i,j})_{1\leq j\leq k_i}$ of the ith individual are seen as random perturbations of points along its specific trajectory.

The group-average trajectory $\boldsymbol{\gamma}_0$ is chosen to be the geodesic

$$\boldsymbol{\gamma}_0 : \begin{cases} \mathbb{R} \longrightarrow \mathbb{M} \\ t \longmapsto \boldsymbol{\gamma}_0(t) = \mathrm{Exp}_{\mathbf{p}_0,t_0}(\mathbf{v})(t) \end{cases} \qquad \text{[iv.47]}$$

which goes through the point $\mathbf{p}_0 \in \mathbb{M}$ at time t_0 and with velocity $\mathbf{v}_0 \in T_{\mathbf{p}_0}\mathbb{M}$. Recall that $\mathrm{Exp}_{\mathbf{p},t_0}(\mathbf{v})(\cdot)$ denotes the Riemannian exponential (see Chapter III) at $\mathbf{p} \in \mathbb{M}$ and with initial velocity $\mathbf{v} \in \mathrm{T}_{\mathbf{p}}\mathbb{M}$. The parameters $\mathbf{p}_0$, t_0 and $\mathbf{v}_0$ are fixed-effects of the model.

Let $i \in \{1, \ldots, p\}$ denote the ith individual. The subject-specific trajectory $\boldsymbol{\gamma}_i$ is defined in two steps. The first step consists in constructing the curve $\eta^{\mathbf{w}_i}(\boldsymbol{\gamma}_0, \cdot)$, which is a *parallel variation* the average trajectory $\boldsymbol{\gamma}_0 = \boldsymbol{\gamma}_{\mathbf{p}_0,t_0,\mathbf{v}_0}$ in the direction of a tangent vector $\mathbf{w}_i \in T_{\mathbf{p}_0}\mathbb{M}$. This tangent vector is chosen orthogonal, for the inner product $g^{\mathbb{M}}_{\mathbf{p}_0}$, to $\dot{\boldsymbol{\gamma}}_0(t_0) = \mathbf{v}_0$. The tangent vectors $(\mathbf{w}_i)_{1\leq i\leq p}$ are random effects of the model, called **space shifts**. The orthogonality condition on the space shifts is discussed in Section IV.3.3.1.

The second step consists in *reparametrizing in time* the parallel variation $\eta^{\mathbf{w}_i}(\boldsymbol{\gamma}, \cdot)$. We consider a subject-specific affine mapping ψ_i of the form $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$, where $\alpha_i > 0$ and $\tau_i \in \mathbb{R}$ are random effects of our model. The trajectory $\boldsymbol{\gamma}_i$ of the ith

individual is

$$\boldsymbol{\gamma}_i : \begin{cases} \mathbb{R} \longrightarrow \mathbb{M} \\ t \longmapsto \boldsymbol{\gamma}_i(t) = \boldsymbol{\eta}^{\mathbf{w}_i}(\boldsymbol{\gamma}_0, \psi_i(t)). \end{cases} \quad \text{[iv.48]}$$

The function ψ_i is called **time reparametrization** and the random effects α_i (respectively τ_i) are called **acceleration factor** (respectively **time shift**). The form of the individual time reparametrization is discussed in Section V.1.1.

In the following, for all $i \in \{1, \ldots, p\}$, the individual time reparametrization ψ_i is of the form $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$ and the acceleration factors are defined by: $\alpha_i = \exp(\xi_i)$.

IV.3.2 Parallel variation and time reparametrization commute

In the previous section, the individual trajectories are constructed as reparametrized parallel variations of the average trajectory $\boldsymbol{\gamma}_0$. Recall that a parallel variation of $\boldsymbol{\gamma}_0$ (in the direction of a space shift $\mathbf{w}_i$) is constructed in two steps. *First*, the "parallel" $\boldsymbol{\eta}^{\mathbf{w}_i}(\boldsymbol{\gamma}_0, \cdot)$ is constructed, *then* it is reparametrized using the affine time reparametrization ψ_i. In this section, we show that individual trajectories could be constructed by considering a parallel of the geodesic $\boldsymbol{\gamma}_0 \circ \psi_i$. In other words, parallel variations and time reparametrization commute, in the following sense:

Proposition IV.7. *For $\mathbf{w}_i \in \mathrm{T}_{\boldsymbol{\gamma}_0(t_0)}\mathbb{M}$, we have:*

$$\forall t \in \mathbb{R},\ \boldsymbol{\eta}^{\mathbf{w}_i}(\boldsymbol{\gamma}_0, \psi_i(t)) = \boldsymbol{\eta}^{\widetilde{\mathbf{w}_i}}(\boldsymbol{\gamma}_0 \circ \psi_i, t) \quad \textit{with} \quad \widetilde{\mathbf{w}_i} = \mathrm{P}_{\boldsymbol{\gamma}_0, t_0, \psi_i(t_0)}(\mathbf{w}_i). \quad \text{[iv.49]}$$

In order to prove this proposition, the following lemma is needed.

Lemma IV.1. *Let $c : I \subset \mathbb{R} \to \mathbb{M}$ be a differentiable curve and $\psi : [a, b] \to I$ $(a < b)$ an affine function. Let $\mathbf{v} \in \mathrm{T}_{(c\circ\psi)(a)}\mathbb{M}$. Then:*

$$\forall s \in [a, b],\ \mathrm{P}_{c\circ\psi, a, b}(\mathbf{v}) = \mathrm{P}_{c, \psi(a), \psi(b)}(\mathbf{v}). \quad \text{[iv.50]}$$

This lemma is a consequence of elementary properties of the covariant derivative along a curve. Note that the result of this lemma holds because ψ is *affine*. If ψ is a polynomial of degree greater or equal to 2, the result no longer holds. The proof of the proposition writes as follows.

Proof. Let $\widetilde{\mathbf{w}_i} = \mathrm{P}_{\boldsymbol{\gamma}_0, t_0, \psi_i(t_0)}(\mathbf{w}_i)$. By definition of a parallel variation of $\boldsymbol{\gamma}_0 \circ \psi_i$ in the direction of $\widetilde{\mathbf{w}}_i$, we have:

$$\forall s \in \mathbb{R},\ \boldsymbol{\eta}^{\widetilde{\mathbf{w}}_i}(\boldsymbol{\gamma}_0 \circ \psi_i, s) = \mathrm{Exp}_{(\boldsymbol{\gamma}_0 \circ \psi_i)(s)}\big(\mathrm{P}_{\boldsymbol{\gamma}_0 \circ \psi_i, t_0, s}(\widetilde{\mathbf{w}}_i)\big). \quad \text{[iv.51]}$$

But, $\mathrm{P}_{\boldsymbol{\gamma}_0\circ\psi_i,t_0,s}(\widetilde{\mathbf{w}}_i) = \mathrm{P}_{\boldsymbol{\gamma}_0\circ\psi_i,t_0,s}\big(\mathrm{P}_{\boldsymbol{\gamma}_0,t_0,\psi_i(t_0)}(\mathbf{w}_i)\big)$. The previous lemma allows to rewrite this last equality as:

$$\begin{aligned}\mathrm{P}_{\boldsymbol{\gamma}_0\circ\psi_i,t_0,s}\big(\mathrm{P}_{\boldsymbol{\gamma}_0,t_0,\psi_i(t_0)}(\mathbf{w}_i)\big) &= \mathrm{P}_{\boldsymbol{\gamma}_0,\psi_i(t_0),\psi_i(s)}\big(\mathrm{P}_{\boldsymbol{\gamma}_0,t_0,\psi_i(t_0)}(\mathbf{w}_i)\big)\\ &= \mathrm{P}_{\boldsymbol{\gamma}_0,t_0,\psi_i(s)}(\mathbf{w}_i).\end{aligned} \qquad \text{[iv.52]}$$

Finally,

$$\begin{aligned}\forall s\in\mathbb{R},\ \boldsymbol{\eta}^{\widetilde{\mathbf{w}}_i}(\boldsymbol{\gamma}_0\circ\psi_i,s) &= \mathrm{Exp}_{(\boldsymbol{\gamma}_0\circ\psi_i)(s)}\big(\mathrm{P}_{\boldsymbol{\gamma}_0,t_0,\psi_i(s)}(\mathbf{w}_i)\big)\\ &= \boldsymbol{\eta}^{\mathbf{w}_i}\big(\boldsymbol{\gamma}_0,\psi_i(s)\big).\end{aligned} \qquad \text{[iv.53]}$$

□

Recall that the space shift $\mathbf{w}_i \in \mathrm{T}_{\boldsymbol{\gamma}_0(t_0)}\mathbb{M}$ are chosen, by definition, orthogonal to $\dot{\boldsymbol{\gamma}}_0(t_0)$ (for the inner product given by the metric $g^{\mathbb{M}}$). It follows that the transformed space shifts $\widetilde{\mathbf{w}_i}$ in Proposition IV.7 remain orthogonal to $\dot{\boldsymbol{\gamma}}_0\big(\psi(t_0)\big)$ because the parallel transport in an *isometry*.

IV.3.3 Definition of the space shifts and orthogonality condition

As emphasized above, the ith ($1 \leq i \leq p$) individual trajectory is defined to be the parallel variation of $\boldsymbol{\gamma}_0$ in the direction of the tangent vector $\mathbf{w}_i \in \mathrm{T}_{\mathbf{p}_0}\mathbb{M}$. The space shifts $(\mathbf{w}_i)_{1\leq i\leq p}$ are required to satisfy to the following orthogonality condition:

$$\forall i\in\{1,\ldots,p\},\ g^{\mathbb{M}}_{\mathbf{p}_0}(\mathbf{w}_i,\mathbf{v}_0) = 0. \qquad \text{[iv.54]}$$

This section discusses different methods which allow to include this orthogonality condition on the space shifts into a statistical model. The methodological challenge raised by this section consists in defining a (nonlinear) mixed-effects model with smooth constraints on some of the random effect of the model. Chapter VI discusses the use of a stochastic algorithm to estimate the parameters of the generic model. The impact of the methods discussed below on the algorithm and its performance are discussed in Section VI.4.4.

In order to ensure the interpretability of the space shifts, we consider an Independent Component Analysis (ICA) [Hyvärinen et al., 2004] decomposition of each tangent vector $\mathbf{w}_i$ as a linear combination of $N_s < N$ statistically independent tangent vectors $(\mathbf{A}_l)_{1\leq l\leq N_s}$ which are called **independent components** or **independent directions**. As a consequence, the space shifts $(\mathbf{w}_i)_{1\leq i\leq p}$ are defined as follows:

$$\forall i\in\{1,\ldots,p\},\ \mathbf{w}_i = \mathbf{A}\mathbf{s}_i = \sum_{l=1}^{N_s} s_{l,i}\mathbf{A}_l \qquad \text{[iv.55]}$$

where $\mathbf{A} = (\mathbf{A}_l)_{1\leq l\leq N_s}$ is such that each $\mathbf{A}_i$ is a vector in $\mathrm{T}_{\dot{\boldsymbol{\gamma}}_0(t_0)}\mathbb{M}$. In the definition Eq. [iv.55], the weights $(s_{l,i})_{1\leq l\leq N_s}$ are random effects of the model called **sources**. By defining the space shifts this way, the generic spatiotemporal model will estimate an ICA decomposition of the space shifts. However, this definition does not ensure the orthogonality of the space shifts. A possible solution to make the vectors $\mathbf{w}_i$ orthogonal to $\mathbf{v}_0 = \dot{\boldsymbol{\gamma}}_0(t_0)$ consists in decomposing each vector in an orthonormal basis of $\mathrm{T}_{\mathbf{p}_0}\mathbb{M}$.

Moreover, it is important to note that the choice of the form of the distribution of the space-shifts does not depend on the reference time-point t_0. Indeed, the $\mathbf{w}_i = \mathbf{A}\mathbf{s}_i$ are defined in the tangent space of the curve at point $\mathbf{p}_0 = \boldsymbol{\gamma}_0(t_0)$. At another point $\mathbf{p}_0' = \boldsymbol{\gamma}_0(t_0')$, space-shifts become $\mathbf{w}_i' = \mathrm{P}_{\boldsymbol{\gamma}_0,t_0,t_0'}\mathbf{w}_i$, where $\mathrm{P}_{\boldsymbol{\gamma}_0,t_0,t_0'}$ is an orthogonal matrix. They are therefore distributed according to $\mathbf{w}_i' = P_{\boldsymbol{\gamma}_0,t_0,t}\mathbf{A}\mathbf{s}_i$: the distribution of the sources $\mathbf{s}_i$ does not change and the independent components (i.e. the columns of $\mathbf{A}$) are adjusted to the new position on the average trajectory. In particular, the variance of the $\mathbf{w}_i'$ is invariant. This property holds for isometric invariant distributions. For instance, if $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, then $\mathbf{w}_i' \sim \mathcal{N}(\mathbf{0}, \mathrm{P}_{\boldsymbol{\gamma}_0,t_0,t_0'}\,\boldsymbol{\Sigma}\,\mathrm{P}^{\top}_{\boldsymbol{\gamma}_0,t_0,t_0'})$.

IV.3.3.1 Construction of an orthonormal basis

Since $\mathbb{M}$ is a N-dimensional Riemannian manifold, the tangent space $\mathrm{T}_{\mathbf{p}_0}\mathbb{M}$ is a N-dimensional vector space and the subspace $\mathrm{Span}\big(\dot{\boldsymbol{\gamma}}_0(t_0)\big)^{\perp}$ is a $(N-1)$-dimensional subspace of $\mathrm{T}_{\mathbf{p}_0}\mathbb{M}$. Let $(\mathcal{B}_k)_{1\leq k\leq N-1}$ denote an orthonormal basis of $\mathrm{Span}\big(\dot{\boldsymbol{\gamma}}_0(t_0)\big)^{\perp}$ and define:

$$\forall l \in \{1,\ldots,N_s\},\ \mathbf{A}_l = \sum_{k=1}^{N-1} \beta_{l,k}\mathcal{B}_k. \qquad \text{[iv.56]}$$

By definition, each independent component $\mathbf{A}_l$ $(1 \leq l \leq N_s)$ satisfies to: $g^{\mathbb{M}}_{\mathbf{p}_0}\big(\mathbf{A}_l, \dot{\boldsymbol{\gamma}}_0(t_0)\big) = 0$, which ensures, by linearity, that the orthogonality condition on the space shifts holds.

Different methods to compute the orthonormal basis $(\mathcal{B}_k)_{1\leq k\leq N-1}$ are reviewed below. These methods exploit the form of the Riemannian metric $g^{\mathbb{M}}$ on $\mathbb{M}$. As a matter of fact, since $\mathbb{M}$ is assumed to be a connected open subset of $\mathbb{R}^N$, for each $\mathbf{p}_0 \in \mathbb{M}$, the tangent space $\mathrm{T}_{\mathbf{p}_0}\mathbb{M}$ can be identified with $\mathbb{R}^N$ itself. It follows that the metric $g^{\mathbb{M}}$ is necessarily of the form

$$\forall \mathbf{p}_0 \in \mathbb{M},\ \forall (\mathbf{u},\mathbf{v}) \in \mathrm{T}_{\mathbf{p}_0}\mathbb{M},\ g^{\mathbb{M}}_{\mathbf{p}_0}(\mathbf{u},\mathbf{v}) = \langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{p}_0} = \mathbf{u}^{\top}\mathbf{G}(\mathbf{p}_0)\mathbf{v}. \qquad \text{[iv.57]}$$

where $\mathbf{p} \in \mathbb{M} \mapsto \mathbf{G}(\mathbf{p})$ is a smooth mapping from $\mathbb{M}$ to $\mathrm{Spd}(N)$. The orthogonality conditions Eq. [iv.54] write:

$$\forall l \in \{1,\ldots,N_s\},\ \langle \mathbf{A}_l, \dot{\boldsymbol{\gamma}}_0(t_0)\rangle_{\mathbf{p}_0} = \mathbf{A}_l^{\top}\mathbf{G}(\mathbf{p}_0)\dot{\boldsymbol{\gamma}}_0(t_0) = 0. \qquad \text{[iv.58]}$$

As a consequence, the problem of constructing an orthonormal basis of $\mathrm{Span}(\dot{\boldsymbol{\gamma}}_0(t_0))^{\perp}$ (for the inner product $\langle\cdot,\cdot\rangle_{\mathbf{p}_0}$) is equivalent to constructing an orthonormal basis of

$\mathrm{Span}\big(\mathbf{G}(\mathbf{p}_0)\dot{\boldsymbol{\gamma}}_0(t_0)\big)^{\perp}$ (for the canonical inner product on $\mathbb{R}^N$). In this dissertation, the Householder method and the Gram-Schmidt algorithm are considered for the construction of an orthonormal basis of $\mathrm{Span}\big(\mathbf{G}(\mathbf{p}_0)\dot{\boldsymbol{\gamma}}_0(t_0)\big)^{\perp}$. The computational cost of these algorithms is discussed in Chapter VI.

IV.3.3.1.1 The Householder method

Let $\mathbf{S}_0 = \mathbf{G}(\mathbf{p}_0)\dot{\boldsymbol{\gamma}}_0(t_0) \in \mathbb{R}^N$ and let $\mathbf{S}_{0,k}$ $(1 \leq k \leq N)$ denote the kth coordinate of $\mathbf{S}_0$. Introduce the vector $\mathbf{a}$ defined by:

$$\mathbf{a} = \mathbf{S}_0 + \mathrm{Sgn}(\mathbf{S}_{0,1})\|\mathbf{S}_0\|\mathbf{e}_1 \qquad \text{[iv.59]}$$

where $\mathbf{e}_1$ denotes the first vector of the canonical basis of $\mathbb{R}^N$ and $\mathrm{sgn}(\mathbf{S}_{0,1})$, the sign of $\mathbf{S}_{0,1}$. Let $\mathbf{Q}$ be the matrix defined by:

$$\mathbf{Q} = \mathrm{I}_N - 2\frac{\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top\mathbf{a}}. \qquad \text{[iv.60]}$$

The following result hold.

Proposition IV.8. *Let $\mathbf{Q}$ be the $N \times N$ matrix defined in Eq. [iv.60]. For $i \in \{1, \ldots, N\}$, let $\mathbf{Q}_i$ denote the ith column of $\mathbf{Q}$. Let $\widetilde{\mathbf{Q}} = \begin{pmatrix}\mathbf{Q}_2 & \ldots & \mathbf{Q}_N\end{pmatrix}$. Then:*

(i) $\mathbf{S}_0 \in \mathrm{Span}(\mathbf{Q}_1)$,

(ii) $\widetilde{\mathbf{Q}}^\top\widetilde{\mathbf{Q}} = \mathrm{I}_{N-1}$,

(iii) $\widetilde{\mathbf{Q}}^\top\mathbf{S}_0 = \mathbf{0}$.

Proof. Given the definition of $\mathbf{a}$, one can easily show that $\mathbf{a}^\top\mathbf{S}_0 = \|\mathbf{S}_0\|(\|\mathbf{S}_0\| + \mathrm{sgn}(\mathbf{S}_{0,1})\mathbf{S}_{0,1})$ and $\mathbf{a}^\top\mathbf{a} = 2\|\mathbf{S}_0\|(\|\mathbf{S}_0\| + \mathrm{sgn}(\mathbf{S}_{0,1})\mathbf{S}_{0,1})$. Therefore:

$$\mathbf{Q}\mathbf{S}_0 = \mathbf{S}_0 - 2\frac{\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top\mathbf{a}} = \mathbf{S}_0 - \mathbf{a} = -\mathrm{sgn}(\mathbf{S}_{0,1})\|\mathbf{S}_0\|\mathbf{e}_1. \qquad \text{[iv.61]}$$

Since $\mathbf{Q}^2 = \mathrm{I}_N$, Eq. [iv.61] yields: $\mathbf{S}_0 = -\mathrm{sgn}(\mathbf{S}_{0,1})\|\mathbf{S}_0\|\mathbf{Q}_1$. Therefore, $\mathbf{S}_0 \in \mathrm{Span}(\mathbf{Q}_1)$.

Let $(\mathbf{e}_1, \ldots, \mathbf{e}_N)$ denote the canonical basis of $\mathbb{R}^N$. Then, for all $(i, j) \in \{2, \ldots, N\}^2$, we have:

$$\mathbf{Q}_i^\top\mathbf{Q}_j = \mathbf{e}_i^\top\mathbf{e}_j - 4\mathbf{e}_i^\top\frac{\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top\mathbf{a}}\mathbf{e}_j + 4\mathbf{e}_i^\top\frac{\mathbf{a}\mathbf{a}^\top}{\mathbf{a}^\top\mathbf{a}}\mathbf{e}_j = \mathbf{e}_i^\top\mathbf{e}_j. \qquad \text{[iv.62]}$$

As a consequence, $\widetilde{\mathbf{Q}}^\top\widetilde{\mathbf{Q}} = \mathrm{I}_{N-1}$.

Finally, for $i \in \{2, \ldots, N\}$, we have:

$$\mathbf{Q}_i^\top\mathbf{S}_0 = \mathbf{S}_{0,i} - 2\frac{\mathbf{S}_{0,i}\|\mathbf{S}_0\|(\|\mathbf{S}_0\| + \mathrm{sgn}(\mathbf{S}_{0,1})\mathbf{S}_{0,1})}{2\|\mathbf{S}_0\|(\|\mathbf{S}_0\| + \mathrm{sgn}(\mathbf{S}_{0,1})\mathbf{S}_{0,1})} = 0. \qquad \text{[iv.63]}$$

Then, $\widetilde{\mathbf{Q}}^\top\mathbf{S}_0 = \mathbf{0}$. □

It follows that an orthonormal basis of $\mathrm{Span}(\dot{\boldsymbol{\gamma}}_0(t_0))^\perp$ is given by the columns of the matrix $\widetilde{\mathbf{Q}}$.

IV.3.3.1.2 The Gram-Schmidt algorithm

The following proposition is to be used with a basis of $\mathrm{Span}\big(\mathbf{G}(\mathbf{p}_0)\dot{\boldsymbol{\gamma}}_0(t_0)\big)^\perp$ computed beforehand.

Proposition IV.9. *Let $k \in \mathbb{N}^*$ and $(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ a set of linearly independent vectors in $\mathbb{R}^N$. There exist a unique orthonormal set of vectors $(\widetilde{\mathbf{v}_1}, \ldots, \widetilde{\mathbf{v}_k})$ in $\mathbb{R}^N$ such that:*

(i) $\forall j \in \{1, \ldots, k\}$, $\mathrm{Span}(\mathbf{v}_1, \ldots, \mathbf{v}_j) = \mathrm{Span}(\widetilde{\mathbf{v}_1}, \ldots, \widetilde{\mathbf{v}_j})$,

(ii) $\forall j \in \{1, \ldots, k\}$, $\mathbf{v}_j^\top \widetilde{\mathbf{v}_j} > 0$.

The vectors $(\widetilde{\mathbf{v}_1}, \ldots, \widetilde{\mathbf{v}_k})$ are given by:

$$\widetilde{\mathbf{v}_1} = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} \quad \textit{and} \quad \forall j \in \{2, \ldots, k\},\ \widetilde{\mathbf{v}_j} = \frac{\mathbf{v}_j - \sum_{l=1}^{j-1} (\mathbf{v}_j^\top \widetilde{\mathbf{v}_l}) \widetilde{\mathbf{v}_l}}{\|\mathbf{v}_j - \sum_{l=1}^{j-1} (\mathbf{v}_j^\top \widetilde{\mathbf{v}_l}) \widetilde{\mathbf{v}_l}\|}. \qquad \text{[iv.64]}$$

Proof. The existence of the set $(\widetilde{\mathbf{v}_1}, \ldots, \widetilde{\mathbf{v}_k})$ is done by induction on k. For $k = 1$, let $\widetilde{\mathbf{v}_1} = \mathbf{v}_1/\|\mathbf{v}_1\|$. Then, it is clear that $\mathrm{Span}(\widetilde{\mathbf{v}_1}) = \mathrm{Span}(\mathbf{v}_1)$ and $\|\widetilde{\mathbf{v}_1}\| = 1$. Let $s \in \mathbb{N}^*$. Assume that $(\widetilde{\mathbf{v}_1}, \ldots, \widetilde{\mathbf{v}_s})$ satisfy to the conditions *(i)* and *(ii)* of the proposition. Let:

$$\widetilde{\mathbf{v}}_{s+1} = \mathbf{v}_{s+1} - \sum_{j=1}^{s} (\mathbf{v}_{s+1}^\top \widetilde{\mathbf{v}_j}) \widetilde{\mathbf{v}_j}. \qquad \text{[iv.65]}$$

Then, the condition $\widetilde{\mathbf{v}}_{s+1} \in \mathrm{Span}(\widetilde{\mathbf{v}_1}, \ldots, \widetilde{\mathbf{v}_s})$ is satisfied since $\widetilde{\mathbf{v}}_{s+1}$ is defined as the orthogonal projection of $\mathbf{v}_{s+1}$ on $\mathrm{Span}(\widetilde{\mathbf{v}_1}, \ldots, \widetilde{\mathbf{v}_s})$. The set $(\widetilde{\mathbf{v}_1}, \ldots, \widetilde{\mathbf{v}}_{s+1})$ satisfies the conditions *(i)* and *(ii)* of the proposition.

The uniqueness of the set $(\widetilde{\mathbf{v}_1}, \ldots, \widetilde{\mathbf{v}_k})$ follows from the condition *(ii)* of the proposition. □

IV.3.4 Statistical model and probability distributions

The generic spatiotemporal model assumes that the jth observation of the ith individual at time $t_{i,j}$ derives from:

$$\mathbf{y}_{i,j} = \eta^{\mathbf{w}_i}(\boldsymbol{\gamma}_0, \psi_i(t_{i,j})) + \boldsymbol{\varepsilon}_{i,j}. \qquad \text{[iv.66]}$$

With the notations introduced above, let $\mathbf{z}_{\mathrm{pop}} = (\mathbf{p}_0, t_0, \mathbf{v}_0, (\beta_{l,k})_{l,k})$ denote the **population variables** and $(\mathbf{z}_i)_{1 \leq i \leq p}$ denote the set of **individual variables** with:

$\mathbf{z}_i = (\xi_i, \tau_i, (s_{l,i})_{l,i})$. Both $\mathbf{z}_{\text{pop}}$ and $(\mathbf{z}_i)_{1\leq i\leq p}$ are **latent** (or random) variables assumed independent of each other and distributed as follows:

$$\mathbf{p}_0 \sim \mathcal{N}(\overline{\mathbf{p}_0}, \sigma^2_{\mathbf{p}_0}), \quad t_0 \sim \mathcal{N}(\overline{t_0}, \sigma^2_{t_0}), \quad \mathbf{v}_0 \sim \mathcal{N}(\overline{v_0}, \sigma^2_{v_0}), \quad \beta_{l,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\overline{\beta}_{l,k}, \sigma^2_\beta) \qquad \text{[iv.67]}$$

and

$$\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_\xi), \quad \tau_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_\tau), \quad s_{l,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1). \qquad \text{[iv.68]}$$

where $\sigma^2_{\mathbf{p}_0}$, $\sigma^2_{t_0}$, $\sigma^2_{\mathbf{v}_0}$ and σ^2_β are fixed variance parameters. The noise variables $(\boldsymbol{\varepsilon}_{i,j})_{i,j}$ are assumed independent of the other random variables and identically distributed:

$$\boldsymbol{\varepsilon}_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2). \qquad \text{[iv.69]}$$

Let $\boldsymbol{\theta}_{\text{var}} = (\sigma^2_\xi, \sigma^2_\tau, \sigma^2)$ denote the variance parameters which are not fixed and $\boldsymbol{\theta} = \big(\overline{\mathbf{p}_0}, \overline{t_0}, \overline{\mathbf{v}_0}, (\overline{\beta}_{l,k}), \boldsymbol{\theta}_{\text{var}}\big)$ be the **parameters** of the model. The domain of $\boldsymbol{\theta}$ is denoted by Θ and defined by:

$$\begin{aligned}\Theta = \big\{\boldsymbol{\theta} = (\overline{\mathbf{p}_0}, \overline{\mathbf{v}_0}, \overline{t_0}, (\overline{\beta}_{l,k})_{l,k}, \boldsymbol{\theta}_{\text{var}}) \,\big/ (\overline{\mathbf{p}_0}, \overline{\mathbf{v}_0}) \in \mathbb{TM},\ \overline{t_0} \in \mathbb{R},\\ (\overline{\beta}_{l,k})_{l,k} \in \mathbb{R}^{(N-1)N_s},\ \boldsymbol{\theta}_{\text{var}} \in]0, +\infty[^3\big\}.\end{aligned} \qquad \text{[iv.70]}$$

The generic spatiotemporal model is described in a Bayesian framework. This means that a probability distribution q_{prior}, called **prior distribution**, is assumed for the parameters of the model and given by:

$$\begin{aligned} q_{\text{prior}}(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}) \propto{}& \exp\Big(-\frac{1}{2s^2_{\mathbf{p}_0}}\|\overline{\mathbf{p}_0} - \overline{\overline{\mathbf{p}_0}}\|^2\Big) \times \exp\Big(-\frac{1}{2s^2_{t_0}}(\overline{t_0} - \overline{\overline{t_0}})^2\Big)\\ &\times \exp\Big(-\frac{1}{2s^2_{\mathbf{v}_0}}\|\overline{\mathbf{v}_0} - \overline{\overline{\mathbf{v}_0}}\|^2\Big) \times \exp\Big(-\frac{1}{2s^2_{\boldsymbol{\beta}}}\|\overline{\boldsymbol{\beta}}\|^2\Big)\\ &\times \left(\frac{1}{\sqrt{\sigma^2_\xi}}\exp\Big(-\frac{\sigma^2_{\xi,0}}{2\sigma^2_\xi}\Big)\right)^{m_\xi} \times \left(\frac{1}{\sqrt{\sigma^2_\tau}}\exp\Big(-\frac{\sigma^2_{\tau,0}}{2\sigma^2_\tau}\Big)\right)^{m_\tau}\\ &\times \left(\frac{1}{\sqrt{\sigma^2}}\exp\Big(-\frac{\sigma^2_0}{2\sigma^2}\Big)\right)^{m_\sigma} \mathrm{d}\overline{\mathbf{p}_0}\,\mathrm{d}\overline{t_0}\,\mathrm{d}\overline{\mathbf{v}_0}\,d\overline{\boldsymbol{\beta}}\,\mathrm{d}\sigma^2_\xi\,\mathrm{d}\sigma^2_\tau\,\mathrm{d}\sigma^2 \end{aligned} \qquad \text{[iv.71]}$$

where m_ξ,m_τ and m_σ are *fixed* hyperparameters strictly greater than 2 and $\overline{\overline{\mathbf{p}_0}}, \overline{\overline{t_0}}, \overline{\overline{\mathbf{v}_0}}, s^2_{\mathbf{p}_0}, s^2_{t_0}, s^2_{\mathbf{v}_0}, s^2_{\boldsymbol{\beta}}, \sigma^2_{\xi,0}$, $\sigma^2_{\tau,0}$ and σ^2_0 *fixed* hyperparameters. The vector $\boldsymbol{\theta}_{\text{hyper}}$ denotes the vector of all the *fixed* hyperparameters used to define the prior density function in Eq. [iv.71]. The dependence between the different variables of the model is represented in Fig. 9. The construction of the generic spatiotemporal model is subject to several hypotheses which are discussed in the next section.

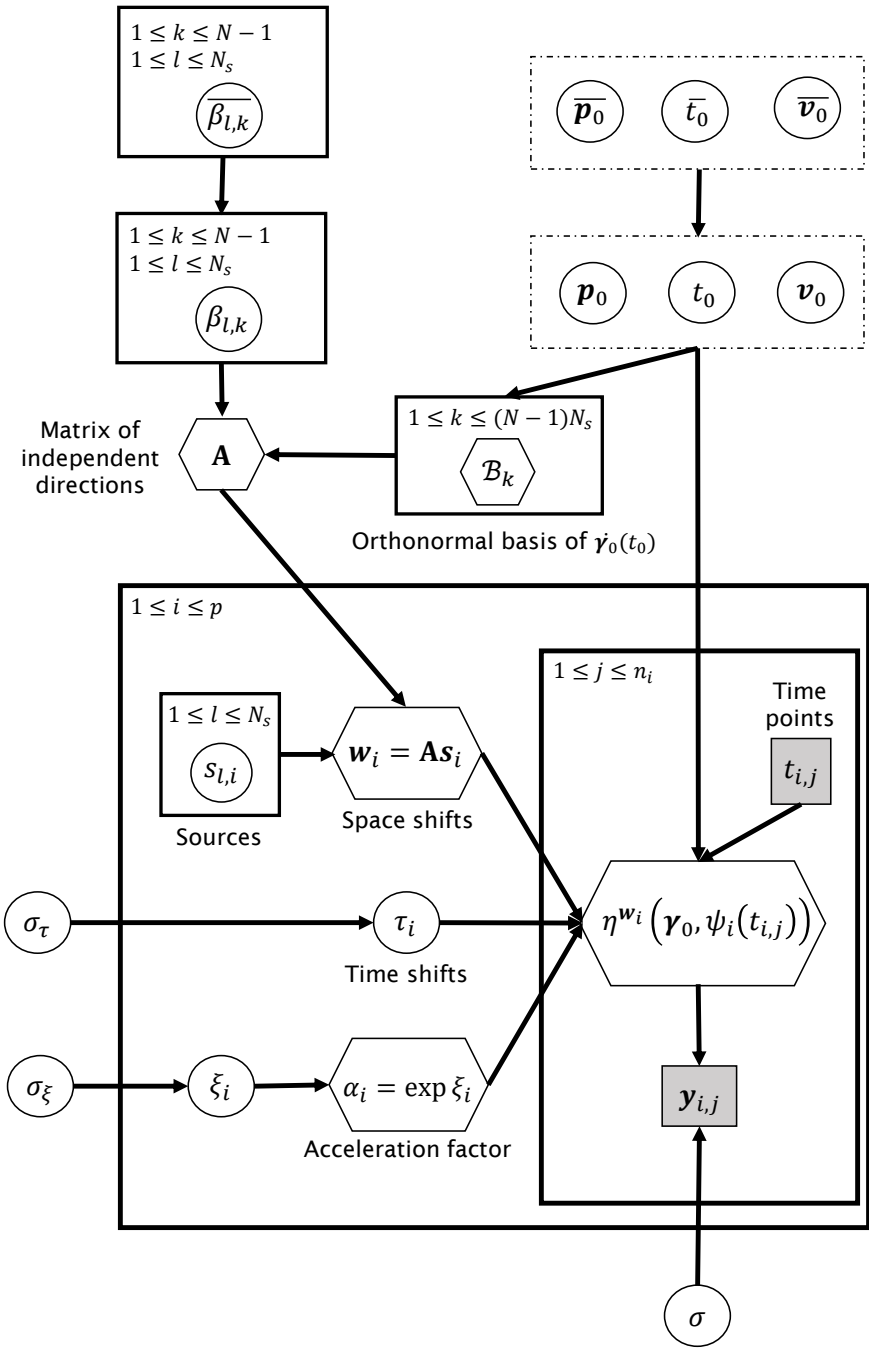


Figure 9 – Graphical representation of the generic spatiotemporal model. Round indicate latent variables of the model. Boxes with indexes in the upper left corner indicate a repetition. Shaded boxed indicates that the quantity is observed. Priors on $\left(\overline{\mathbf{p}_0}, \overline{t_0}, \overline{\mathbf{v}_0}, (\beta_{l,k})_{1 \leq l \leq N_s, 1 \leq k \leq N-1}, \sigma_\xi, \sigma_\tau, \sigma\right)$ are omitted for clarity.

IV.3.5 Discussion

IV.3.5.1 The noise model

The additive, or *extrinsic*, noise model in Eq. [iv.66] makes sense because we assumed that $\mathbb{M}$ is a subset of the Euclidean space $\mathbb{R}^N$. The term $\eta^{\mathbf{w}_i}(\boldsymbol{\gamma}, \psi_i(t_{i,j}))$ belongs to the manifold $\mathbb{M}$ while the noise term $\boldsymbol{\varepsilon}_{i,j}$ is added in the underlying Euclidean space. However, the noise model is not *intrinsic* in the sense that the noise term $\boldsymbol{\varepsilon}_{i,j}$ is not added on the manifold. In [Fletcher, 2011], the author have considered an intrinsic noise model which would write:

$$\mathbf{y}_{i,j} = \mathrm{Exp}_{\boldsymbol{\eta}^{\mathbf{w}_i}(\boldsymbol{\gamma}_0, \psi_i(t_{i,j}))}(\boldsymbol{\varepsilon}_{i,j}). \tag{iv.72}$$

This noise model allows to remain on the manifold. Still, obtaining maximum *a posteriori* estimates of the parameters with this intrinsic noise model is more difficult as the model likelihood might not be available in closed-form.

In addition to this, the Gaussian random variables $(\boldsymbol{\varepsilon}_{i,j})_{1\leq i\leq p,\, 1\leq j\leq k_i}$ are assumed *independent* of each other and *identically distributed*. This assumption may be too simplistic since it implies that for each individual, $(\mathbf{y}_{i,j})_{1\leq j\leq k_i}$ are not correlated. A more realistic noise model would consist in assuming that the residuals $(\boldsymbol{\varepsilon}_{i,j})_{1\leq j\leq k_i}$ are correlated but independent of $(\boldsymbol{\varepsilon}_{l,j})_{1\leq j\leq k_l}$ for $l \neq i$. In [Chi and Reinsel, 1989], the authors addressed this problem by considering a regression model with autocorrelated errors. However, they note that the correlation between measurements tends to decrease exponentially with the temporal distance between the measurements occasions.

IV.3.5.2 On the choice of probability distributions

IV.3.5.2.1 For the point $\mathbf{p}_0$

Note that the Gaussian prior on $\mathbf{p}_0$ does not take into account the fact that $\mathbf{p}_0$ is a point on a Riemannian manifold. Indeed, the prior on $\mathbf{p}_0$ is defined in the Euclidean space $\mathbb{R}^N$. In [Pennec et al., 2006], the author generalizes the multivariate Gaussian distribution to Riemannian manifolds. Following these ideas, an *intrinsic* prior distribution on $\mathbf{p}_0 \in \mathbb{M}$ would be given by:

$$p(\mathbf{p}_0) \propto \exp\Big(-\frac{1}{2\sigma^2_{\mathbf{p}_0}} d(\mathbf{p}_0, \overline{\mathbf{p}}_0)^2\Big) \tag{iv.73}$$

where $\overline{\mathbf{p}}_0 \in \mathbb{M}$ denotes the mean of the distribution, $\sigma^2_{\mathbf{p}_0}$ its variance and $d(\cdot,\cdot)$, the Riemannian distance function. For all $(\mathbf{p}, \mathbf{q}) \in \mathbb{M}$, the **Riemannian distance function** d is defined by: $d(p,q) = \|\mathrm{Log}_{\mathbf{p}}(\mathbf{q})\|_{\mathbf{p}}$ ($\mathrm{Log}_{\mathbf{p}}(\cdot)$ is defined in III.1.5). In addition to being intrinsic to the Riemannian manifold $\mathbb{M}$, this probability distribution reduces to the Gaussian distribution $\mathcal{N}(\overline{\mathbf{p}}_0, \sigma^2_{\mathbf{p}_0}\mathrm{I}_N)$ if $\mathbb{M}$ is the Euclidean space $\mathbb{R}^N$. However,

this *Riemannian Gaussian distribution* requires that the Riemannian manifold $\mathbb{M}$ be an *homogeneous space*. If $\mathbb{M}$ does not have this property, the normalizing constant of the probability distribution may depend on $\overline{\mathbf{p}}_0$, which makes *a posteriori* estimation of $\overline{\mathbf{p}}_0$ more difficult. Moreover, computing the log-probability density function of $\mathbf{p}_0$ requires to be able to compute the Riemannian logarithm. In longitudinal shape analysis, where the Riemannian manifold $\mathbb{M}$ is given by the group of diffeomorphisms, this would not be possible. Also note that the prior $\mathbf{p}_0 \sim \mathcal{N}(\overline{\mathbf{p}}_0, \sigma^2_{\mathbf{p}_0} \mathrm{I}_N)$ makes sense since we assume that the manifold $\mathbb{M}$ is an open subset of $\mathbb{R}^N$. As a consequence, if the mean $\overline{\mathbf{p}}_0$ is chosen on $\mathbb{M}$ with a "sufficiently small" variance $\sigma^2_{\mathbf{p}_0}$, $\mathbf{p}_0$ shall remain in $\mathbb{M}$.

IV.3.5.2.2 For the acceleration factors α_i

In Section IV.3, we assume that the acceleration factors $(\alpha_i)_{1 \leq i \leq p}$ follow a log-normal distribution. This choice of prior distribution aims at ensuring the positiveness of the acceleration factors. Indeed, an individual time reparametrization ψ_i with a negative acceleration factor would reverse time, which does not make sense for the generic spatiotemporal model. Note that other continuous prior distributions could have been considered, such as the exponential distribution.

Part V

Particular cases of the model

Summary

In this chapter, particular cases of the generic model are presented. These particular cases are obtained by describing the model for specific Riemannian manifolds. The models described below can be used to analyze univariate or multivariate normalized measurements, symmetric positive definite matrices or the temporal progression of a family of biological characteristics.

In the previous chapter, the model is described in a generic framework where $\mathbb{M}$ can be any open subset of the Euclidean space $\mathbb{R}^N$. By specifying a manifold $\mathbb{M}$ and a Riemannian metric on it, one specifies a model. Indeed, choosing a Riemannian metric on $\mathbb{M}$ defines the geodesics as well as the parallel transport. Several particular cases of the generic spatiotemporal model are considered. In Section V.1, two particular cases which allow to analyze longitudinal *univariate* observations are studies. These models provide a different insight on the form of the time reparametrization used above. Section V.3 introduces a model which allows to analyze the temporal progression of a family of biological characteristics. Finally, the generic spatiotemporal model is described for 3×3 symmetric positive definite matrices in Section V.2. The relations between the generic model and the ones presented hereafter is summarized in Figure 10.

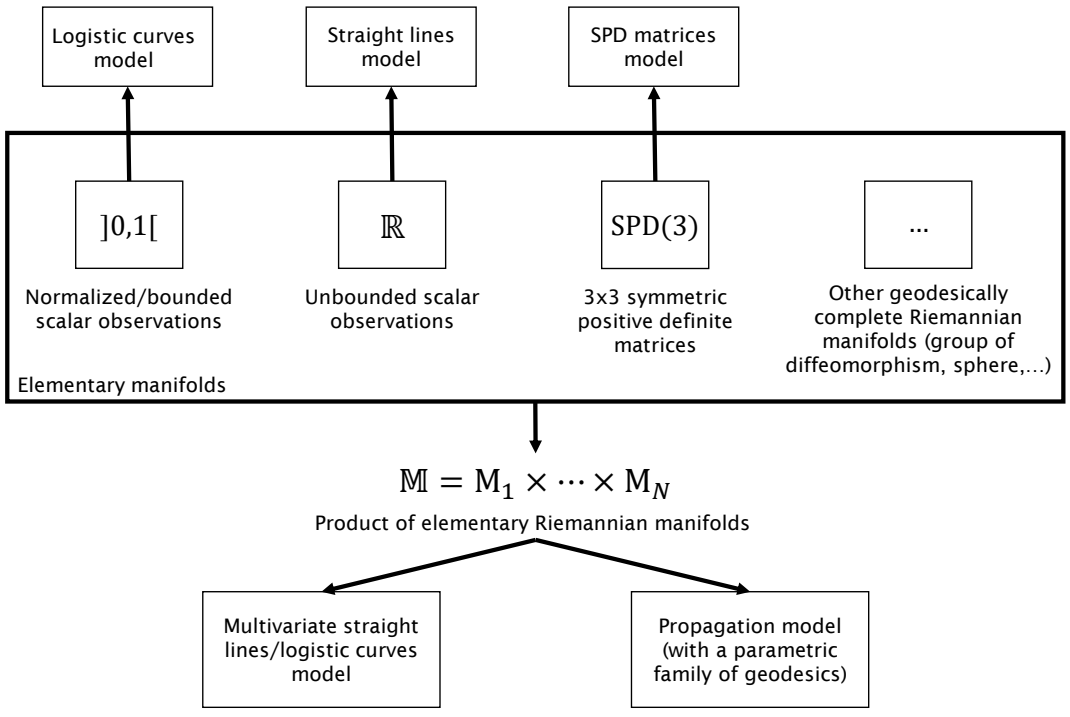


Figure 10 – Particular cases of the generic spatiotemporal model

V.1 One-dimensional geodesically complete Riemannian manifolds

Let M be an open interval of $\mathbb{R}$ equipped with a Riemannian metric g^M, for which it is geodesically complete. The case of one dimensional manifolds is particular because, for all $p_0 \in M$, $\mathrm{T}_p M \simeq \mathbb{R}$ and given $v_0 \in \mathrm{T}_{p_0} M$, there is only one tangent vector w at p_0 which is orthogonal (for the inner product $g^M_{p_0}$) to v_0 : $w = 0$. As a corollary of Proposition IV.6, if γ is a geodesic of M, $t_0 \in \mathbb{R}$ and $w = 0$, then for all $s \in \mathbb{R}$, $\eta^w(\gamma, s) = \gamma(s)$. Therefore, with the notations of Section IV.3, the generic spatiotemporal model Eq. [iv.66] writes:

$$y_{i,j} = \gamma \circ \psi_i(t_{i,j}) + \varepsilon_{i,j} \qquad [v.1]$$

with, for all $i \in \{1, \ldots, p\}$, $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$ and $\alpha_i = \exp(\xi_i)$. The prior on the parameters of this model and probability distribution of its latent variables are the same as in iv.71.

V.1.1 An alternative presentation of the generic model

The alternative presentation below provides a different insight on the role of the latent variables $(\alpha_i, \tau_i)_{1 \leq i \leq p}$. Let $p_0 \in M$, $t_0 \in \mathbb{R}$ and $v_0 \in \mathrm{T}_{p_0} M \simeq \mathbb{R}$. With the notations introduced in Section III.1.5, let γ_0 be the group-average trajectory defined as the geodesic $t \in \mathbb{R} \mapsto \mathrm{Exp}_{p_0,t_0}(v_0)(t)$. γ_0 is the geodesic of $\mathbb{M}$ which goes through the point p_0 at time t_0 and with velocity v_0. Let $1 \leq i \leq p$. The trajectory γ_i of the ith individual is defined as the geodesic $\gamma_i(t) = \mathrm{Exp}_{p_0,t_0+\tau_i}(\alpha_i v_0)(t)$. Hence, γ_i is the geodesic which goes through the point p_0 at time $t_0 + \tau_i$ and with velocity $\alpha_i v_0$. Having defined individual trajectories of progression, the observations are seen as random samples along these trajectories:

$$y_{i,j} = \gamma_i(t_{i,j}) + \varepsilon_{i,j}. \qquad [v.2]$$

In this definition, the acceleration factor α_i allows to characterize whether the ith individual is progressing faster ($\alpha_i > 1$) or slower ($\alpha_i < 1$) than the average trajectory. The time shift τ_i allows to determine whether the ith individual is evolving ahead ($\tau_i < 0$) or behind ($\tau_i > 0$) the average trajectory.

The following proposition clarifies the link between this definition of individual trajectories of progression and the time reparametrizations introduced in the previous section.

Proposition V.1. *Let $p_0 \in M$, $t_0 \in \mathbb{R}$ and $v_0 \in \mathrm{T}_{p_0} M$. Let $\alpha_i > 0$ and $\tau_i \in \mathbb{R}$ and define the affine function $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$. Then,*

$$\forall s \in \mathbb{R},\ \mathrm{Exp}_{p_0,t_0+\tau_i}(\alpha_i v_0)(s) = \mathrm{Exp}_{p_0,t_0}(v_0)\big(\psi_i(s)\big). \qquad [v.3]$$

Proof. Introduce the curves $c_1 : s \in \mathbb{R} \mapsto \mathrm{Exp}_{p_0,t_0}(v_0)\big(\psi_i(s)\big)$ and $c_2 : s \in \mathbb{R} \mapsto \mathrm{Exp}_{p_0,t_0+\tau_i}(\alpha_i v_0)(s)$. Note that c_1 and c_2 are geodesics of M which, by definition, satisfy to:

$$c_1(t_0+\tau_i) = c_2(t_0+\tau_i) = p_0 \quad \text{and} \quad \dot{c_1}(t_0+\tau_i) = \dot{c_2}(t_0+\tau_i) = \alpha_i v_0. \qquad \text{[v.4]}$$

By unicity, $c_1 = c_2$. □

In addition to giving a simple interpretation of the acceleration factors $(\alpha_i)_{1\leq i\leq p}$ and the time shifts $(\tau_i)_{1\leq i\leq p}$, this proposition legitimates the choice of affine time reparametrizations of the form $\psi_i : t \mapsto \alpha_i(t - t_0 - \tau_i) + t_0$.

V.1.2 The "straight lines model"

Unbounded observations can be considered as points on the real line. The real line $M = \mathbb{R}$ equipped with its canonical metric is a geodesically complete one-dimensional Riemannian manifold. For the canonical metric, the geodesics are of the form $t \in \mathbb{R} \mapsto at + b$ with $(a, b) \in \mathbb{R}^2$. The generic model Eq. [v.1] writes:

$$y_{i,j} = p_0 + \alpha_i v_0(t_{i,j} - t_0 - \tau_i) + \varepsilon_{i,j}. \qquad \text{[v.5]}$$

This model is referred to as the **univariate straight lines model**. Note that, even though the average and individual trajectories are straight lines, the model is *not* linear due to the multiplication between the random effects α_i and τ_i.

V.1.2.1 Discussion

We propose to compare the nonlinear straight lines model to the linear mixed-effects model discussed in the introduction of this dissertation: the random slope and intercept model. Recall that this linear mixed-effects model writes:

$$y_{i,j} = (\overline{a} + a_i)(t_{i,j} - t_0) + (\overline{b} + b_i) + \varepsilon_{i,j}. \qquad \text{[v.6]}$$

This linear model analyzes the distribution of the observations at a fixed reference time t_0. In comparison, the straight lines model analyzes the distribution of the times at which the observations reach a given value of the measurements. These two different approaches are illustrated in Figure 11.

V.1.3 The "logistic curves model"

If the observations are bounded, such as percentages or scores to a test, the measurements can be normalized to produce new observations in the open interval $M =]0, 1[$.

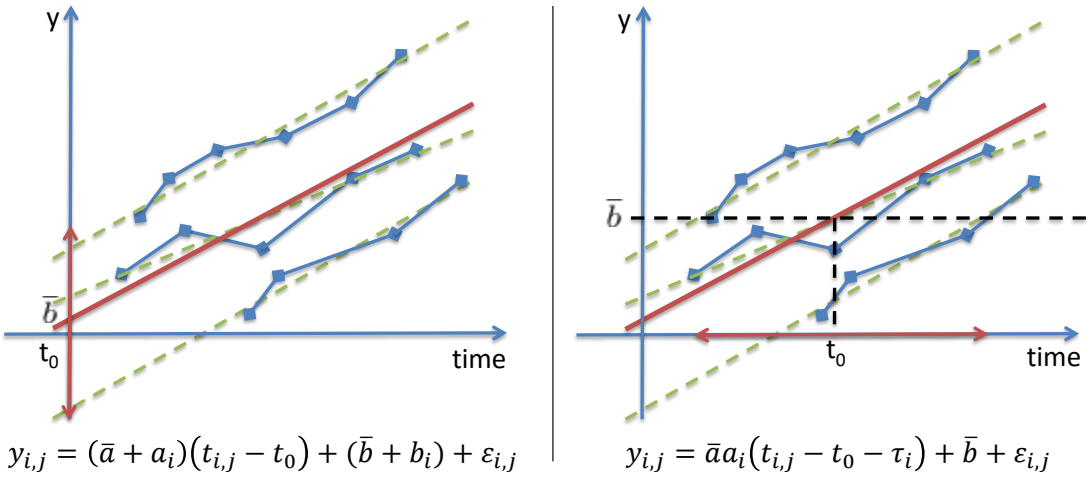


Figure 11 – Schematic example of a random slope and intercept linear mixed-effects model (left) and straight lines model (right).

Then, if $M =]0,1[$ is equipped with the Riemannian metric defined in Eq. [iv.5], the generic model Eq. [v.1] writes:

$$y_{i,j} = \left(1 + \left(\frac{1}{p_0} - 1\right)\exp\left(\frac{-\alpha_i v_0 (t_{i,j} - t_0 - \tau_i)}{p_0(1-p_0)}\right)\right)^{-1} + \varepsilon_{i,j}. \qquad \text{[v.7]}$$

This model for normalized longitudinal observations is referred to as the **logistic curves model**.

V.1.3.1 Discussion

In this framework, the Riemannian logarithm at $p \in]0,1[$ is given by: $\forall q \in]0,1[$, $\mathrm{Log}_p(q) = -p(1-p)\ln(p(1-q)) + p(1-p)\ln(q(1-p))$. In particular, at the point $p = 1/2$, which corresponds to the inflexion point of the logistics, the Riemannian logarithm is given by: $\forall q \in]0,1[$, $\mathrm{Log}_{1/2}(q) = (1/4)\mathrm{logit}(q)$. However, in Eq. [v.7], the point p_0 is not fixed to $1/2$. p_0 is a parameter of the model which is estimated along with the other parameters. Therefore, the logistic curves model is not equivalent to a linear model on the logit transform of the observations. The model, lifted up on the tangent space at $p = 1/2$ is *still* not linear due to the multiplication between the random effects α_i and τ_i. Instead of fixing p_0, the logistic curves model will estimate the best p_0, and therefore the best tangent space, at which describe the observations. In Section VI.4.2.1, we compare the logistic curves model with a linear mixed-effects model on logit-transformed observations. We show, on a longitudinal dataset of health data, that the logistic curves model explains a greater percentage of the total variance than the linear model.

V.2 The "SPD matrices model"

We describe how the generic spatiotemporal model can be used to analyze longitudinal datasets of 3×3 symmetric positive definite matrices. Such datasets may arise in Diffusion Tensor Imaging (DTI) or when observing the temporal evolution of stochastic process of covariance matrices. The space of 3×3 symmetric positive definite matrices is usually denoted by SDP(3), which is an open subset of the vector space of $(3, 3)$ symmetric real matrices, denoted by Sym(3). By identifying Sym(3) with $\mathbb{R}^6$, $\mathbb{M} =$ SDP(3) can be considered as an open subset of $\mathbb{R}^6$. In order to obtain a geodesically complete Riemannian manifold, $\mathbb{M}$ is equipped with a Riemannian metric called *affine-invariant* metric.

It follows from Eq. [iv.31] and Eq. [iv.32] that the generic spatiotemporal model Eq. [iv.66] for symmetric positive definite matrices writes:

$$\mathbf{Y}_{i,j} = \mathbf{P}_i(t_{i,j})^{1/2} \exp\big(\mathbf{P}_i(t_{i,j})^{-1/2}\mathbf{V}_i(t_{i,j})\mathbf{P}_i(t_{i,j})^{-1/2}\big)\mathbf{P}_i(t)^{1/2} + \boldsymbol{\varepsilon}_{i,j} \qquad \text{[v.8]}$$

with, for all $t \in \mathbb{R}$,

$$\mathbf{P}_i(t) = \mathbf{P}_0^{1/2} \exp\big(\alpha_i(t - t_0 - \tau_i)\mathbf{P}_0^{-1/2}\mathbf{V}_0\mathbf{P}_0^{-1/2}\big)\mathbf{P}_0^{1/2} \qquad \text{[v.9]}$$

and:

$$\mathbf{V}_i(t) = \exp\Big(\frac{\alpha_i(t - t_0 - \tau_i)}{2}\mathbf{V}_0\mathbf{P}_0^{-1}\Big)\mathbf{W}_i \exp\Big(\frac{\alpha_i(t - t_0 - \tau_i)}{2}\mathbf{P}_0^{-1}\mathbf{V}_0\Big). \qquad \text{[v.10]}$$

The probability distributions of the matrices $\mathbf{P}_0$, $\mathbf{V}_0$ and $(\boldsymbol{\varepsilon}_{i,j})_{i,j}$ are defined as follows: $\mathbf{P}_0 \sim \mathcal{SN}(\overline{\mathbf{P}_0}, \sigma_{P_0}^2)$, $\mathbf{V}_0 \sim \mathcal{SN}(\overline{\mathbf{V}_0}, \sigma_{V_0}^2)$ and $\boldsymbol{\varepsilon}_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{SN}(\mathbf{0}, \sigma^2)$, where $\mathcal{SN}$ denotes the Gaussian distribution on the vector space $\mathrm{Sym}(n)$. Given $\overline{\mathbf{M}} \in \mathrm{Sym}(n)$, the probability distribution $\mathcal{SN}(\overline{\mathbf{M}}, \sigma^2)$ on $\mathrm{Sym}(n)$ is defined by the density function q:

$$q(\mathbf{M}) = \frac{1}{(2\pi)^{m/2}\sigma^m} \exp\Big(-\frac{1}{2\sigma^2}\mathrm{tr}\big[(\overline{\mathbf{M}} - \mathbf{M})^2\big]\Big),\ \mathbf{M} \in \mathrm{Sym}(n) \qquad \text{[v.11]}$$

with $m = n(n + 1)/2$. The "standard" distribution $\mathcal{SN}(\mathbf{0}, 1)$ is used in physics and in the theory of random matrices. It is sometimes called **Gaussian Orthogonal Ensemble**. The probability distribution of the other fixed or random effects of the model are defined as in Eq. [iv.71]. This model will be referred to as the **symmetric positive definite matrices model** or $\mathrm{Spd}(n)$ **matrices model**.

V.3 Propagation models

This section presents a particular case of the generic spatiotemporal model which can be used to study the temporal progression of a set of N ($N \geq 1$) features which characterize

the evolution of a biological phenomenon. We assume that each feature is described by repeated *univariate* observations, which are random perturbations of quantities lying in a one-dimensional geodesically complete Riemannian manifold (M, g^M), open subset of $\mathbb{R}$. For each individual, at each time point, the observations $(\mathbf{y}_{i,j})_{1\leq i\leq p,\, 1\leq j\leq k_i}$ consist in a N-dimensional vector of univariate features. Hence, for this propagation model, the observations $(\mathbf{y}_{i,j})_{1\leq i\leq p,\, 1\leq j\leq k_i}$ are considered as random perturbations of quantities which belong to the product manifold $\mathbb{M} = M \times \ldots \times M = M^N$. Since each Riemannian manifold (M, g^M) is geodesically complete, $\mathbb{M}$ equipped with the product metric Eq. [8] is geodesically complete.

On the product manifold $\mathbb{M} = M^N$, the geodesics and parallel transport are given by the results in Section IV.1.2. In particular, these results show that a geodesic of $\mathbb{M}$ is of the form $t \mapsto \big(\gamma_1(t), \ldots, \gamma_N(t)\big)$, where $\gamma_1, \ldots, \gamma_N$ are geodesics of the one-dimensional Riemannian manifold M. Because we would like to model the joint temporal progression of N features, we propose to choose the group-average trajectory among a parametric family of geodesics of $\mathbb{M}$. This family is of the form:

$$\left\{ \boldsymbol{\gamma}_{0,\boldsymbol{\delta}} : t \in \mathbb{R} \mapsto \big(\gamma_0(t), \gamma_0(t+\delta_1), \ldots, \gamma_0(t+\delta_{N-1})\big) \right\} \qquad \text{[v.12]}$$

with $\boldsymbol{\delta} = (0, \delta_1, \ldots, \delta_{N-1})^\top$, $\delta_i \in \mathbb{R}$ and γ_0 denotes a geodesic of the one-dimensional Riemannian manifold g^M which goes through a point $p_0 \in M$ at time t_0 with velocity v_0. Following the ideas of the previous sections, if $M = \mathbb{R}$ (equipped with the canonical metric), γ shall be a straight line and if $M =]0, 1[$ (equipped with the Riemannian metric Eq. [iv.5]), γ shall be a logistic curve. The relative delay between two consecutive biomarkers is given by the parameters δ_i $(1 \leq i \leq N-1)$. The vector $\boldsymbol{\delta}$ is to be estimated as a fixed effect of the model. The first component of the vector $\boldsymbol{\delta}$ is chosen equal to zero to ensure the identifiability of the model.

Let $\boldsymbol{\delta} = (0, \delta_1, \ldots, \delta_{N-1})^\top \in \mathbb{R}^N$ and $\boldsymbol{\gamma}_{0,\boldsymbol{\delta}}(t) = (\gamma_0(t), \ldots, \gamma_0(t+\delta_{N-1}))$ be the group-average trajectory. The result of Proposition Eq. [IV.6] allows to compute a parallel variation of $\boldsymbol{\gamma}_{0,\boldsymbol{\delta}}$ in the direction of a tangent vector $\mathbf{w}_i \in \mathrm{T}_{\boldsymbol{\gamma}_{0,\boldsymbol{\delta}}(t_0)}\mathbb{M}$. The generic spatiotemporal model with the parametric family of geodesics write:

$$(\mathbf{y}_{i,j})_k = \gamma_0\Big(\frac{(\mathbf{w}_i)_k}{\dot{\gamma}_0(t_0+\delta_{k-1})} + \delta_{k-1} + \psi_i(t)\Big) + (\boldsymbol{\varepsilon}_{i,j})_k. \qquad \text{[v.13]}$$

where, for all $k \in \{1, \ldots, N\}$, $(\mathbf{y}_{i,j})_k$ denotes the kth component of $\mathbf{y}_{i,j}$. In other words, $(\mathbf{y}_{i,j})_k$ is the observation associated to the kth biomarker, for the ith individual, at the jth time point. Similarly, $(\mathbf{w}_i)_k$ denotes the kth component of the space shift $\mathbf{w}_i$. For all $i \in \{1, \ldots, p\}$, $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$ is the individual specific time reparametrization introduced in IV.3.1. This model is referred to as the *propagation model*. For this model, the latent variables are: $\mathbf{z}_{\text{pop}} = \big(p_0, t_0, v_0, (\delta_k)_{1\leq k\leq N-1}, (\beta_{l,k})_{l,k}\big)$ and, for all $i \in \{1, \ldots, p\}$, $\mathbf{z}_i = \big(\xi_i, \tau_i, (s_{l,i})_{l,i}\big)$. The definition of the individual latent variables $(\mathbf{z}_i)_{1\leq i\leq p}$ remains unchanged. For the population latent variables $\mathbf{z}_{\text{pop}}$, the variables

$(\delta_k)_{1\leq k\leq N-1}$ are added. We assume that the latent variables $\mathbf{z}_{\text{pop}}$ are distributed as follows:

$$p_0 \sim \mathcal{N}(\overline{p_0}, \sigma^2_{p_0}),\ t_0 \sim \mathcal{N}(\overline{t_0}, \sigma^2_{t_0}),\ v_0 \sim \mathcal{N}(\overline{v_0}, \sigma^2_{v_0}) \qquad \text{[v.14]}$$

and

$$\beta_{l,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\overline{\beta}_{l,k}, \sigma^2_\beta),\ \delta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\overline{\delta}_k, \sigma^2_\delta) \qquad \text{[v.15]}$$

where $\sigma^2_{p_0}$, $\sigma^2_{t_0}$, $\sigma^2_{v_0}$ and σ^2_δ are *fixed* variance parameters. Similarly to the generic spatiotemporal model, the latent variables are assumed independent of each other and independent of the noise variables $\boldsymbol{\varepsilon}_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathrm{I}_N)$. For the propagation model, the parameters space is defined as:

$$\begin{aligned}\Theta = \big\{\boldsymbol{\theta} = \big(&\overline{p_0}, \overline{t_0}, \overline{v_0}, (\overline{\delta}_k)_{1\leq k\leq N-1}, (\overline{\beta}_{l,k}), \boldsymbol{\theta}_{\text{var}}\big),\\ &(\overline{p_0}, \overline{v_0}) \in \mathrm{T}M,\ \overline{t_0} \in \mathbb{R},\ (\overline{\delta}_k)_k \in \mathbb{R}^{N-1},\ (\overline{\beta}_{l,k})_{l,k} \in \mathbb{R}^{(N-1)N_s},\ \boldsymbol{\theta}_{\text{var}} \in]0, +\infty[^3\big\}\end{aligned} \qquad \text{[v.16]}$$

where $\boldsymbol{\theta}_{\text{var}} = \big(\sigma^2_\xi, \sigma^2_\tau, \sigma^2\big)$ as defined for the generic spatiotemporal model. The prior assumed on the parameters of the propagation model writes:

$$\begin{aligned} q_{\text{prior}}(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}) \propto\ & \exp\Big(-\frac{1}{2s^2_{p_0}}(\overline{p_0} - \overline{\overline{p_0}})^2\Big) \times \exp\Big(-\frac{1}{2s^2_{t_0}}(\overline{t_0} - \overline{\overline{t_0}})^2\Big)\\ & \times \exp\Big(-\frac{1}{2s^2_{v_0}}(\overline{v_0} - \overline{\overline{v_0}})^2\Big) \times \exp\Big(-\frac{1}{2s^2_{\boldsymbol{\beta}}}\|\overline{\boldsymbol{\beta}}\|^2\Big)\\ & \times \exp\Big(-\frac{1}{2s^2_\delta}\sum_{k=1}^{N-1}(\overline{\delta}_k - \overline{\overline{\delta}}_k)^2\Big)\\ & \times \left(\frac{1}{\sqrt{\sigma^2_\xi}}\exp\Big(-\frac{\sigma^2_{\xi,0}}{2\sigma^2_\xi}\Big)\right)^{m_\xi} \times \left(\frac{1}{\sqrt{\sigma^2_\tau}}\exp\Big(-\frac{\sigma^2_{\tau,0}}{2\sigma^2_\tau}\Big)\right)^{m_\tau}\\ & \times \left(\frac{1}{\sqrt{\sigma^2}}\exp\Big(-\frac{\sigma^2_0}{2\sigma^2}\Big)\right)^{m_\sigma} \mathrm{d}\overline{\mathbf{p}_0}\,\mathrm{d}\overline{t_0}\,\mathrm{d}\overline{\mathbf{v}_0}\,d\overline{\boldsymbol{\beta}}\,\mathrm{d}\overline{\boldsymbol{\delta}}\,\mathrm{d}\sigma^2_\xi\,\mathrm{d}\sigma^2_\tau\,\mathrm{d}\sigma^2 \end{aligned} \qquad \text{[v.17]}$$

where m_ξ, m_τ, m_σ are fixed hyperparameters strictly greater that 2 and $\overline{\overline{p_0}}$, $\overline{\overline{t_0}}$, $\overline{\overline{v_0}}$, $(\overline{\overline{\delta}}_k)_{1\leq k\leq N-1}$, $s^2_{p_0}$, $s^2_{t_0}$, $s^2_{v_0}$, s^2_δ, $s^2_{\boldsymbol{\beta}}$, $\sigma^2_{\xi,0}$, $\sigma^2_{\tau,0}$ and σ^2_0 are fixed hyperparameters. The vector $\boldsymbol{\theta}_{\text{hyper}}$ denotes the vector of the *fixed* hyperparameters.

V.3.1 The "straight lines propagation model"

If $M = \mathbb{R}$, as mentioned in Section IV.1.1.2, the geodesics of the real line are of the form $t \in \mathbb{R} \mapsto p_0 + v_0(t - t_0)$. Therefore, the propagation model Eq. [v.13] writes (in matrix form):

$$\mathbf{y}_{i,j} = p_0 \mathbb{1}_N + v_0 \alpha_i (t_{i,j} - t_0 - \tau_i)\mathbb{1}_N + v_0 \boldsymbol{\delta} + \mathbf{w}_i + \boldsymbol{\varepsilon}_{i,j} \qquad \text{[v.18]}$$

with $\boldsymbol{\delta} = (0, \delta_1, \ldots, \delta_{N-1})^\top$ and $\alpha_i = \exp(\xi_i)$. This model is called the **straight lines propagation model**.

V.3.2 The "logistic curves propagation model"

If $M =]0,1[$ is equipped with the Riemannian metric introduced in Section IV.1.1.3, then the geodesics of $]0,1[$ are logistic curves. In this situation, the propagation model Eq. [v.13] writes:

$$(\mathbf{y}_{i,j})_k = \left(1 + \left(\frac{1}{p_0} - 1\right)\exp\left(-\frac{v_0\alpha_i(t_{i,j} - t_0 - \tau_i) + v_0\delta_k + v_0\frac{(\mathbf{w}_i)_k}{\dot{\gamma}_0(t_0+\delta_k)}}{p_0(1-p_0)}\right)\right)^{-1} + (\boldsymbol{\varepsilon}_{i,j})_k \qquad \text{[v.19]}$$

with $\alpha_i = \exp(\xi_i)$. This model is called the **logistic curves propagation model**. Experimental results obtained with this model are presented in Section VII.

V.3.3 Discussion

Assuming that the group-average geodesic $\boldsymbol{\gamma}_{0,\boldsymbol{\delta}}$ belongs to the parametric family Eq. [v.12] is equivalent to assuming the progression of the biomarkers is described by trajectories which have the same shape but are shifted in time.

It is interesting to note that for this model, the individual trajectories are of the same shape as the group-average trajectory. More precisely, each individual trajectory $\boldsymbol{\gamma}_i$ is of the form $\boldsymbol{\gamma}_{0,\tilde{\boldsymbol{\delta}}_i}$ with $\tilde{\boldsymbol{\delta}}_i = (0, \tilde{\delta}_1, i, \ldots, \tilde{\delta}_{(N-1),i})$ and:

$$\forall k \in \{1, \ldots, N-1\},\ \tilde{\delta}_k = \frac{(\mathbf{w}_i)_k}{\dot{\gamma}_0(t_0 + \delta_{k-1})} + \delta_{k-1}. \qquad \text{[v.20]}$$

Given the definition of the space shift $\mathbf{w}_i$, we have: $\mathbb{E}[\mathbf{w}_i] = 0$. As a consequence, the delays $\tilde{\delta}_k$ $(1 \leq k \leq N-1)$ can be interpreted as random perturbations of $(\delta_k)_{1\leq k\leq N-1}$: $\mathbb{E}[\tilde{\delta}_k] = \delta_k$. This shows that constructing a parallel variation of $\boldsymbol{\gamma}_{0,\boldsymbol{\delta}}$ (in the direction of a tangent vector $\mathbf{w}_i$) may change the relative delay, and possibly the ordering, between the different biomarkers.

Part VI

Statistical inference and algorithms

Summary

VI.1 Existence of a *maximum a posteriori*

In Section IV.3.4, the generic spatiotemporal model is defined in a Bayesian framework. In this context, given some observations $\mathbf{y}$, the parameters of the generic model can be estimated in a *maximum a posteriori* (MAP) approach. Let $\widehat{\boldsymbol{\theta}}_{\text{MAP}}$ denote the MAP estimates of the parameters of the model, defined by:

$$\widehat{\boldsymbol{\theta}}_{\text{MAP}} \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, q(\boldsymbol{\theta} \mid \mathbf{y}). \tag{vi.1}$$

where $q(\boldsymbol{\theta} \mid \mathbf{y})$ is the posterior distribution of the parameters $\boldsymbol{\theta}$ given the observations $\mathbf{y}$. This section provides theoretical results regarding the existence of a maximum a posteriori, given a dataset $\mathbf{y}$. The main result of this section is given in Theorem VI.1.

VI.1.1 Main result

Lemma VI.1. *Given the generic spatiotemporal model (see Eq. [iv.66]), the choice of probability distributions for the parameters (see Eq. [iv.71]) and latent variables of the model, the posterior $\boldsymbol{\theta} \in \Theta \mapsto q(\boldsymbol{\theta} \mid \mathbf{y})$ is continuous on the parameters space Θ.*

Proof. Using the notations introduced in Section IV.3.4, let $\mathcal{Z}$ denote the space of latent variables in the generic spatiotemporal model:

$$\mathcal{Z} = \left\{ \left(\mathbf{z}_{\text{pop}}, (\mathbf{z}_i)_{1\leq i\leq p}\right),\ \mathbf{z}_{\text{pop}} \in \mathbb{M} \times \mathbb{R}^{(N-1)N_s+N+1}; \forall i \in \{1,\dots,p\},\ \mathbf{z}_i \in \mathbb{R}^{N_s+2} \right\}. \tag{vi.2}$$

The posterior $\boldsymbol{\theta} \in \Theta \mapsto q(\boldsymbol{\theta} \mid \mathbf{y})$ is defined by:

$$\begin{aligned} \forall \boldsymbol{\theta} \in \Theta,\ q(\boldsymbol{\theta} \mid \mathbf{y}) &= \frac{1}{q(\mathbf{y})} q(\mathbf{y} \mid \boldsymbol{\theta}) q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}) \\ &= \frac{1}{q(\mathbf{y})} \left(\int_{\mathcal{Z}} q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z} \mid \boldsymbol{\theta})\, d\mathbf{z} \right) q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}). \end{aligned} \tag{vi.3}$$

The probability distribution q_{prior} is defined in Eq. [iv.71]. This density function $\boldsymbol{\theta} \in \Theta \mapsto q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}})$ is continuous on Θ since it is a product of continuous functions. In order to prove the continuity of $\boldsymbol{\theta} \in \Theta \mapsto q(\boldsymbol{\theta} \mid \mathbf{y})$, it suffices to prove that, for any compact $L \subset \Theta$, there exist a positive function φ_L, defined and integrable on Θ, such that:

$$\forall \mathbf{z} \in \mathcal{Z},\ \forall \boldsymbol{\theta} \in L \subset \Theta,\ q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z} \mid \boldsymbol{\theta}) \leq \varphi_L(\mathbf{z}). \tag{vi.4}$$

Given that:

$$q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}) = \frac{1}{(\sigma\sqrt{2\pi})^{NK}} \exp\left(-\frac{1}{2\sigma^2} \sum_{\substack{1\leq i\leq p \\ 1\leq j\leq k_i}} \left\| \mathbf{y}_{i,j} - \boldsymbol{\eta}^{\mathbf{w}_i}\left(\boldsymbol{\gamma}_0, \psi_i(t_{i,j})\right) \right\|^2 \right) \tag{vi.5}$$

we have $q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}) \leq 1$ for all $\mathbf{z} \in \mathcal{Z}$ and all $\boldsymbol{\theta} \in \Theta$. Therefore, the inequality Eq. [vi.4] holds by continuity of the mapping $\boldsymbol{\theta} \in \Theta \mapsto q(\mathbf{z} \mid \boldsymbol{\theta})$ for every $\mathbf{z} \in \mathcal{Z}$. □

The main result of this section is given in the following theorem.

Theorem VI.1. *Given the generic spatiotemporal model and the choice of probability distributions for the parameters and latent variables of the model, for any dataset* $(t_{i,j}, \mathbf{y}_{i,j})_{1\leq i\leq p,\, 1\leq j\leq k_i}$*, there exist* $\widehat{\boldsymbol{\theta}}_{\text{MAP}}$ *such that:* $\widehat{\boldsymbol{\theta}}_{\text{MAP}} \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, q(\boldsymbol{\theta} \mid \mathbf{y})$.

Proof. For clarity, let $\overline{\boldsymbol{\beta}}$ denote the vector $(\overline{\beta}_{l,k})_{1\leq l\leq N_s,\, 1\leq k\leq (N-1)}$. Recall from Eq. [iv.70] that the parameters space Θ is defined by:

$$\begin{aligned}\Theta &= \Big\{\boldsymbol{\theta} = \big(\overline{\mathbf{p}_0}, \overline{t_0}, \overline{\mathbf{v}_0}, \overline{\boldsymbol{\beta}}, \boldsymbol{\theta}_{\text{var}}\big),\ /\ (\overline{\mathbf{p}_0}, \overline{\mathbf{v}_0}) \in \mathbb{TM},\ \overline{t_0} \in \mathbb{R},\\ &\qquad \overline{\boldsymbol{\beta}} \in \mathbb{R}^{(N-1)N_s},\ \boldsymbol{\theta}_{\text{var}} = (\sigma_\xi^2, \sigma_\tau^2, \sigma^2) \in]0, +\infty[^2\Big\} \qquad \text{[vi.6]}\\ &= \big(\mathbb{M}\times\mathbb{R}\big)\times\mathbb{R}\times\mathbb{R}^{(N-1)N_s}\times]0, +\infty[^2.\end{aligned}$$

because $\mathbb{M}$ is assumed to be an open subset of $\mathbb{R}^N$ and therefore: $\mathbb{TM} = \mathbb{M}\times\mathbb{R}^N$. Below, $\mathbb{M}$ is equipped with the induced norm from $\mathbb{R}^N$. Given the result of Lemma VI.1, in order to prove that $\boldsymbol{\theta} \in \Theta \mapsto \log q(\boldsymbol{\theta} \mid \mathbf{y})$ has a maximum, we prove:

$$\lim_{\substack{\|\overline{\mathbf{p}_0}\|, |t_0|, \|\overline{\mathbf{v}_0}\|, \|\overline{\boldsymbol{\beta}}\|\to+\infty\\ \sigma_\xi^2+(1/\sigma_\xi^2)\to+\infty\\ \sigma_\tau^2+(1/\sigma_\tau^2)\to+\infty\\ \sigma^2+(1/\sigma^2)\to+\infty}} \log q(\boldsymbol{\theta} \mid \mathbf{y}) = -\infty. \qquad \text{[vi.7]}$$

Using Bayes rule,

$$q(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{1}{q(\mathbf{y})}\, q(\mathbf{y} \mid \boldsymbol{\theta}) q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}) \qquad \text{[vi.8]}$$

where $q(\mathbf{y} \mid \boldsymbol{\theta})$ is the observed likelihood and $q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}})$ the prior of $\boldsymbol{\theta}$ (see Eq. [iv.71]). The constant $q(\mathbf{y})$ is called *model evidence* and does not depend on the parameters $\boldsymbol{\theta}$. As in Section IV.3.4, let $\mathbf{z}_{\text{pop}} = (\overline{\mathbf{p}_0}, \overline{t_0}, \overline{\mathbf{v}_0}, \overline{\boldsymbol{\beta}})$ denote the *population* latent variables of the model and, for $1 \leq i \leq p$, $\mathbf{z}_i = \big(\xi_i, \tau_i, (s_{l,i})_{l,i}\big)$, the *individual* latent variables. Let $\mathbf{z} = \big(\mathbf{z}_{\text{pop}}, (\mathbf{z}_i)_{1\leq i\leq p}\big)$. By definition:

$$q(\mathbf{y} \mid \boldsymbol{\theta}) = \int_{\mathcal{Z}} q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z} \mid \boldsymbol{\theta})\, d\mathbf{z}. \qquad \text{[vi.9]}$$

The *model likelihood* $q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta})$ writes:

$$q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}) = \frac{1}{(\sigma\sqrt{2\pi})^{NK}} \exp\left(-\frac{1}{2\sigma^2} \sum_{\substack{1\leq i\leq p\\ 1\leq j\leq k_i}} \big\|\mathbf{y}_{i,j} - \boldsymbol{\eta}^{\mathbf{w}_i}\big(\boldsymbol{\gamma}_0, \psi_i(t_{i,j})\big)\big\|^2 \right). \qquad \text{[vi.10]}$$

Bounding the exponential in Eq. [vi.10] above by 1 leads to:

$$q(\mathbf{y} \mid \boldsymbol{\theta}) \leq \frac{1}{(\sigma\sqrt{2\pi})^{NK}} \int_{\mathcal{Z}} q(\mathbf{z} \mid \boldsymbol{\theta})\, d\mathbf{z} \qquad \text{[vi.11]}$$

and the term on the right integrates to 1. It follows from Eq. [vi.8] that:

$$q(\boldsymbol{\theta} \mid \mathbf{y}) \leq \frac{1}{q(\mathbf{y})} \frac{1}{(\sigma\sqrt{2\pi})^{NK}} q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}). \tag{vi.12}$$

Or, equivalently:

$$\log q(\boldsymbol{\theta} \mid \mathbf{y}) \leq C(\mathbf{y}) - NK \log(\sigma) + \log q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}). \tag{vi.13}$$

where $C(\mathbf{y}) = -\log q(\mathbf{y}) - \dfrac{NK}{2} \log(2\pi)$. And:

$$\begin{aligned} \log q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}) = C &- \frac{1}{2s_{\mathbf{p}_0}^2} \|\overline{\mathbf{p}_0} - \overline{\overline{\mathbf{p}_0}}\|^2 - \frac{1}{2s_{t_0}^2} (\overline{t_0} - \overline{\overline{t_0}})^2 - \frac{1}{2s_{\mathbf{v}_0}^2} \|\overline{\mathbf{v}_0} - \overline{\overline{\mathbf{v}_0}}\|^2 \\ &- \frac{1}{2s_{\boldsymbol{\beta}}^2} \|\overline{\boldsymbol{\beta}}\|^2 - m_\xi \log(\sigma_\xi) - m_\xi \frac{\sigma_{\xi,0}^2}{2\sigma_\xi^2} - m_\tau \log(\sigma_\tau) - m_\tau \frac{\sigma_{\tau,0}^2}{2\sigma_\tau^2} \\ &- m_\sigma \log(\sigma) - m_\sigma \frac{\sigma_0^2}{2\sigma^2}. \end{aligned} \tag{vi.14}$$

In Eq. [vi.14], C is a sum of terms which only depends on the *fixed* hyper parameters of the model. Putting Eq. [vi.13] and Eq. [vi.14] together gives:

$$\begin{aligned} \log q(\boldsymbol{\theta} \mid \mathbf{y}) \leq \tilde{C}(\mathbf{y}) &- NK \log(\sigma) - \frac{1}{2s_{\mathbf{p}_0}^2} \|\overline{\mathbf{p}_0} - \overline{\overline{\mathbf{p}_0}}\|^2 - \frac{1}{2s_{t_0}^2} (\overline{t_0} - \overline{\overline{t_0}})^2 \\ &- \frac{1}{2s_{\mathbf{v}_0}^2} \|\overline{\mathbf{v}_0} - \overline{\overline{\mathbf{v}_0}}\|^2 - \frac{1}{2s_{\boldsymbol{\beta}}^2} \|\overline{\boldsymbol{\beta}}\|^2 - m_\xi \log(\sigma_\xi) - m_\xi \frac{\sigma_{\xi,0}^2}{2\sigma_\xi^2} \\ &- m_\tau \log(\sigma_\tau) - m_\tau \frac{\sigma_{\tau,0}^2}{2\sigma_\tau^2} - m_\sigma \log(\sigma) - m_\sigma \frac{\sigma_0^2}{2\sigma^2}. \end{aligned} \tag{vi.15}$$

Finally, the existence result holds since:

$$\begin{aligned} \lim_{\|\mathbf{p}_0\|, |t_0|, \|\mathbf{v}_0\|, \|\boldsymbol{\beta}\| \to +\infty} \Bigg(&- \frac{1}{2s_{\mathbf{p}_0}^2} \|\overline{\mathbf{p}_0} - \overline{\overline{\mathbf{p}_0}}\|^2 - \frac{1}{2s_{t_0}^2} (\overline{t_0} - \overline{\overline{t_0}})^2 \\ &- \frac{1}{2s_{\mathbf{v}_0}^2} \|\overline{\mathbf{v}_0} - \overline{\overline{\mathbf{v}_0}}\|^2 - \frac{1}{2s_{\boldsymbol{\beta}}^2} \|\overline{\boldsymbol{\beta}}\|^2 \Bigg) = -\infty \end{aligned} \tag{vi.16}$$

And:

$$\begin{aligned} \lim_{\substack{\sigma_\xi^2 + (1/\sigma_\xi^2) \to 0 \\ \sigma_\tau^2 + (1/\sigma_\tau^2) \to 0 \\ \sigma^2 + (1/\sigma^2) \to 0}} \Bigg(&- NK \log(\sigma) - m_\xi \log(\sigma_\xi) - m_\xi \frac{\sigma_{\xi,0}^2}{2\sigma_\xi^2} \\ &- m_\tau \log(\sigma_\tau) - m_\tau \frac{\sigma_{\tau,0}^2}{2\sigma_\tau^2} - m_\sigma \log(\sigma) - m_\sigma \frac{\sigma_0^2}{2\sigma^2} \Bigg) = -\infty. \end{aligned} \tag{vi.17}$$

□

VI.2 Inference in nonlinear mixed-effects models

This section reviews several algorithms for the statistical inference in nonlinear mixed-effects models and discusses the advantages and drawbacks of each method. The methods and algorithms presented below are grouped into two classes: the *deterministic algorithms* and the *stochastic algorithms*. Unlike deterministic algorithms, stochastic methods require to generate random samples. Still, methods from both classes aim at obtaining maximum likelihood estimates (MLE) or maximum a posteriori estimates (MAP), in a Bayesian framework.

VI.2.1 A brief review of nonlinear mixed-effects models

Contrary to linear mixed-effects models, nonlinear mixed-effects models assume that the fixed and random effects contribute nonlinearly to the response variable $\mathbf{y}$. Nonlinear mixed-effects (NLME) models first appeared in the work of Sheiner and Beal [Sheiner and Beal, 1980] and have been a blooming topic of research since 1990. These models are now popular tools in a large variety of areas, such as pharmacokinetic modeling, medicine, etc. NLME models assume that a longitudinal dataset $(\mathbf{y}_{i,j}, t_{i,j})_{1\leq i\leq p,\, 1\leq j\leq k_i}$, with $\mathbf{y}_{i,j} \in \mathbb{R}^d$, arises from:

$$\mathbf{y}_{i,j} = f(\boldsymbol{\psi}_{i,j}, t_{i,j}) + \boldsymbol{\varepsilon}_{i,j} \qquad \text{[vi.18]}$$

where $\boldsymbol{\psi}_{i,j} = \mathbf{X}_{i,j}\boldsymbol{\alpha} + \mathbf{Z}_{i,j}\boldsymbol{\beta}_i$. $(\mathbf{X}_{i,j})_{i,j}$ and $(\mathbf{Z}_{i,j})_{i,j}$ are design matrices linking the fixed (respectively random) effects $\boldsymbol{\alpha}$ (respectively $\boldsymbol{\beta}_i$) to $\boldsymbol{\psi}_{i,j}$. The random effects are distributed as follows: $\boldsymbol{\beta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{D})$ and independent of the noise $\boldsymbol{\varepsilon}_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$. f is a nonlinear mapping from $\mathbb{R}^{p+q} \times \mathbb{R}$ to $\mathbb{R}^d$. One can easily note that the LME models appear as a particular case of NLME models.

The greater flexibility of the NLME models comes at the price of an intractable likelihood. Indeed, if Eq. [vi.18] is written $\mathbf{y}_i = f(\boldsymbol{\psi}_i, t_i) + \boldsymbol{\varepsilon}_i$ in matrix form, then the likelihood $q(\mathbf{y}_i \mid \boldsymbol{\alpha}, \boldsymbol{\theta})$ writes:

$$q(\mathbf{y}_i \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) = \int q(\mathbf{y}_i \mid \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta}_i) q(\boldsymbol{\beta}_i \mid \boldsymbol{\theta})\, d\boldsymbol{\beta}_i \qquad \text{[vi.19]}$$

where the integral over the random effects is often intractable. As a consequence, the methods to estimate the parameters of LME models cannot be used directly. Methods specific to NLME models are reviewed in the following section. We shall see that some of these methods consist in linearizing the nonlinear model in order to "approximate" it with a linear model. Other methods consist in approximations of the integral in Eq. [vi.19]. Another notable difference between LME and NLME models is that NLME models are usually more sensitive to the initialization of the fixed effects. Finally, the increased flexibility and complexity of the model implies a higher computational cost to fit these models.

VI.2.2 Deterministic algorithms

Since the 1990's, many methodological contributions have been made to the topic of inference in NLME models. There are currently several methods to address this problem. As mentioned above, the likelihood in NLME models is not available in closed-form. Therefore, several methods to estimate the parameters of NLME models consist in approximations of the likelihood and linearization of the model.

Using the notations introduced in VI.2.1, NLME models write (in matrix form) $\mathbf{y}_i = f(\boldsymbol{\psi}_i, t_i) + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\psi}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{D})$ is independent of $\boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 \mathrm{I}_{k_i})$. Similarly to LME models, the parameters to be estimated are $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \mathbf{D}, \sigma^2)$, with $\boldsymbol{\alpha} \in \mathbb{R}^p$, $\mathbf{D} \in \mathrm{Spd}(q)$ and $\sigma^2 \in]0, +\infty[$. If $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_p)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$, the likelihood $q(\mathbf{y} \mid \boldsymbol{\theta})$ writes:

$$q(\mathbf{y} \mid \boldsymbol{\theta}) = \int q(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) q(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \, d\boldsymbol{\beta}. \qquad \text{[vi.20]}$$

In [Lindstrom and Bates, 1990], Lindstrom and colleagues proposed a two-steps algorithm called "LME approximation", which approximates the likelihood in Eq. [vi.20] with the one of a LME model. This algorithm is implemented in the R package `nlme` and in the MATLAB function `nlmefit`. Davidian et al., in [Davidian and Gallant, 1992], used an adaptive Gaussian quadrature to approximate the likelihood Eq. [vi.20]. This is implemented in the SAS Software procedure `proc nlmixed`. Laplacian approximation, which is equivalent to the adaptive Gaussian quadrature method with only one quadrature point is discussed in [Pinheiro, 1994] and also implemented in the SAS procedure `proc nlmixed`. Note that the LME approximation, Laplacian approximation and adaptive Gaussian quadrature are also detailed in the book [Pinheiro and Bates, 2006].

For Bayesian inference, the famous Expectation-Maximization (EM, [Dempster et al., 1977]) algorithm is a natural choice. However, the nonlinearity and complexity of the model makes the "E-step" usually intractable. Therefore, a stochastic version of the EM algorithm, called the Monte Carlo Markov Chains Stochastic Approximation EM (MCMC-SAEM) algorithm can be used to address this problem. The MCMC-SAEM is implemented in the `Monolix` software, dedicated especially to pharmacokinetic/pharmacodynamic (PK-PD) modeling [Lavielle and Mentré, 2007, Savic et al., 2011].

The adaptive Gaussian quadrature and importance sampling are not reviewed in this dissertation. In the following, the LME approximation and Laplace approximation are discussed. A short remainder on the EM algorithm and the MCMC-SAEM are presented below.

VI.2.2.1 The LME approximation

In order to discuss the LME approximation of Lindstrom and Bates, let $\mathbf{D}^{-1} = \sigma^{-2}\boldsymbol{\Delta}^\top\boldsymbol{\Delta}$. The matrix $\boldsymbol{\Delta}$ is called *precision matrix*. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\Delta}, \sigma^2)$ and note that the likelihood $q(\mathbf{y} \mid \boldsymbol{\theta})$ of a NLME model writes:

$$\begin{aligned} q(\mathbf{y} \mid \boldsymbol{\theta}) &= \prod_{i=1}^{p} \int q(\mathbf{y}_i \mid \boldsymbol{\beta}_i, \boldsymbol{\theta}) q(\boldsymbol{\beta}_i \mid \boldsymbol{\theta})\, d\boldsymbol{\beta}_i \\ &= \frac{1}{(2\pi\sigma^2)^{(K+qp)/2}} (\det \boldsymbol{\Delta})^p \prod_{i=1}^{p} \int \exp\Big(-\frac{1}{2\sigma^2} \Big[\|\mathbf{y}_i - f(\boldsymbol{\psi}_i, t_i)\|^2 \\ &\qquad + \frac{1}{2}\boldsymbol{\beta}_i^\top \boldsymbol{\Delta}^\top \boldsymbol{\Delta} \boldsymbol{\beta}_i \Big] \Big)\, d\boldsymbol{\beta}_i. \end{aligned} \tag{vi.21}$$

The LME approximation consists in two steps: the first is called "Penalized Nonlinear Least-Squares (PNLS) step" and the second step is called "LME step". The PNLS step consists in solving the optimization problem

$$\big(\widehat{\boldsymbol{\alpha}}, (\widehat{\boldsymbol{\beta}})_{1\le i\le p}\big) = \underset{\boldsymbol{\alpha}, (\boldsymbol{\beta}_i)_{1\le i\le p}}{\operatorname{argmin}} \sum_{i=1}^{p} \|\mathbf{y}_i - f(\boldsymbol{\psi}_i, t_i)\|^2 + \|\boldsymbol{\Delta}\boldsymbol{\beta}_i\|^2 \tag{vi.22}$$

where the precision matrix $\boldsymbol{\Delta}$ is considered fixed. Let $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\Delta}, \sigma^2)$ and assume an improper prior on $\boldsymbol{\alpha}$ (*id est* $q(\boldsymbol{\alpha} \mid \tilde{\boldsymbol{\theta}}) \propto 1$), then this step is equivalent to maximizing the conditional distribution $q(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{y}, \tilde{\boldsymbol{\theta}}) \propto q(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \tilde{\boldsymbol{\theta}}) q(\boldsymbol{\beta} \mid \tilde{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This explains why $\widehat{\boldsymbol{\alpha}}, (\widehat{\boldsymbol{\beta}}_i)_{1\le i\le p}$ are called *conditional modes*. The LME step considers the LME model

$$\tilde{\mathbf{y}}_i = \widehat{\mathbf{X}}_i \boldsymbol{\alpha} + \widehat{\mathbf{Z}}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \tag{vi.23}$$

where $\tilde{\mathbf{y}}_i = \mathbf{y}_i - f(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}_i) + \widehat{\mathbf{X}}_i \widehat{\boldsymbol{\alpha}} + \widehat{\mathbf{Z}}_i \widehat{\boldsymbol{\beta}}_i$ and:

$$\widehat{\mathbf{X}}_i = \frac{\partial f}{\partial \boldsymbol{\alpha}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}_i) \text{ and } \widehat{\mathbf{Z}}_i = \frac{\partial f}{\partial \boldsymbol{\beta}_i}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}_i). \tag{vi.24}$$

This LME model is obtained with a first-order Taylor expansion of the model function f around the conditional modes obtained from the PNLS step. Using methods presented in Appendix B, one can write the likelihood of this LME model and obtain estimates of the variance parameters $(\boldsymbol{\Delta}, \sigma^2)$. The two steps are repeated until convergence. One should note that Lindstrom and Bates do not provide order of convergence and control regarding the approximations which are made to derive this algorithm.

VI.2.2.2 The Laplacian approximation

The Laplacian approximation [Tierney and Kadane, 1986] is used to approximate the individual likelihoods $q(\mathbf{y}_i \mid \boldsymbol{\theta})$. Similarly to the LME approximation algorithm, it

is based on a first-order Taylor expansion. For the ith individual ($1 \leq i \leq p$), the likelihood $q(\mathbf{y}_i \mid \boldsymbol{\theta})$ is given by:

$$q(\mathbf{y}_i \mid \boldsymbol{\theta}) = \frac{\det \boldsymbol{\Delta}}{(2\pi\sigma^2)^{(k_i+q)/2}} \int \exp\Big(-\frac{1}{2\sigma^2}\Big[\|\mathbf{y}_i - f(\boldsymbol{\psi}_i, t_i)\|^2 + \|\boldsymbol{\Delta}\boldsymbol{\beta}_i\|^2\Big]\Big)\, d\boldsymbol{\beta}_i. \quad \text{[vi.25]}$$

Let $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\Delta}, \sigma^2)$ and:

$$g_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_i) = \|\mathbf{y}_i - f(\boldsymbol{\psi}_i, t_i)\|^2 + \|\boldsymbol{\Delta}\boldsymbol{\beta}_i\|^2 \text{ and } \widehat{\boldsymbol{\beta}}_i(\boldsymbol{\alpha}, \tilde{\boldsymbol{\theta}}) = \underset{\boldsymbol{\beta}_i}{\operatorname{argmin}}\, g(\boldsymbol{\alpha}, \boldsymbol{\beta}_i, \tilde{\boldsymbol{\theta}}). \quad \text{[vi.26]}$$

Using a first-order Taylor expansion of $g_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\alpha}, \cdot)$ around $\widehat{\boldsymbol{\beta}}_i$, we have:

$$g_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_i) \simeq g_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i) + \frac{1}{2}(\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i)^\top \frac{\partial^2 g_{\tilde{\boldsymbol{\theta}}}}{\partial \boldsymbol{\beta}_i^2}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i)(\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i) \quad \text{[vi.27]}$$

where $\partial^2 g_{\tilde{\boldsymbol{\theta}}}/\partial \boldsymbol{\beta}_i^2(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i)$ denotes the Hessian matrix of $g_{\tilde{\boldsymbol{\theta}}}$ at $(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i)$. Plugging Eq. [vi.27] into Eq. [vi.25], $q(\mathbf{y}_i \mid \boldsymbol{\theta})$ writes (up to some normalization terms) as the integral of a multivariate Gaussian density function. The Laplacian approximation consists in approximating $q(\mathbf{y}_i \mid \boldsymbol{\theta})$ with:

$$q(\mathbf{y}_i \mid \boldsymbol{\theta}) \simeq \frac{\det \boldsymbol{\Delta}}{(\sigma^2 2\pi)^{k_i/2}} \exp\Big(-\frac{1}{2\sigma^2} g_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i)\Big)\Big(\det \frac{\partial^2 g_{\tilde{\boldsymbol{\theta}}}}{\partial \boldsymbol{\beta}_i^2}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i)\Big)^{-1/2}. \quad \text{[vi.28]}$$

A notable drawback of this approximation is that it requires the computation of the inverse of the Hessian matrix of $g_{\tilde{\boldsymbol{\theta}}}$ at $(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_i)$. In practice, for complex models, this matrix cannot be computed in closed-form and its approximation using numerical schemes is usually very costly. Therefore, to address this problem, Pinheiro proposed to approximate the Hessian matrix. With this additional approximation, Eq. [vi.28] reduces to a more tractable form which can be maximized using gradient descent to provide estimates of the fixed effects and variance parameters.

VI.2.2.3 The Expectation-Maximization (EM) algorithm

The Expectation-Maximization (EM) algorithm [Dempster et al., 1977] is a popular algorithm which allows to obtain *maximum likelihood* (or *maximum a posteriori*) estimates of the parameters of a statistical model. The EM algorithm was introduced in the context of statistical models with latent variables. These are models for which the observed likelihood $q(\mathbf{y} \mid \boldsymbol{\theta})$ writes:

$$q(\mathbf{y} \mid \boldsymbol{\theta}) = \int q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})\, d\mu(\mathbf{z}) \quad \text{[vi.29]}$$

where the integral is taken over a set of *unobserved* random variables $\mathbf{z} \in \mathcal{Z}$ called *latent variables*. The function is integrated with respect to a measure μ, which is

defined by the latent variables $\mathbf{z}$. With continuous latent variables, μ could be the Lebesgue measure on $\mathbb{R}^D$ (where D corresponds to the dimension of the set of the latent variables). For discrete random variables (mixture of distributions, for example), then μ is the counting measure. The joint likelihood $q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$ of $\mathbf{y}, \mathbf{z}$ conditionally on the parameters $\boldsymbol{\theta}$ is the *complete likelihood*. Clearly, mixed-effects models belong to the class of latent variables models since the random effects can be considered as latent variables.

In general, with such models, the *observed likelihood* is not available in closed-form as it writes as an integral, over the latent variables, of the *complete likelihood*. Because this integral is often intractable, the idea of the EM algorithm is to maximize a *lower bound* on the observed likelihood. Under generic conditions described in [Dempster et al., 1977], corrected in [Wu, 1983] and generalized in [Delyon et al., 1999], the algorithm converges to critical point of the observed likelihood.

The EM algorithm iterates, until convergence, between two steps: the "E-step" and the "M-step". Let $\mathbf{y}$ (respectively $\mathbf{z}$) denote the observations (respectively latent variables) of the generic spatiotemporal model. Let $k \in \mathbb{N}^*$ and $\boldsymbol{\theta}^{(k)}$ denote the estimate of the parameters of the model at the kth iteration of the algorithm. Let $q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$ denote the distribution of the observations and latent variables conditionally on the parameters $\boldsymbol{\theta}$. The "E-step" consists in computing the function $\boldsymbol{\theta} \in \Theta \mapsto \mathbf{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ defined by:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \mathbb{E}_{q(\cdot|\mathbf{y},\boldsymbol{\theta}^{(k)})}\big[\log q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})\big]. \qquad \text{[vi.30]}$$

In Eq. [vi.30], the expectation is taken with respect to the conditional distribution of the latent variables knowing the observations and current estimate of the parameters. This function $\mathbf{Q}(\cdot \mid \boldsymbol{\theta}^{(k)})$ is then used in the "M-step" to update the estimate of the parameters as follows:

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \big(\mathbf{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) + q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}})\big). \qquad \text{[vi.31]}$$

However, for most nonlinear mixed-effects models, the "E-step" of the EM algorithm is intractable. To address this problem, the *Monte Carlo Markov Chain - Stochastic Approximation EM* (MCMC-SAEM), introduced in [Kuhn and Lavielle, 2004] and proved convergent in [Allassonnière et al., 2010], replaces this step with a stochastic approximation of the expectation in Eq. [vi.30]. The MCMC-SAEM is described in the next section.

VI.2.2.4 Other deterministic algorithms

Other possible methods such as adaptive Gaussian quadrature are discussed in Pinheiro's Ph.D. dissertation [Pinheiro, 1994]. This method are compared to the Laplacian approximation and LME approximation in terms of number of iterations requires to reach convergence. The adaptive Gaussian quadrature is far more computationally intensive than the other methods described above. Since the cost of these methods is

already important with univariate observations, it can be expected that these methods would not scale up to multivariate manifold-valued observations.

VI.2.3 Stochastic algorithms

VI.2.3.1 Towards a stochastic algorithm

The LME approximation algorithm is based on an approximation of the likelihood Eq. [vi.21], without any control or theoretical guarantee of the convergence towards a local maximum of the likelihood. Moreover, the methods has to be adapted in order to be used in a Bayesian framework. The MATLAB `nlmefit` code was tested with the univariate cases of the generic spatiotemporal model, without priors on the parameters of the model, and raised several numerical issues. As a matter of fact, the LME approximation requires (in the LME step) to compute the Jacobian matrix of the model function f at the conditional modes $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}_i)$. For individuals whose acceleration factor $\exp(\xi_i)$ is close to 1, the Jacobian matrix becomes badly conditioned thus leading to numerical instabilities which arise from trying to invert this Jacobian matrix in a linear system. Moreover, the least-squares criterion in the PNLS step is sometimes difficult to minimize since coordinates of the gradient of this criterion are almost proportional. To a greater extent, the LME approximation requires to compute derivative of the model function f with respect to its parameters and random effects. For the generic spatiotemporal model, this might not be possible since the Riemannian exponential and parallel transport may not be known in closed-form. Even if these quantities are known in closed-form, computing the derivatives of the parallel transport (along the average trajectory) with respect to its initial conditions may be a difficult problem. Computing these derivatives would result in a set of coupled ordinary differential equations (ODE) with second-order covariant derivative, similar to what is done in [Durrleman et al., 2011, Durrleman et al., 2013]. Such equations are computationally expensive to implement and compute.

The Laplace approximation of the likelihood, through the SAS procedure `nlmixed`, was used for the univariate straight lines model and the univariate logistic curves model. The results obtained with the SAS Software are presented in Section VII.4 and in [Schiratti et al., 2015d]. However, similarly to the `nlmefit` or `nlme` code, these functions are written to be used with univariate longitudinal observations. Their use with multivariate longitudinal observations is not straightforward and would require several modifications.

Stochastic algorithms such as the MCMC-SAEM offer the advantage that, with a particular family of MCMC samplers, these algorithms do not require to compute the derivative of the model function with respect to its parameters. Still, other methods for the Bayesian inference VI.2.3.3 may require to compute the gradient of the likelihood with respect to the parameters of the model, which may be computationally intensive.

VI.2.3.2 The MCMC-SAEM algorithm

Similarly to the EM algorithm VI.2.2.3, the Monte Carlo Markov Chain Stochastic Approximation EM (MCMC-SAEM) algorithm is presented for latent variables models. The MCMC-SAEM is an "EM-like" algorithm which iterates, until convergence, between three steps: *simulation*, *stochastic approximation* and *maximization*. These three steps are reviewed in the following sections.

VI.2.3.2.1 Sampling step

Let $\boldsymbol{\theta}^{(k)}$ denote the estimate of the parameters at the kth iteration of the algorithm. The first step of the algorithm, namely the **simulation** step, consists in sampling a set of latent variables $\mathbf{z}^{(k)}$ from the transition kernel $\boldsymbol{\pi}_{\boldsymbol{\theta}^{(k-1)}}$ of an ergodic Markov chain whose stationary distribution is the conditional distribution $q(\cdot \mid \mathbf{y}, \boldsymbol{\theta}^{(k-1)})$ of the latent variables knowing the observations $\mathbf{y}$ and the estimate of the parameters at the previous iteration. This step writes:

$$\mathbf{z}^{(k)} \sim \boldsymbol{\pi}_{\mathbf{y},\boldsymbol{\theta}^{(k-1)}}(\mathbf{z}^{(k-1)}, \cdot) \qquad \text{[vi.32]}$$

and is achieved using a MCMC sampler. A large variety of MCMC samplers can be used for this sampling step. However, theoretical results on the convergence of the MCMC-SAEM with unbounded latent variables [Allassonnière et al., 2010] require that the sampler produces an ergodic Markov chain whose convergence to its stationary distribution is uniformly geometric. As proven in [Allassonnière et al., 2010], this property holds for several samplers. In particular, the hybrid Metropolis-Hastings-within-Gibbs sampler is likely to satisfy the property in our model. The sampling step of the MCMC-SAEM is detailed below with the Metropolis-Hastings-within-Gibbs sampler. Still, other samplers such as the slice sampler or the Hit-and-Run sampler could be considered, even though they generally lead to numerous evaluations of the target distribution function. Hamiltonian Monte Carlo samplers, which require the computation of the gradient of the target distribution, may lead to heavy computations.

Using the Metropolis-Hastings-within-Gibbs sampler to sample $\mathbf{z}^{(k)}$ would write:

Algorithm 3 Gibbs sampler to sample from the transition kernel $\boldsymbol{\pi}_{\mathbf{y},\boldsymbol{\theta}^{(k-1)}}(\mathbf{z}^{(k-1)}, \cdot)$ in Eq. [vi.32].

Require: Set of latent variables $\mathbf{z}^{(k-1)}$, current estimate of the parameters $\boldsymbol{\theta}^{(k-1)}$

Ensure: Set of latent variables $\mathbf{z}^{(k)}$

1: Sample $z_1^{(k)}$ from $q(z_1 \mid \mathbf{y}, z_1^{(k-1)}, \ldots, z_L^{(k-1)}, \boldsymbol{\theta}^{(k-1)})$
2: Sample $z_2^{(k)}$ from $q(z_2 \mid \mathbf{y}, z_1^{(k)}, z_3^{(k-1)}, \ldots, z_L^{(k-1)}, \boldsymbol{\theta}^{(k-1)})$
3: ...
4: Sample $z_L^{(k)}$ from $q(z_L \mid \mathbf{y}, z_1^{(k)}, \ldots, z_{L-1}^{(k)}, \boldsymbol{\theta}^{(k-1)})$

The lth step ($1 \leq l \leq L$) of the Gibbs sampler presented above requires to sample $z_l^{(k)}$ from its target distribution, the *full conditional* $\pi_{k,l}$:

$$\pi_{k,l}(\cdot) = q(\cdot \mid \mathbf{y}, z_1^{(k)}, \ldots, z_{l-1}^{(k)}, z_{l+1}^{(k-1)}, \ldots, z_L^{(k-1)}, \boldsymbol{\theta}^{(k-1)}). \quad \text{[vi.33]}$$

The full conditional $\pi_{k,l}$ is the conditional distribution of z_l knowing $\mathbf{y}, \boldsymbol{\theta}^{(k-1)}$ and the most recent state of *all* the other latent variables of the model. For complex nonlinear models, the full conditionals are usually known up to a normalizing constant. In addition to this, $\pi_{k,l}$ does not belong to a standard family of distributions. Therefore, the *Metropolis-Hastings algorithm* is particularly suited to sample from $\pi_{k,l}$ since the normalizing constant of $\pi_{k,l}$ may be intractable. As a consequence, the sampling step is done using the (deterministic scan) Gibbs sampler combined with the Metropolis-Hastings (MH) algorithm. At each step, the *proposal distribution* of MH is chosen to be a Gaussian distribution centered at the current state, with a fixed variance. This results in an algorithm called: *Symmetric Random Walk Metropolis-Hastings* (SRW-MH). Other possible choices of proposal distributions and their influence on the convergence or computational cost of the MCMC-SAEM is discussed in Section VI.3.5. The SRW-MH within Gibbs sampler writes:

Algorithm 4 The Metropolis-Hastings-within-Gibbs sampler

Require: Set of latent variables $\mathbf{z}^{(k-1)}$, current estimate of the parameters $\boldsymbol{\theta}^{(k-1)}$, variances $(\sigma_l^2)_{1 \leq l \leq L}$ for the proposal distributions

Ensure: Set of latent variables $\mathbf{z}^{(k)}$

1: **for** $l = 1 \ldots L$ **do**

2: Draw a candidate z_l^* using the proposal distribution: $z_l^* \sim \mathcal{N}(z_l^{(k-1)}, \sigma_l^2)$

3: Compute the acceptance ratio $\alpha(z_l^{(k-1)}, z_l^*)$:

$$\alpha(z_l^{(k-1)}, z_l^*) = \frac{\pi_{k,l}(z_l^*)}{\pi_{k,l}(z_l^{(k-1)})} \wedge 1 \quad \text{[vi.34]}$$

4: Draw $U \sim \text{Uniform}([0,1])$

5: Set:

$$z_l^{(k)} = \begin{cases} z_l^* & \text{if } U \leq \alpha(z_l^{(k-1)}, z_l^*) \\ z_l^{(k-1)} & \text{otherwise.} \end{cases}$$

6: **end for**

7: **Return**: $\mathbf{z}^{(k)} = (z_1^{(k)}, \ldots, z_L^{(k)})$.

In the algorithm 4, the acceptance ratio can be written in terms of the

model likelihood. For clarity, let $\mathbf{z}_{-l}^{(k-1),(k)}$ denote the set of latent variables $(z_1^{(k)}, \ldots, z_{l-1}^{(k)}, z_{l+1}^{(k-1)}, \ldots, z_L^{(k-1)})$. Then, the full conditional distribution $\pi_{k,l}$ writes: $q(\cdot \mid \mathbf{y}, \mathbf{z}_{-l}^{(k-1),(k)}, \boldsymbol{\theta}^{(k-1)})$. Using the Bayes rule, we have:

$$\pi_{k,l}(z_l^*) \propto q(\mathbf{y} \mid \mathbf{z}_{-l}^{(k-1),(k)}, z_l^*, \boldsymbol{\theta}^{(k-1)}) q(\mathbf{z}_{-l}^{(k-1),(k)} \mid \boldsymbol{\theta}^{(k-1)}) q(z_l \mid \boldsymbol{\theta}^{(k-1)}). \quad \text{[vi.35]}$$

In Eq. [vi.35], the term $q(z_l^* \mid \boldsymbol{\theta}^{(k-1)})$ (respectively $q(\mathbf{z}_{-l}^{(k-1),(k)} \mid \boldsymbol{\theta}^{(k-1)})$) denotes the likelihood of the probability distribution specified for z_l (respectively joint probability distribution of $(z_1, \ldots, z_{l-1}, z_{l+1}, \ldots, z_L)$) evaluated at z_l^* (respectively $\mathbf{z}_{-l}^{(k-1),(k)}$). As a consequence, the acceptance ratio in Eq. [vi.34] simplifies to:

$$\alpha(z_l^{(k-1)}, z_l^*) = \frac{q(\mathbf{y} \mid z_l^*, \mathbf{z}_{-l}^{(k-1),(k)}, \boldsymbol{\theta}^{(k-1)}) q(z_l^* \mid \boldsymbol{\theta}^{(k-1)})}{q(\mathbf{y} \mid z_l^{(k-1)}, \mathbf{z}_{-l}^{(k-1),(k)}, \boldsymbol{\theta}^{(k-1)}) q(z_l^{(k-1)} \mid \boldsymbol{\theta}^{(k-1)})} \wedge 1. \quad \text{[vi.36]}$$

VI.2.3.2.2 Stochastic approximation step

This paragraph on the stochastic approximation step starts with an important remark about theoretical results regarding the convergence of the MCMC-SAEM.

Remark. The convergence of the MCMC-SAEM is proved, in [Kuhn and Lavielle, 2004] (for bounded latent variables) and in [Allassonnière et al., 2010] (for unbounded latent variables), for statistical models which belong to the *curved exponential family*. That is to say, models for which the log complete likelihood $q(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta})$ writes:

$$\forall \boldsymbol{\theta} \in \Theta, \ \log q(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) = -\Phi(\boldsymbol{\theta}) + \langle \mathbf{S}(\mathbf{y}, \mathbf{z}), \Psi(\boldsymbol{\theta}) \rangle \quad \text{[vi.37]}$$

where Φ, Ψ are smooth functions of the parameters, $\mathbf{S}(\mathbf{y}, \mathbf{z})$ is a measurable function of the observations and latent variables called **sufficient statistic of the model** and $\langle \cdot, \cdot \rangle$ is an inner product on a product space. Proving that the model belongs to the curved exponential family is necessary because if the model does not have this property, the MCMC-SAEM might not converge.

The stochastic approximation step consists in constructing a sequence of functions $\big(\mathbf{Q}_k(\cdot)\big)_{k \geq 0}$ defined on Θ. Let $k \in \mathbb{N}^*$ denote the kth iteration of the MCMC-SAEM. Using the set $\mathbf{z}^{(k)}$ of latent variables obtained from the sampling step, the function $\boldsymbol{\theta} \in \Theta \mapsto \mathbf{Q}_k(\boldsymbol{\theta})$ is defined as follows:

$$\forall \boldsymbol{\theta} \in \Theta, \ \mathbf{Q}_k(\boldsymbol{\theta}) = \mathbf{Q}_{k-1}(\boldsymbol{\theta}) + \varepsilon_k \big(\log q(\mathbf{y}, \mathbf{z}^{(k)} \mid \boldsymbol{\theta}) - \mathbf{Q}_{k-1}(\boldsymbol{\theta}) \big). \quad \text{[vi.38]}$$

where $\mathbf{Q}_0 = 0$ and $(\varepsilon_k)_{k \geq 0}$ is a sequence of positive step-sizes which are such that: $\sum_{k \geq 0} \varepsilon_k = +\infty$ and $\sum_{k \geq 0} \varepsilon_k^2 < +\infty$. The choice of the sequence $(\varepsilon_k)_{k \geq 0}$ is discussed in Section VI.3.6. Note that Eq. [vi.38] is a *stochastic approximation* of the Robbins-Monro "like" [Robbins and Monro, 1951] which converges to the expectation

$\mathbb{E}_{q(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(k-1)})}\big[\log q(\mathbf{y},\mathbf{z}\mid\boldsymbol{\theta})\big]$. As a consequence, the *simulation* and *stochastic approximation* steps of the MCMC-SAEM are asymptotically equivalent to the "E-step" of the classical EM algorithm.

Assuming that the statistical model at hand belongs to the curved exponential family, the stochastic approximation in Eq. [vi.38] can be done on the sufficient statistics of the model. For each sufficient statistic of the model, initialize it with $\mathbf{S}_0 = 0$ and let:

$$\mathbf{S}_k = \mathbf{S}_{k-1} + \varepsilon_k\big(\mathbf{S}(\mathbf{y},\mathbf{z}^{(k)}) - \mathbf{S}_{k-1}\big) \qquad \text{[vi.39]}$$

during the "stochastic approximation" step of the algorithm. In that case, the function $\mathbf{Q}_k(\boldsymbol{\theta})$ is defined by:

$$\forall\boldsymbol{\theta}\in\Theta,\ \mathbf{Q}_k(\boldsymbol{\theta}) = -\Phi(\boldsymbol{\theta}) + \langle\mathbf{S}_k,\Psi(\boldsymbol{\theta})\rangle\,. \qquad \text{[vi.40]}$$

From the stochastic approximation (Eq. [vi.39]) on the sufficient statistics, one can note that if $\varepsilon_k = 1$, then $\mathbf{S}_k$ does not depend on $\mathbf{S}_{k-1}$. Intuitively, the sequence $(\mathbf{S}_k)_{k\geq 0}$ has "no memory" as long as $\varepsilon_k = 1$ and the MCMC-SAEM explores freely the parameters space during this period. Usually, an integer $N_b \in \mathbb{N}^*$ is chosen and the sequence $(\varepsilon_k)_{k\geq 0}$ is defined by:

$$\forall k\in\mathbb{N},\ \varepsilon_k = \begin{cases} 1 & \text{if } 0\leq k\leq N_b \\ (k-N_b)^{-\alpha} & \text{otherwise} \end{cases} \qquad \text{[vi.41]}$$

where $\alpha\in[1/2,1[$. The condition on α is necessary to ensure the convergence of the MCMC-SAEM (see [Allassonnière et al., 2010, Allassonniere and Kuhn, 2015]). The integer N_b is called **burn-in parameter**. Contrary to Bayesian inference, where "burn-in" traditionally refers to a certain amount of samples which are discarded, here the term "burn-in" refers to memoryless approximation steps. In other words, during the burn-in phase, the information contained in $\mathbf{z}^{(k)}$ is not used in the approximation of the sufficient statistics. In practice, the burn-in period is often chosen to be half of the maximum number of iterations.

Remark. The convergence of the algorithm is proved in [Kuhn and Lavielle, 2004] as long as the latent variables of the model belong to a compact of the Euclidean space. However, for latent variables models in which the latent variables have a non-compact support, a step called *truncation on random boundaries* (see [Andrieu et al., 2005]) is necessary. Let $\mathcal{S}$ denote the space of the sufficient statistics and consider an increasing sequence $(\mathcal{K}_n)_{n\geq 0}$ of *compact* subsets of $\mathcal{S}$ such that $\bigcup_n \mathcal{K}_n = \mathcal{S}$ and, for all n, $\mathcal{K}_n \subset \text{int}(\mathcal{K}_{n+1})$. At the beginning of the algorithm, one considers the compact $\mathcal{K}_0$. As long as the stochastic approximation $\mathbf{S}_{k-1} + \varepsilon_k\big(\mathbf{S}(\mathbf{y},\mathbf{z}^{(k)}) - \mathbf{S}_{k-1}$ is not "too far" from its previous value $\mathbf{S}_{k-1}$ and as long as this stochastic approximation states in the current compact, the algorithm continues. If one of the two conditions is not satisfies, the sequences $(\mathbf{z}^{(k)},\mathbf{S}_k)_{k\geq 0}$ are re-initialized using a projection and the size of the current compact is increased. These steps are described in [Allassonnière et al., 2010], where the convergence is proved for any latent variables. In practice, this

step is never required because the computer itself does not allow for arbitrarily large numbers. The results obtained with the MCMC-SAEM and presented in Chapter VII were obtained without this control.

VI.2.3.2.3 Maximization step

The last step of the MCMC-SAEM consists in updating the current estimate of the parameters of the model. This is done by maximizing, with respect to $\boldsymbol{\theta} \in \Theta$, the function $\mathbf{Q}_k(\cdot)$ defined in Eq. [vi.40]. The, the maximization step writes:

$$\boldsymbol{\theta}^{(k)} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \big(-\Phi(\boldsymbol{\theta}) + \langle \mathbf{S}_k, \Psi(\boldsymbol{\theta}) \rangle \big). \qquad \text{[vi.42]}$$

where $\mathbf{S}_k$ denotes the stochastic approximation on the sufficient statistics of the model, obtained in the "stochastic approximation step" of the algorithm.

VI.2.3.2.4 Overview of the algorithm

The MCMC-SAEM algorithm, whose steps are detailed above, writes as follows:

Algorithm 5 The MCMC-SAEM algorithm.

Require: Data $\mathbf{y}$, initial guess $\boldsymbol{\theta}^{(0)}$

Ensure: ML or MAP estimate $\boldsymbol{\theta}$ of the parameters of the model

1: Initializations: $\mathbf{z} \leftarrow \mathbf{0}$, $\mathbf{S}_0 \leftarrow \mathbf{0}$, $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)}$ and step-sizes $(\varepsilon_k)_{k\geq 0}$.

2: **repeat**

3: **Simulation** using a Metropolis-Hastings-within Gibbs sampler:

$$\mathbf{z}^{(k)} \sim \boldsymbol{\pi}_{\mathbf{y},\boldsymbol{\theta}^{(k-1)}}(\mathbf{z}^{(k-1)}, \cdot)$$

4: **Stochastic approximation**:

$$\mathbf{S}_k \leftarrow \mathbf{S}_k + \varepsilon_k \big(\mathbf{S}(\mathbf{y}, \mathbf{z}^{(k)}) - \mathbf{S}_{k-1} \big)$$

5: **Maximization**:

$$\boldsymbol{\theta}^{(k)} \leftarrow \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \big(-\Phi(\boldsymbol{\theta}) + \langle \mathbf{S}_k, \Psi(\boldsymbol{\theta}) \rangle \big)$$

with closed-form updates given in VI.3.3.

6: **until** convergence.

VI.2.3.3 Full-Bayesian inference

Finally, another class of stochastic algorithms can be considered for the statistical inference in nonlinear mixed-effects models. These algorithms, grouped under the term "Full-Bayesian methods" aim at learning the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$ and not only its modes. In Chapter III, we mentioned that MCMC samplers can be used to produce an ergodic Markov chain whose stationary distribution is the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$ of $\boldsymbol{\theta}$ given the observations $\mathbf{y}$. After a sufficiently large number of iterations, the sampler will produce samples which are approximately distributed as $q(\boldsymbol{\theta} \mid \mathbf{y})$. This approach differs from that of the MCMC-SAEM in the sense that the MCMC-SAEM aims at producing maximum a posteriori estimates, or in other words, "point estimates". Indeed, the MCMC-SAEM (as would the EM algorithm) will converge to a critical point of the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$, which is a local maximum (thanks to the randomness of the algorithm which avoids saddle points). Full-Bayesian methods draw a large number of samples approximately distributed as $q(\boldsymbol{\theta} \mid \mathbf{y})$ and then, these samples allow to construct approximations of the posterior using, for example, Kernel Density Estimates (KDE). But other informations on the posterior distribution can be derived from the samples.

Hamiltonian Monte Carlo (HMC) is a popular MCMC method which can be used for Bayesian inference. In [Hoffman and Gelman, 2014], the authors propose an adaptive HMC sampler called the *No U-Turns Sampler* (NUTS). This sampler is implemented in a R/C++ library called `STAN`. In this paper, the authors mention that the NUTS sampler offers much better performance in high-dimensional settings that classical samplers such as the Gibbs sampler or the Metropolis-Hastings algorithm. However, HMC samplers require to integrate a system of Hamiltonian equations and compute gradients of the posterior with respect to the parameters of the model, which can be computationally intensive.

VI.2.3.4 Other stochastic algorithms

In [Pinheiro, 1994], the author discusses the use of importance sampling (with Gaussian importance distribution) for inference in NLME models. This method is implemented in the `nlmixed` procedure of the SAS Software. In the work of Pinheiro, importance sampling is used to approximate the observed likelihood (see Eq. [vi.29]) of the model. Similarly, other Importance Sampling methods, such as Population Monte Carlo [Cappé et al., 2012], could be used to approximate the observed likelihood of the model and perform Bayesian inference. In [Pinheiro, 1994], the importance function is chosen to be a Gaussian distribution while, in Population Monte Carlo (PMC), the algorithm adapts the importance functions based on past samples. Even though importance sampling methods usually produce accurate results, in comparison to other methods to approximate the observed likelihood of the model, the computational cost of these methods tends to become prohibitive when the dimension of the space of latent variables

becomes "large". Importance sampling methods are not considered in this work.

Mean-field variational inference [Jordan et al., 1999, Consonni and Marin, 2007] aim at approximating the posterior $q(\boldsymbol{\theta} \mid \mathbf{y})$ through an optimization problem. The "best" approximating distribution $q(\boldsymbol{\theta})$ is estimated among a family of probability distributions $\mathcal{Q}$ by minimizing the Kullback-Leibler (KL) divergence $\mathrm{KL}\big(q(\boldsymbol{\theta}) \,||\, q(\boldsymbol{\theta} \mid \mathbf{y})\big)$. Since the KL divergence depends on the model evidence $q(\mathbf{y})$, which is usually not available in closed-form, these methods aim at minimizing a function called *Evidence Lower Bound* (ELBO), which is equal to the KL divergence up to a constant. To make the optimization problem tractable, the "mean-field assumption" consists in assuming that the approximating distribution $q(\boldsymbol{\theta})$ is a product of simple distributions, of the form $\prod_i q_i(\theta_i)$. Even though variational methods are usually faster than "Fully-Bayesian" (MCMC) methods, there is currently no theoretical guarantee of convergence for complex posterior distributions, such as the one we consider with the generic spatiotemporal model. As a result, we chose not to consider variational inference methods in this work.

VI.3 The MCMC-SAEM for the Bayesian generic spatiotemporal model

As mentioned in Section VI.2.3.2, the convergence of the MCMC-SAEM is proved for models which belong to the curved exponential family. In the following sections, we prove that the generic spatiotemporal model belongs to this curved exponential family and give the sufficient statistics of the model. In Section VI.3.6 we discuss the methodological challenges which arise when using the MCMC-SAEM algorithm in a Riemannian framework.

VI.3.1 Sufficient statistics

For the generic spatiotemporal model(Eq. [iv.66]), recall that the parameters are $\boldsymbol{\theta} = (\overline{\mathbf{p}_0}, \overline{\mathbf{v}_0}, \overline{t_0}, (\overline{\beta}_k)_{1\leq k\leq (N-1)N_s}, \boldsymbol{\theta}_{\text{var}})$, with $\boldsymbol{\theta}_{\text{var}} = (\sigma_\xi^2, \sigma_\tau^2, \sigma^2)$. The latent variables of the model are $\mathbf{z} = (\mathbf{z}_{\text{pop}}, (\mathbf{z}_i)_{1\leq i\leq p})$ with: $\mathbf{z}_{\text{pop}} = \big(\mathbf{p}_0, t_0, \mathbf{v}_0, (\beta_k)_{1\leq k\leq (N-1)N_s}\big)$ and $(\mathbf{z}_i)_{1\leq i\leq p} = \big(\xi_i, \tau_i, (s_{l,i})_{1\leq l\leq N_s}\big)_{1\leq i\leq p}\big)$. Let $\boldsymbol{\theta}_{\text{hyper}}$ denote the vector of *fixed* hyperparameters of the model. Recall that these hyperparameters are used to define the prior distribution q_{prior} in Eq. [iv.71]. For the generic spatiotemporal model, the joint likelihood $q(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}})$ writes:

$$\begin{aligned} q(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}) &= q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\theta}_{\text{hyper}}) q(\mathbf{z}, \boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}) \\ &= q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\theta}_{\text{hyper}}) q(\mathbf{z} \mid \boldsymbol{\theta}) q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}) \\ &= q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z}_{\text{pop}} \mid \boldsymbol{\theta}) q\big((\mathbf{z}_i)_{1\leq i\leq p} \mid \boldsymbol{\theta}\big) q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}) \\ &= q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}) q(\mathbf{z}_{\text{pop}} \mid \boldsymbol{\theta}) q\big((\mathbf{z}_i)_{1\leq i\leq p} \mid \boldsymbol{\theta}\big) q_{\text{prior}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{hyper}}) \end{aligned} \qquad \text{[vi.43]}$$

and it follows from modeling assumptions that:

$$q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}) q\big((\mathbf{z}_i)_{1\leq i\leq p} \mid \boldsymbol{\theta}\big) = \left(\prod_{\substack{1\leq i\leq p \\ 1\leq j\leq k_i}} q(\mathbf{y}_{i,j} \mid \mathbf{z}_i, \boldsymbol{\theta}) \right) \prod_{1\leq i\leq p} q(\mathbf{z}_i \mid \boldsymbol{\theta}). \qquad \text{[vi.44]}$$

Therefore,

$$
\begin{aligned}
\log q(\mathbf{y},\mathbf{z},\boldsymbol{\theta}\mid\boldsymbol{\theta}_{\text{hyper}}) = \; & \underbrace{-NK\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{\substack{1\le i\le p\\ 1\le j\le n_i}}\|\mathbf{y}_{i,j}-\eta^{\mathbf{w}_i}(\boldsymbol{\gamma}_0,\psi_i(t_{i,j}))\|^2}_{①} \\
& - p\log(\sigma_\xi\sqrt{2\pi}) - \frac{1}{2\sigma_\xi^2}\sum_{i=1}^{p}\xi_i^2 - p\log(\sigma_\tau\sqrt{2\pi}) - \frac{1}{2\sigma_\tau^2}\sum_{i=1}^{p}\tau_i^2 \\
& - \frac{pN_s}{2}\log(2\pi) \; \underbrace{-N\log(\sigma_{\mathbf{p}_0}\sqrt{2\pi}) - \frac{1}{2\sigma_{\mathbf{p}_0}^2}\|\mathbf{p}_0-\overline{\mathbf{p}_0}\|^2}_{②} \\
& - \log(\sigma_{t_0\sqrt{2\pi}}) - \frac{1}{2\sigma_{t_0}^2}(t_0-\overline{t_0})^2 - (N-1)N_s\log(\sigma_\beta\sqrt{2\pi}) \\
& \underbrace{-N\log(\sigma_{\mathbf{v}_0}\sqrt{2\pi}) - \frac{1}{2\sigma_{\mathbf{v}_0}^2}\|\mathbf{v}_0-\overline{\mathbf{v}_0}\|^2}_{③} - \frac{1}{2\sigma_\beta^2}\sum_{k=1}^{(N-1)N_s}(\beta_k-\overline{\beta}_k)^2 \\
& - \frac{1}{2}\sum_{\substack{1\le i\le p\\ 1\le l\le N_s}} s_{l,i}^2 - m_\xi\log(\sigma_\xi) - m_\xi\frac{\sigma_{\xi,0}^2}{2\sigma_\xi^2} - m_\tau\log(\sigma_\tau) - m_\tau\frac{\sigma_{\tau,0}^2}{2\sigma_\tau^2} \\
& - m_\sigma\log(\sigma) - m_\sigma\frac{\sigma_0^2}{2\sigma^2} \; \underbrace{-\frac{1}{2s_{\mathbf{p}_0}^2}\|\overline{\mathbf{p}_0}-\overline{\overline{\mathbf{p}_0}}\|^2}_{④} - \frac{1}{2s_{t_0}^2}(\overline{t_0}-\overline{\overline{t_0}})^2 \\
& \underbrace{-\frac{1}{2s_{\mathbf{v}_0}^2}\|\overline{\mathbf{v}_0}-\overline{\overline{\mathbf{v}_0}}\|^2}_{⑤} - \frac{1}{2s_{\boldsymbol{\beta}}^2}\|\overline{\boldsymbol{\beta}}\|^2 + C_{\text{prior}}\left[\!\left[\; -(N-1)\log(\sigma_\delta\sqrt{2\pi}) \right.\right. \\
& \left.\left. - \frac{1}{2\sigma_\delta^2}\sum_{k=1}^{N-1}(\delta_k-\overline{\delta_k})^2 \; \right]\!\right]
\end{aligned}
\qquad \text{[vi.45]}
$$

where k_i $(1 \le i \le p)$ denotes the number of time points for the ith individual and $K = \sum_{1\le i\le p} k_i$. In Eq. [vi.45], C_{prior} is the normalization constant of q_{prior}. Since it depends only on the *fixed* hyperparameters $\boldsymbol{\theta}_{\text{hyper}}$ of the model, C_{prior} is not written in closed-form. The term between double brackets in Eq. [vi.45] shall be considered only for the propagation model (see Eq. [v.13]). Note that for this model, the terms ① – ⑤

write:

$$
\begin{aligned}
① &= -\frac{1}{2\sigma^2}\sum_{\substack{1\leq i\leq p\\ 1\leq j\leq k_i}}\|\mathbf{y}_{i,j}-\boldsymbol{\eta}^{\mathbf{w}_i}(\boldsymbol{\gamma}_{0,\boldsymbol{\delta}},\psi_i(t_{i,j}))\|^2\\
② &= -\log(\sigma_{p_0}\sqrt{2\pi})-\frac{1}{2\sigma_{p_0}^2}(p_0-\overline{p_0})^2\\
③ &= -\log(\sigma_{v_0}\sqrt{2\pi})-\frac{1}{2\sigma_{v_0}^2}(v_0-\overline{v_0})^2\\
④ &= -\frac{1}{2s_{p_0}^2}(\overline{p_0}-\overline{\overline{p_0}})^2\\
⑤ &= -\frac{1}{2s_{v_0}^2}(\overline{v_0}-\overline{\overline{v_0}})^2.
\end{aligned}
\qquad \text{[vi.46]}
$$

Whereas, for the $\mathrm{Spd}(n)$ matrices model (Eq. [v.8]), the terms ① − ⑤ write:

$$
\begin{aligned}
① &= -\frac{n(n+1)}{2}K\log(\sigma\sqrt{2\pi})-\frac{1}{2\sigma^2}\sum_{\substack{1\leq i\leq p\\ 1\leq j\leq k_i}}\mathrm{tr}\Big(\big[\mathbf{Y}_{i,j}-\boldsymbol{\eta}^{\mathbf{W}_i}(\boldsymbol{\gamma}_0,\psi_i(t_{i,j}))\big]^2\Big)\\
② &= -\frac{n(n+1)}{2}\log(\sigma_{\mathbf{P}_0}\sqrt{2\pi})-\frac{1}{2\sigma_{\mathbf{P}_0}^2}\mathrm{tr}\Big(\big[\,\mathbf{P}_0-\overline{\mathbf{P}_0}\,\big]^2\Big)\\
③ &= -\frac{n(n+1)}{2}\log(\sigma_{\mathbf{V}_0}\sqrt{2\pi})-\frac{1}{2\sigma_{\mathbf{V}_0}^2}\mathrm{tr}\Big(\big[\,\mathbf{V}_0-\overline{\mathbf{V}_0}\,\big]^2\Big)\\
④ &= -\frac{1}{2s_{\mathbf{P}_0}^2}\mathrm{tr}\Big(\big[\,\overline{\mathbf{P}_0}-\overline{\overline{\mathbf{P}_0}}\,\big]^2\Big)\\
⑤ &= -\frac{1}{2s_{\mathbf{V}_0}^2}\mathrm{tr}\Big(\big[\,\overline{\mathbf{V}_0}-\overline{\overline{\mathbf{V}_0}}\,\big]^2\Big).
\end{aligned}
\qquad \text{[vi.47]}
$$

In order to prove that the generic spatiotemporal model belongs to the curved exponential family (see Eq. [vi.37]), let:

$$
\begin{aligned}
\mathbf{S}_1(\mathbf{y},\mathbf{z}) &= \left[\|\mathbf{y}_{i,j}\|^2\right]_{i,j} \in \mathbb{R}^K,\\
\mathbf{S}_2(\mathbf{y},\mathbf{z}) &= \left[\mathbf{y}_{i,j}^\top \eta^{\mathbf{w}_i}(\boldsymbol{\gamma}_0, \psi_i(t_{i,j}))\right]_{i,j} \in \mathbb{R}^K,\\
\mathbf{S}_3(\mathbf{y},\mathbf{z}) &= \left[\|\eta^{\mathbf{w}_i}(\boldsymbol{\gamma}_0, \psi_i(t_{i,j}))\|^2\right]_{i,j} \in \mathbb{R}^K,\\
\mathbf{S}_4(\mathbf{y},\mathbf{z}) &= \left[\xi_i^2\right]_i \in \mathbb{R}^p,\\
\mathbf{S}_5(\mathbf{y},\mathbf{z}) &= \left[\tau_i^2\right]_i \in \mathbb{R}^p,\\
\mathbf{S}_6(\mathbf{y},\mathbf{z}) &= \left[s_{l,i}^2\right]_{l,i} \in \mathbb{R}^{pN_s},\\
\mathbf{S}_7(\mathbf{y},\mathbf{z}) &= \mathbf{p}_0 \in \mathbb{R}^N,\\
S_8(\mathbf{y},\mathbf{z}) &= t_0 \in \mathbb{R},\\
\mathbf{S}_9(\mathbf{y},\mathbf{z}) &= \mathbf{v}_0 \in \mathbb{R}^N,\\
\mathbf{S}_{10}(\mathbf{y},\mathbf{z}) &= \left[\beta_k\right]_k \in \mathbb{R}^{(N-1)N_s},\\
S_{11}(\mathbf{y},\mathbf{z}) &= \|\mathbf{p}_0\|^2 \in \mathbb{R},\\
S_{12}(\mathbf{y},\mathbf{z}) &= t_0^2 \in \mathbb{R},\\
S_{13}(\mathbf{y},\mathbf{z}) &= \|\mathbf{v}_0\|^2 \in \mathbb{R},\\
\mathbf{S}_{14}(\mathbf{y},\mathbf{z}) &= \left[\beta_k^2\right]_k \in R^{(N-1)N_s},\\
\left[\left[\ \mathbf{S}_{15}(\mathbf{y},\mathbf{z}) = \left[\delta_k\right]_k \in \mathbb{R}^{N-1},\ \right]\right],&\\
\left[\left[\ \mathbf{S}_{16}(\mathbf{y},\mathbf{z}) = \left[\delta_k^2\right]_k \in \mathbb{R}^{N-1}\ \right]\right].&
\end{aligned}
\qquad [\text{vi.48}]
$$

For the propagation model, the sufficient statistic $\mathbf{S}_2$ and $\mathbf{S}_3$ are defined by: $\mathbf{S}_2(\mathbf{y},\mathbf{z}) = \left[\mathbf{y}_{i,j}^\top \boldsymbol{\eta}^{\mathbf{w}_i}(\boldsymbol{\gamma}_{0,\boldsymbol{\delta}}, \psi_i(t_{i,j}))\right]_{i,j}$ and $\mathbf{S}_3(\mathbf{y},\mathbf{z}) = \left[\|\boldsymbol{\eta}^{\mathbf{w}_i}(\boldsymbol{\gamma}_{0,\boldsymbol{\delta}}, \psi_i(t_{i,j}))\|^2\right]_{i,j}$. The sufficient statistics $\mathbf{S}_{15}$ and $\mathbf{S}_{16}$ shall be considered only for this propagation model. Regarding the $\mathrm{Spd}(n)$ matrices model, the sufficient statistics $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_3$, $\mathbf{S}_7$, $\mathbf{S}_9$, $\mathbf{S}_{11}$ and

$\mathbf{S}_{13}$ are defined by:

$$
\begin{aligned}
\mathbf{S}_1(\mathbf{Y},\mathbf{z}) &= \sum_{i,j} \mathbf{Y}_{i,j}^2,\\
\mathbf{S}_2(\mathbf{Y},\mathbf{z}) &= \sum_{i,j} \mathbf{Y}_{i,j}^\top \boldsymbol{\eta}^{\mathbf{W}_i}(\boldsymbol{\gamma}_0, \psi_i(t_{i,j})),\\
\mathbf{S}_3(\mathbf{Y},\mathbf{z}) &= \sum_{i,j} \big(\, \boldsymbol{\eta}^{\mathbf{W}_i}(\boldsymbol{\gamma}_0, \psi_i(t_{i,j}))\,\big)^2,\\
\mathbf{S}_7(\mathbf{Y},\mathbf{z}) &= \mathbf{P}_0,\\
\mathbf{S}_9(\mathbf{Y},\mathbf{z}) &= \mathbf{V}_0,\\
\mathbf{S}_{11}(\mathbf{Y},\mathbf{z}) &= \mathbf{P}_0^2,\\
\mathbf{S}_{13}(\mathbf{Y},\mathbf{z}) &= \mathbf{V}_0^2.
\end{aligned}
\qquad \text{[vi.49]}
$$

For $n \in \mathbb{N}^*$, let $\mathbb{1}_n$ denote the vector in $\mathbb{R}^n$ with all its components equal to 1 and I_n be the $n \times n$ identity matrix. For the generic spatiotemporal model, the inner product $\langle \mathbf{S}(\mathbf{y},\mathbf{z}), \Psi(\boldsymbol{\theta})\rangle$ is given by:

$$
\begin{aligned}
\langle \mathbf{S}(\mathbf{y},\mathbf{z}), \Psi(\boldsymbol{\theta})\rangle = {} & \big(\mathbf{S}_1(\mathbf{y},\mathbf{z}) - 2\mathbf{S}_2(\mathbf{y},\mathbf{z}) + \mathbf{S}_3(\mathbf{y},\mathbf{z})\big)^\top \Big(-\frac{1}{2\sigma^2}\mathbb{1}_K\Big)\\
& + \mathbf{S}_4(\mathbf{y},\mathbf{z})^\top \Big(-\frac{1}{2\sigma_\xi^2}\mathbb{1}_p\Big) + \mathbf{S}_5(\mathbf{y},\mathbf{z})^\top \Big(-\frac{1}{2\sigma_\tau^2}\mathbb{1}_p\Big)\\
& + \mathbf{S}_6^\top(\mathbf{y},\mathbf{z})\Big(-\frac{1}{2}\mathbb{1}_{pN_s}\Big) + S_{11}(\mathbf{y},\mathbf{z})\Big(-\frac{1}{2\sigma_{\mathbf{p}_0}^2}\Big)\\
& + \mathbf{S}_7(\mathbf{y},\mathbf{z})^\top \Big(\frac{1}{\sigma_{\mathbf{p}_0}^2}\overline{\mathbf{p}_0}\Big) + S_{12}(\mathbf{y},\mathbf{z})\Big(-\frac{1}{2\sigma_{t_0}^2}\Big)\\
& + S_8(\mathbf{y},\mathbf{z})\Big(\frac{1}{\sigma_{t_0}^2}\overline{t_0}\Big) + S_{13}(\mathbf{y},\mathbf{z})\Big(-\frac{1}{2\sigma_{\mathbf{v}_0}^2}\Big) + \mathbf{S}_9(\mathbf{y},\mathbf{z})^\top \Big(\frac{1}{\sigma_{\mathbf{v}_0}^2}\overline{\mathbf{v}_0}\Big)\\
& + \mathbf{S}_{14}(\mathbf{y},\mathbf{z})^\top \Big(-\frac{1}{2\sigma_\beta^2}\mathbb{1}_{(N-1)N_s}\Big) + \mathbf{S}_{10}(\mathbf{y},\mathbf{z})^\top \Big(\frac{1}{\sigma_\beta^2}\overline{\boldsymbol{\beta}}\Big)\\
& \Big[\Big[+ \mathbf{S}_{16}(\mathbf{y},\mathbf{z})^\top \Big(\frac{-1}{2\sigma_\delta^2}\mathbb{1}_{N-1}\Big) + \mathbf{S}_{15}(\mathbf{y},\mathbf{z})^\top \Big(\frac{1}{\sigma_\delta^2}\overline{\boldsymbol{\delta}}\Big)\ \Big]\Big]
\end{aligned}
\qquad \text{[vi.50]}
$$

with $\overline{\boldsymbol{\beta}} = \big[\overline{\beta}_k\big]_k$, $\overline{\boldsymbol{\delta}} = \big[\overline{\delta}_k\big]_{1\le k\le N-1}$. The two terms between double brackets in Eq. [vi.50] shall be considered only for the propagation model. Regarding the $\mathrm{Spd}(n)$ matrices

model, the dot product $\langle \mathbf{S}(\mathbf{Y},\mathbf{z}), \Psi(\boldsymbol{\theta})\rangle$ write:

$$\begin{aligned}\langle \mathbf{S}(\mathbf{Y},\mathbf{z}), \Psi(\boldsymbol{\theta})\rangle = &\Big\langle \mathbf{S}_1(\mathbf{Y},\mathbf{z}) - 2\mathbf{S}_2(\mathbf{Y},\mathbf{z}) + \mathbf{S}_3(\mathbf{Y},\mathbf{z}), -\frac{1}{2\sigma^2}\mathrm{I}_n\Big\rangle_F \\ &+ \mathbf{S}_4(\mathbf{Y},\mathbf{z})^\top\Big(-\frac{1}{2\sigma_\xi^2}\mathbb{1}_p\Big) + \mathbf{S}_5(\mathbf{Y},\mathbf{z})^\top\Big(-\frac{1}{2\sigma_\tau^2}\mathbb{1}_p\Big) \\ &+ \mathbf{S}_6^\top(\mathbf{Y},\mathbf{z})\Big(-\frac{1}{2}\mathbb{1}_{pN_s}\Big) + \Big\langle \mathbf{S}_{11}(\mathbf{Y},\mathbf{z}), -\frac{1}{2\sigma_{\mathbf{P}_0}^2}\mathrm{I}_n\Big\rangle_F \\ &+ \Big\langle \mathbf{S}_7(\mathbf{Y},\mathbf{z}), \frac{1}{\sigma_{\mathbf{P}_0}^2}\overline{\mathbf{P}_0}\Big\rangle_F + S_{12}(\mathbf{Y},\mathbf{z})\Big(-\frac{1}{2\sigma_{t_0}^2}\Big) \\ &+ S_8(\mathbf{Y},\mathbf{z})\Big(\frac{1}{\sigma_{t_0}^2}\overline{t_0}\Big) + \Big\langle \mathbf{S}_{13}(\mathbf{Y},\mathbf{z}), -\frac{1}{2\sigma_{\mathbf{V}_0}^2}\mathrm{I}_n\Big\rangle_F \\ &+ \Big\langle \mathbf{S}_9(\mathbf{Y},\mathbf{z}), \frac{1}{\sigma_{\mathbf{V}_0}^2}\overline{\mathbf{V}_0}\Big\rangle_F + \mathbf{S}_{14}(\mathbf{y},\mathbf{z})^\top\Big(-\frac{1}{2\sigma_\beta^2}\mathbb{1}_{(N-1)N_s}\Big) \\ &+ \mathbf{S}_{10}(\mathbf{y},\mathbf{z})^\top\Big(\frac{1}{\sigma_\beta^2}\overline{\boldsymbol{\beta}}\Big)\end{aligned} \qquad \text{[vi.51]}$$

where $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius inner product defined by:
$\forall(\mathbf{A},\mathbf{B}) \in \mathrm{Mat}_n(\mathbb{R})$, $\langle \mathbf{A},\mathbf{B}\rangle_F = \mathrm{tr}\big(\mathbf{A}^\top\mathbf{B}\big)$.

For the generic spatiotemporal model, the term $\Phi(\boldsymbol{\theta})$ writes:

$$\begin{aligned}\Phi(\boldsymbol{\theta}) = &-(NK+m_\sigma)\log(\sigma) - (p+m_\xi)\log(\sigma_\xi) - (p+m_\tau)\log(\sigma_\tau) \\ &\underbrace{-\frac{1}{2\sigma_{\mathbf{p}_0}^2}\|\overline{\mathbf{p}_0}\|^2 - \frac{1}{2s_{\mathbf{p}_0}^2}\|\overline{\mathbf{p}_0} - \overline{\overline{\mathbf{p}_0}}\|^2}_{⑥} - \frac{1}{2\sigma_{t_0}^2}\overline{t_0}^2 - \frac{1}{2s_{t_0}^2}(\overline{t_0} - \overline{\overline{t_0}})^2 \\ &\underbrace{-\frac{1}{2\sigma_{\mathbf{v}_0}^2}\|\overline{\mathbf{v}_0}\|^2 - \frac{1}{2s_{\mathbf{v}_0}^2}\|\overline{\mathbf{v}_0} - \overline{\overline{\mathbf{v}_0}}\|^2}_{⑦} - \frac{1}{2\sigma_\beta^2}\|\overline{\boldsymbol{\beta}}\|^2 - \frac{1}{2s_{\boldsymbol{\beta}}^2}\|\overline{\boldsymbol{\beta}}\|^2 - \frac{1}{2}\big(NK+2p \\ &+ pN_s + 2N + 1 + (N-1)N_s\big)\log(2\pi) - m_\xi\frac{\sigma_{\xi,0}^2}{2\sigma_\xi^2} - m_\tau\frac{\sigma_{\tau,0}^2}{2\sigma_\tau^2} - \\ &m_\sigma\frac{\sigma_0^2}{2\sigma^2} + C_{\text{prior}}\Big[\Big[-\frac{1}{2}(N-1)\log(2\pi) - \frac{1}{2\sigma_\delta^2}\sum_{k=1}^{N-1}\overline{\delta_k}^2\Big]\Big]\end{aligned} \qquad \text{[vi.52]}$$

where the terms between double brackets are to be considered only for the propagation model. With the $\mathrm{Spd}(n)$ matrices model, the terms ⑥ and ⑦ write:

$$
\begin{aligned}
⑥ &= -\frac{1}{2\sigma^2_{\mathbf{P}_0}}\mathrm{tr}\big(\overline{\mathbf{P}_0}^2\big) - \frac{1}{2s^2_{\mathbf{P}_0}}\mathrm{tr}\Big(\big[\,\overline{\mathbf{P}_0} - \overline{\overline{\mathbf{P}_0}}\,\big]^2\Big) \\
⑦ &= -\frac{1}{2\sigma^2_{\mathbf{V}_0}}\mathrm{tr}\big(\overline{\mathbf{V}_0}^2\big) - \frac{1}{2s^2_{\mathbf{V}_0}}\mathrm{tr}\Big(\big[\,\overline{\mathbf{V}_0} - \overline{\overline{\mathbf{V}_0}}\,\big]^2\Big).
\end{aligned}
\qquad \text{[vi.53]}
$$

Given the expressions of the sufficient statistics Eq. [vi.48], the stochastic approximation step of the MCMC-SAEM can be conducted as written in Section VI.2.3.2.2. The next section discusses sampling strategies for the sampling step of the MCMC-SAEM.

VI.3.2 On the sampling step of the MCMC-SAEM

In this section, we derive a Block Metropolis-Hastings-within-Gibbs (Block MHwG) sampler for the sampling step of the MCMC-SAEM. Each Metropolis-Hastings step of the algorithm consists in a multivariate symmetric random walk. The Block MHwG sampler updates simultaneously *block* (or sets) of latent variables then, at each iteration, each block is updated conditionally on the others. Even though the latent variables can be grouped in several ways, we chose to group the latent variables as follows: $\{\mathbf{z}_{\text{pop}}\}$ and $\{\mathbf{z}_i\}_{1\leq i\leq p}$. This grouping being given by the hierarchical structure of the model. Note that the latent variables also could have been grouped as follows: $\{\mathbf{p}_0, t_0, \mathbf{v}_0\}$, $\{(\beta_{l,k})_{l,k}\}$ and $\{\mathbf{z}_i\}_{1\leq i\leq p}$. In the case of the propagation models, the delay variables $(\delta_k)_{1\leq k\leq N-1}$ were grouped with $\mathbf{z}_{\text{pop}}$, although they could also be considered as a block in itself. For each block, the proposal in the Metropolis-Hastings step is chosen to be a multivariate Gaussian distribution centered at the current state of the block. Each variance-covariance matrix of a proposal distribution is chosen to be diagonal matrix: $\mathbf{D}_{\text{pop}} = \mathrm{Diag}\big(\zeta^2_{\mathbf{p}_0}\mathrm{I}_N, \zeta^2_{t_0}, \zeta^2_{\mathbf{v}_0}\mathrm{I}_N, \zeta^2_{\boldsymbol{\beta}}\mathrm{I}_{(N-1)N_s}\big)$ for the proposal distribution associated to $\mathbf{z}_{\text{pop}}$ and $\mathbf{D}_{\text{indiv}} = \mathrm{Diag}\big(\zeta^2_\xi, \zeta^2_\tau, \zeta^2_s\big)$ for the proposal distribution associated to $\mathbf{z}_i$ $(1 \leq i \leq p)$. The variances parameters $\zeta^2_{\mathbf{p}_0}, \zeta^2_{t_0}, \zeta^2_{\mathbf{v}_0}, \zeta^2_{\boldsymbol{\beta}}$ and $\zeta^2_\xi, \zeta^2_\tau$ are adjusted by hand to ensure an average acceptance rate for each block around 23% [Roberts et al., 1997]. The Block MHwG sampler is described in Algorithm 6. Let $\boldsymbol{\theta}_{\text{hyper}} = (\sigma^2_{\mathbf{p}_0}, \sigma^2_{t_0}, \sigma^2_{\mathbf{v}_0}, \sigma^2_{\boldsymbol{\beta}})$ denote the fixed hyperparameters which appear in the probability distribution of the latent variables in $\mathbf{z}_{\text{pop}}$.

Algorithm 6 The Block Metropolis-Hastings-within-Gibbs sampler

Require: Set of latent variables $\mathbf{z}^{(k-1)} = \big(\mathbf{z}_{\text{pop}}^{(k-1)}, (\mathbf{z}_i^{(k-1)})_{1\leq i\leq p}\big)$, current estimate of the parameters $\boldsymbol{\theta}^{(k-1)}$, variance-covariance matrices $\mathbf{D}_{\text{pop}}$ and $(\mathbf{D}_i)_{1\leq i\leq p}$ and $\boldsymbol{\theta}_{\text{hyper}}$

Ensure: Set of latent variables $\mathbf{z}^{(k)}$

Block $\mathbf{z}_{\text{pop}}^{(k)}$ of population latent variables:

Draw a candidate $\mathbf{z}_{\text{pop}}^{*} \sim \mathcal{N}\big(\mathbf{z}_{\text{pop}}^{(k-1)}, \mathbf{D}_{\text{pop}}\big)$

Compute the acceptance ratio $\alpha(\mathbf{z}_{\text{pop}}^{(k-1)}, \mathbf{z}_{\text{pop}}^{*})$ defined by:

$$\alpha(\mathbf{z}_{\text{pop}}^{(k-1)}, \mathbf{z}_{\text{pop}}^{*}) = \frac{q(\mathbf{y} \mid \mathbf{z}_{\text{pop}}^{*}, (\mathbf{z}_i^{(k-1)})_{1\leq i\leq p}, \boldsymbol{\theta}^{(k-1)}) q_{\text{pop}}(\mathbf{z}_{\text{pop}}^{*} \mid \boldsymbol{\theta}^{(k-1)})}{q(\mathbf{y} \mid \mathbf{z}_{\text{pop}}^{(k-1)}, (\mathbf{z}_i^{(k-1)})_{1\leq i\leq p}, \boldsymbol{\theta}^{(k-1)}) q_{\text{pop}}(\mathbf{z}_{\text{pop}}^{(k-1)} \mid \boldsymbol{\theta}^{(k-1)})} \wedge 1.$$

Draw $U \sim \text{Uniform}\big([0,1]\big)$

Set:

$$\mathbf{z}_{\text{pop}}^{(k)} = \begin{cases} \mathbf{z}_{\text{pop}}^{(k-1)} & \text{if } U \leq \alpha(\mathbf{z}_{\text{pop}}^{(k-1)}, \mathbf{z}_{\text{pop}}^{*}) \\ \mathbf{z}_{\text{pop}}^{*} & \text{otherwise.} \end{cases}$$

for $i = 1 \dots p$ **do**

  Blocks $(\mathbf{z}_i^{(k)})_{1\leq i\leq p}$ of individual latent variables:

  Draw a candidate $\mathbf{z}_i^{*} \sim \mathcal{N}\big(\mathbf{z}_i^{(k-1)}, \mathbf{D}_{\text{indiv}}\big)$

  Compute the acceptance ratio $\alpha(\mathbf{z}_i^{(k-1)}, \mathbf{z}_i^{*})$ defined by:

$$\alpha(\mathbf{z}_i^{(k-1)}, \mathbf{z}_i^{*}) = \frac{q(\mathbf{y} \mid \mathbf{z}_{\text{pop}}^{(k)}, \mathbf{z}_{-i}^{(k-1),(k)}, \mathbf{z}_i^{*}, \boldsymbol{\theta}^{(k-1)}) q_i(\mathbf{z}_i^{*} \mid \boldsymbol{\theta}^{(k-1)})}{q(\mathbf{y} \mid \mathbf{z}_{\text{pop}}^{(k)}, \mathbf{z}_{-i}^{(k-1),(k)}, \mathbf{z}_i^{(k-1)}, \boldsymbol{\theta}^{(k-1)}) q_i(\mathbf{z}_i^{(k-1)} \mid \boldsymbol{\theta}^{(k-1)})} \wedge 1.$$

  Draw $U \sim \text{Uniform}([0,1])$

  Set:

$$\mathbf{z}_i^{(k)} = \begin{cases} \mathbf{z}_i^{*} & \text{if } U \leq \alpha(\mathbf{z}_i^{(k-1)}, \mathbf{z}_i^{*}) \\ z_i^{(k-1)} & \text{otherwise.} \end{cases}$$

end for

Return: $\mathbf{z}^{(k)} = \big(\mathbf{z}_{\text{pop}}^{(k)}, (\mathbf{z}_i^{(k)})_{1\leq i\leq p}\big)$.

Let $i \in \{1, \dots, p\}$ and $q_{\text{pop}}\big(\cdot \mid \boldsymbol{\theta}\big)$ (respectively $q_i(\cdot \mid \boldsymbol{\theta})$ denote the density function of the joint distribution of the latent variables $\mathbf{z}_{\text{pop}}$ (respectively $\mathbf{z}_i$) as specified in the

generative model (Eq. [iv.67] and Eq. [iv.68]):

$$q_{\text{pop}}(\mathbf{z}_{\text{pop}} \mid \boldsymbol{\theta}) \propto \exp\Big(-\frac{1}{2\sigma_{\mathbf{p}_0}^2}\|\mathbf{p}_0 - \overline{\mathbf{p}_0}\|^2\Big)\exp\Big(-\frac{1}{2\sigma_{t_0}^2}(t_0 - \overline{t_0})^2\Big) \\ \exp\Big(-\frac{1}{2\sigma_{\mathbf{v}_0}^2}\|\mathbf{v}_0 - \overline{\mathbf{v}_0}\|^2\Big)\exp\Big(-\frac{1}{2\sigma_{\boldsymbol{\beta}}^2}\|\boldsymbol{\beta} - \overline{\boldsymbol{\beta}}\|^2\Big) \qquad \text{[vi.54]}$$

and

$$q_i(\mathbf{z}_i \mid \boldsymbol{\theta}) \propto \exp\Big(-\frac{1}{2\sigma_\xi^2}\xi_i^2\Big)\exp\Big(-\frac{1}{2\sigma_\tau^2}\tau_i^2\Big)\exp\Big(-\frac{1}{2}\|\mathbf{s}_i\|^2\Big) \qquad \text{[vi.55]}$$

with: $\boldsymbol{\beta} = [\beta_{l,k}]_{1\leq l\leq N_s, 1\leq k\leq N-1}$ and for all $i \in \{1,\ldots,p\}$, $\mathbf{s}_i = [s_{l,i}]_{1\leq l\leq N_s}$. The probability distributions q_{pop} and q_i $(1 \leq i \leq p)$ are given up to a constant. Indeed, the normalizing constant of q_{pop} or q_i $(1 \leq i \leq p)$ depends only on the parameters $\boldsymbol{\theta}$. Therefore, these constants can be omitted for the computation of the acceptance ratio in Algorithm 6. For $k \in \mathbb{N}^*$ and $i \in \{1,\ldots,p\}$, $z_{-i}^{(k-1),(k)}$ denotes: $\big(z_1^{(k)},\ldots,z_{i-1}^{(k)},z_{i+1}^{(k-1)},\ldots,z_p^{(k-1)}\big)$.

VI.3.2.1 Discussion

In the sampling step of the MCMC-SAEM, the computation of the model likelihood $q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta})$, defined by:

$$\log q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta}) = -\frac{1}{2\sigma^2}\sum_{\substack{1\leq i\leq p\\ 1\leq j\leq k_i}} \big\|\mathbf{y}_i - \boldsymbol{\eta}^{\mathbf{w}_i}\big(\boldsymbol{\gamma}_0, \psi_i(t_{i,j})\big)\big\|^2 \qquad \text{[vi.56]}$$

is computationally costly for large datasets. The runtime of the MCMC-SAEM on specific examples is detailed in Section VI.4.1. Since computing the log-likelihood $q(\mathbf{y} \mid \mathbf{z}, \boldsymbol{\theta})$ is expensive, it is preferable use a sampler which requires few computation of this log-likelihood. As a consequence, the Block MHwG sampler is computationally more interesting than a deterministic (or "one-at-a-time") MHwG sampler as the algorithm proposed above requires $2+2p$ computations of the model likelihood against a minimum of $4+(N-1)(N_s+1)+(3+N_s)p$ for the "one-at-a-time" Gibbs sampler. In Algorithm 6, the variance-covariance matrices of the proposal distributions are diagonal and the variance coefficients on the diagonal are tuned by hand. Using adaptive sampling algorithms, such as the one described in [Atchadé, 2006], the variance parameters on the diagonal could be automatically adjusted by the algorithm. Finally, note that the steps `7:` to `12:` in Algorithm 6 could be done in parallel, for large datasets.

VI.3.3 On the maximization step of the MCMC-SAEM

The maximization step of the MCMC-SAEM consists, at the kth iteration, in solving the following optimization problem:

$$\boldsymbol{\theta}^{(k)} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\big(-\Phi(\boldsymbol{\theta}) + \langle \mathbf{S}_k, \Psi(\boldsymbol{\theta})\rangle\big). \qquad \text{[vi.57]}$$

where $\mathbf{S}_k$ denotes the stochastic approximation on the sufficient statistics of the model obtained from the "stochastic approximation step". Recall, from Section IV.3.4, that:

$$\Theta = \Big\{\boldsymbol{\theta} = (\overline{\mathbf{p}_0}, \overline{\mathbf{v}_0}, \overline{t_0}, (\overline{\beta}_{l,k})_{l,k}, \boldsymbol{\theta}_{\text{var}}) \,\big/ (\overline{\mathbf{p}_0}, \overline{\mathbf{v}_0}) \in \mathrm{T}\mathbb{M},\ \overline{t_0} \in \mathbb{R}, \\ (\overline{\beta}_{l,k})_{l,k} \in \mathbb{R}^{(N-1)N_s},\ \boldsymbol{\theta}_{\text{var}} \in]0, +\infty[^3\Big\} \qquad \text{[vi.58]}$$

with $\boldsymbol{\theta}_{\text{var}} = (\sigma_\xi^2, \sigma_\tau^2, \sigma^2)$. Since we assumed that $\mathbb{M}$ is a convex open subset of $\mathbb{R}^N$, for all $\mathbf{p} \in \mathbb{M}$, one has the following identification: $\mathrm{T}_{\mathbf{p}}\mathbb{M} = \mathbb{R}^N$. Moreover, as $\mathbb{M} \subset \mathbb{R}^N$, one can consider that the function $\mathbf{Q}_k$, defined by $\forall \boldsymbol{\theta} \in \Theta$, $\mathbf{Q}_k(\boldsymbol{\theta}) = -\Phi(\boldsymbol{\theta}) + \langle \mathbf{S}_k, \Psi(\boldsymbol{\theta}) \rangle$, is defined and differentiable on an open subset of the Euclidean space. As a consequence, $\mathbf{Q}_k$ is maximized by looking for its critical points.

Computing the gradient of $\mathbf{Q}_k$ shows that there is a unique critical point in Θ. Therefore, the maximization step of the MCMC-SAEM for the generic spatiotemporal model writes:

$$\begin{aligned}
\overline{\mathbf{p}_0}^{(k+1)} &= \Big(\frac{1}{s^2_{\mathbf{p}_0}} + \frac{1}{\sigma^2_{\mathbf{p}_0}}\Big)^{-1}\Big(\frac{1}{\sigma^2_{\mathbf{p}_0}}\mathbf{S}_7(\mathbf{y}, \mathbf{z}^{(k)}) + \frac{1}{s^2_{\mathbf{p}_0}}\overline{\overline{\mathbf{p}_0}}\Big), \\
\overline{\mathbf{v}_0}^{(k+1)} &= \Big(\frac{1}{s^2_{\mathbf{v}_0}} + \frac{1}{\sigma^2_{\mathbf{v}_0}}\Big)^{-1}\Big(\frac{1}{\sigma^2_{\mathbf{v}_0}}\mathbf{S}_9(\mathbf{y}, \mathbf{z}^{(k)}) + \frac{1}{s^2_{\mathbf{v}_0}}\overline{\overline{\mathbf{v}_0}}\Big), \\
\overline{t_0}^{(k+1)} &= \Big(\frac{1}{s^2_{t_0}} + \frac{1}{\sigma^2_{t_0}}\Big)^{-1}\Big(\frac{1}{\sigma^2_{t_0}}S_8(\mathbf{y}, \mathbf{z}^{(k)}) + \frac{1}{s^2_{t_0}}\overline{\overline{t_0}}\Big), \\
\overline{\beta}_{l,j}^{(k+1)} &= \Big(\frac{1}{s^2_{\boldsymbol{\beta}}} + \frac{1}{\sigma^2_{\boldsymbol{\beta}}}\Big)^{-1}\Big(\frac{1}{\sigma^2_{\boldsymbol{\beta}}}\mathbf{S}_{10}(\mathbf{y}, \mathbf{z}^{(k)})\Big) \\
(\sigma^2_\xi)^{(k+1)} &= \frac{1}{p + m_\xi}\Big(\mathbf{S}_4(\mathbf{y}, \mathbf{z}^{(k)})^\top \mathbb{1}_p + m_\xi \sigma^2_{\xi,0}\Big), \\
(\sigma^2_\tau)^{(k+1)} &= \frac{1}{p + m_\tau}\Big(\mathbf{S}_5(\mathbf{y}, \mathbf{z}^{(k)})^\top \mathbb{1}_p + m_\tau \sigma^2_{\tau,0}\Big), \\
(\sigma^2)^{(k+1)} &= \frac{1}{NK + m_\sigma}\Big(\big[\mathbf{S}_1(\mathbf{y}, \mathbf{z}^{(k)}) - 2\mathbf{S}_2(\mathbf{y}, \mathbf{z}^{(k)}) + \\
&\qquad \mathbf{S}_3(\mathbf{y}, \mathbf{z}^{(k)})\big]^\top \mathbb{1}_K + m_\sigma \sigma_0^2\Big), \\
\Big[\Big[\ (\overline{\delta}_j)_j^{(k+1)} &= \mathbf{S}_{15}(\mathbf{y}, \mathbf{z}^{(k)})\ \Big]\Big]
\end{aligned} \qquad \text{[vi.59]}$$

Eq. [vi.59] shows that each update can be interpreted as a barycenter between its corresponding sufficient statistic and the mean of its corresponding prior. For instance, in the update

$$\overline{\mathbf{p}_0}^{(k+1)} = \Big(\frac{1}{s^2_{\mathbf{p}_0}} + \frac{1}{\sigma^2_{\mathbf{p}_0}}\Big)^{-1}\Big(\frac{1}{\sigma^2_{\mathbf{p}_0}}\mathbf{S}_7(\mathbf{y}, \mathbf{z}^{(k)}) + \frac{1}{s^2_{\mathbf{p}_0}}\overline{\overline{\mathbf{p}_0}}\Big) \qquad \text{[vi.60]}$$

the larger the variance $s^2_{\mathbf{p}_0}$, the less the influence of the prior $\overline{\mathbf{p}_0}^{(k+1)}$. On the contrary, a narrow prior ($s^2_{\mathbf{p}_0}$ small) will force the parameter $\mathbf{p}_0$ to remain close to $\overline{\overline{\mathbf{p}_0}}$. The update between double brackets is to be considered only for the propagation models.

Note that the Riemannian metric on $\mathbb{M}$ does not appear in the parameter update of $\overline{\mathbf{p}_0}$. Because $\mathbb{M}$ is a convex open subset of $\mathbb{R}^N$, the proposed update (in Eq. [vi.60]) remains in $\mathbb{M}$ as long as, at the kth iteration, $\mathbf{p}_0^{(k)} = \mathbf{S}_7(\mathbf{y}, \mathbf{z}^{(k)})$ is in $\mathbb{M}$. In the previous section, we saw that, at each step, the latent variable $\mathbf{p}_0$ is updated with a symmetric random walk. If the variance of the proposal $\zeta^2_{\mathbf{p}_0}$ is "small enough", the assumption that $\mathbb{M}$ is *open* ensures that a small perturbation of a point in $\mathbb{M}$ remains in $\mathbb{M}$. In the case where $\mathbb{M} =]0, 1[$, the validation of the MCMC-SAEM in Section VI.4 and the experimental results in Chapter VII show that the parameter $\overline{p_0}$ remains within the range of the observations and is never estimated close to 0 or close to 1.

VI.3.4 Choice of the hyperparameters

As discussed in Section VI.1.1, the priors assumed for the generic spatiotemporal model (see Eq. [iv.71]) ensure the existence of a maximum *a posteriori*. In order to use the MCMC-SAEM, the hyperparameters $\overline{\overline{\mathbf{p}_0}}, \overline{\overline{t_0}}, \overline{\overline{\mathbf{v}_0}}, \sigma_{\xi,0}, \sigma_{\tau,0}, \sigma_0, m_\xi, m_\tau, m_\sigma, s_{\mathbf{p}_0}, s_{t_0}, s_{\mathbf{v}_0}, s_{\boldsymbol{\beta}}$ have to be specified. The prior distribution could also be used to constrain the parameters space and avoid identifiability problems. In this section, we describe heuristics to choose these hyperparameters, based only on the longitudinal dataset which is to be analyzed with the MCMC-SAEM.

In a naive heuristic, the hyperparameter $\overline{\overline{\mathbf{p}_0}}$ can be chosen to be the median of the observations. In order to ensure that $\overline{\overline{\mathbf{p}_0}}$ belongs to the manifold $\mathbb{M}$, the notion of median shall be generalized to smooth manifolds. On a Riemannian manifold, the median can be defined as the minimizer of the sum of geodesic distances to the data points [Fletcher et al., 2009]. In practice, the hyperparameter $s_{\mathbf{p}_0}$ is chosen "large" ($s_{\mathbf{p}_0} \simeq 1$) to ensure that the prior on $\overline{\mathbf{p}_0}$ is not informative and does not force $\overline{\mathbf{p}_0}$ to remain in a close neighborhood of $\overline{\overline{\mathbf{p}_0}}$. For the other hyperparameters, consider $\overline{\overline{\mathbf{p}_0}} \in \mathbb{M}$ fixed and fit a geodesic of the form $\overline{\overline{\boldsymbol{\gamma}}}_{0,i}(\cdot) = \mathrm{Exp}_{\overline{\overline{\mathbf{p}_0}}, t_i}(\mathbf{v}_i)(\cdot)$ to the individual observations $(\mathbf{y}_{i,j}, t_{i,j})_{1 \leq j \leq k_i}$ for each individual: this is typically p independent regression problems, where the observations of each individual are regressed against age. Each fit can be done by minimizing a *nonlinear* least-squares criterion which writes:

$$\forall 1 \leq i \leq p, \ (t_i, v_i) = \underset{t \in \mathbb{R}, \mathbf{v} \in \mathbb{R}^N}{\mathrm{argmin}} \sum_{j=1}^{k_i} \| \mathbf{y}_{i,j} - \mathrm{Exp}_{\overline{\overline{\mathbf{p}_0}}, t}(\mathbf{v})(t_{i,j}) \|^2 \qquad \text{[vi.61]}$$

This nonlinear least squares criterion is minimized using gradient descent, where the gradient of the objective function is approximated numerically using central finite differences. For each individual, t_i (respectively $\mathbf{v}_i$) corresponds to the time at which (respectively the speed at which) the fitted trajectory goes through $\overline{\overline{\mathbf{p}_0}}$. Naturally,

$\overline{\overline{t_0}}$ can be chosen to be the median of the times $(t_i)_{1\leq i\leq p}$ and $\overline{\overline{\mathbf{v}_0}}$ the median of the speeds $(\mathbf{v}_i)_{1\leq i\leq p}$. Similarly, the associated variance hyperparameters s_{t_0} and $s_{\mathbf{v}_0}$ are chosen to provide narrow or non-informative priors. The variance hyperparameter $s_{\boldsymbol{\beta}}$ is, by default, chosen equal to 1 in the experimental results and validations presented hereafter. In the simple heuristic presented above, we choose to consider the median instead of the mean. This is because the median is known to be more robust to outliers than the mean.

In the case of the propagation model, recall that the observation $(\mathbf{y}_{i,j})_{1\leq i\leq p,\,1\leq j\leq k_i}$ are points in the product manifold $\mathbb{M} = M^N$, where M is a one-dimensional geodesically complete Riemannian manifold. For this model, the hyperparameter $\overline{\overline{p_0}}$ is initialized as the Riemannian median of the observations $\big((\mathbf{y}_{i,j})_1\big)_{1\leq i\leq p,\,1\leq j\leq k_i}$, where $(\mathbf{y}_{i,j})_1$ denotes the first component of $\mathbf{y}_{i,j}$. This choice is motivated by the specific form of the average trajectory in the propagation model. Similarly to the procedure described above, for each $k \in \{1,\ldots,N\}$, times $(t_i^k)_{1\leq i\leq p}$ (in years) and velocities $(v_i^k)_{1\leq i\leq p}$ are estimated by minimizing the least squares criterion

$$\forall 1 \leq k \leq N-1,\ (t_i^k, v_i^k) = \underset{t\in\mathbb{R}, v\in\mathbb{R}}{\operatorname{argmin}} \sum_{j=1}^{k_i} \big((\mathbf{y}_{i,j})_k - \operatorname{Exp}_{\overline{\overline{p_0}},t,v}(t_{i,j})\big)^2 \qquad \text{[vi.62]}$$

where $\operatorname{Exp}_{\overline{\overline{p_0}},t,v}(\cdot)$ denotes the geodesic of the one-dimensional manifold M, which goes through the point $\overline{\overline{p_0}} \in M$ at time t and with velocity v. The hyperparameter $\overline{\overline{t_0}}$ (respectively $\overline{\overline{v_0}}$) is initialized as the median of the times $(t_i^1)_{1\leq i\leq p}$ (respectively velocities $(v_i^1)_{1\leq i\leq p}$). For $k \in \{2,\ldots,N\}$, the hyperparameter $\overline{\overline{\delta}}_k$ is initialized as the difference between $\overline{\overline{t_0}}$ and the median of the times $(t_i^k)_{1\leq i\leq p}$.

Regarding the variance parameters, the hyperparameters $\sigma_{\xi,0}$, $\sigma_{\tau,0}$ and σ_0 really play an important role when the number of individuals is small. In that case, these hyperparameters bound below the variances σ_ξ^2, σ_τ^2, σ^2 and avoid variance parameters going to 0. But if the number of individuals is large enough, the variance parameters in the MCMC-SAEM are not expected to degenerate. The hyperparameters m_ξ, m_τ and m_ξ control the shape of the Inverse-Gamma prior on these variance parameters. The larger these hyperparameters are, the narrower the priors are. In the applications presented in this dissertation, these shape hyperparameters are chosen equal to 3, which corresponds to a wide prior.

VI.3.5 Stopping criterion and convergence assessment

VI.3.5.1 Impact of several variables on the overall runtime

We discuss below the impact of several variables on the runtime of the MCMC-SAEM. These variables are: the number of independent sources N_s, the dimension of the observations N and the number of individuals p. The variables N_s and N characterize

the total number of parameters to estimate. The number of individuals p influences (linearly) the number of latent variables in the model, and therefore the sampling step of the algorithm.

In the MCMC-SAEM algorithm (see Algorithm 5), the sampling step is, by far, the most computationally demanding step of the algorithm. Indeed, if the longitudinal dataset contains a large number of individuals and/or high-dimensional data, computing the model likelihood may have a high computational cost. Eq. [vi.29] shows that the logarithm of the model likelihood consists in a double sum where each term requires to compute a Riemannian exponential, the parallel transport of a tangent vector and an orthonormal basis (see Section IV.3.3.1).

For instance, with the (logistic curves or straight lines) propagation model Eq. [v.13], the MCMC-SAEM would have to estimate the following parameters: $\boldsymbol{\theta} = (\overline{p_0}, \overline{t_0}, \overline{v_0}, \overline{\delta}_1, \ldots, \overline{\delta}_{N-1}, \overline{\beta}_1, \ldots, \overline{\beta}_{(N-1)N_s}, \sigma_\eta, \sigma_\tau, \sigma)$. In this example, we see that the number of parameters to estimate is $6 + (N-1)(N_s+1)$. As the dimension N of the product manifold $\mathbb{M}$ increases, the number of parameters increases linearly. Moreover, as N increases, the number N_s of independent sources has a greater impact on the number of parameters to estimate. In a high-dimensional setting (N very large), considering a large number of independent components may be computationally very costly since each additional independent component requires to estimate $N-1$ additional parameters.

VI.3.5.2 Convergence monitoring

At the kth iteration of the algorithm, let $\boldsymbol{\theta}^{(k)}$ denote the current estimate of the parameters of the generic spatiotemporal model. For $1 \leq j \leq |\boldsymbol{\theta}^{(k)}|$, $\boldsymbol{\theta}_j^{(k)}$ denotes the jth coordinate of the vector $\boldsymbol{\theta}^{(k)}$. Monitoring the convergence of empirical averages defined by:

$$\left\{ \left(\frac{1}{k} \sum_{s=1}^{k} \boldsymbol{\theta}_j^{(s)} \right)_{k \geq 1} \right\}_j \qquad \text{[vi.63]}$$

can inform on whether the MCMC-SAEM has converged. To avoid biased averages, the empirical averages in Eq. [vi.63] should be computed on a moving window. Plotting the evolution of these empirical averages while the MCMC-SAEM is running provides a graphical mean of assessing the convergence of the algorithm. Another solution is to plot the evolution of the parameters $(\boldsymbol{\theta}^{(k)})_{k \geq 0}$. This visualization is proposed in MONOLIX. When the evolution of the parameters stabilizes, the MCMC-SAEM may have converged.

In [Booth and Hobert, 1999], the authors use a numerical stopping rule for their Monte Carlo EM (MCEM) algorithm. Given small positive constants c_1 and c_2 (for

example, $c_1 = 10^{-3}$ and $c_2 = 10^{-4}$), the algorithm is stopped as soon as

$$\max_j \left(\frac{|\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\theta}_j^{(k-1)}|}{|\boldsymbol{\theta}_j^{(k)}| + c_1} \right) < c_2 \qquad \text{[vi.64]}$$

is satisfied for several consecutive iterations.

In the following experiments, the MCMC-SAEM was run several times with various limits on the maximum number of iterations. If the MCMC-SAEM could not converge in the given number of iterations, the algorithm was run again from the previous estimates. Even though the plots of empirical averages, computed on a moving window, offers a way of assessing the convergence of the algorithm, deriving a generic and automatic convergence criterion would help to save computational time by stopping the algorithm at the right moment. However, proposing such an automatic stopping rule remains an open problem.

VI.3.6 Discussion

VI.3.6.1 Sampling and optimization on a Riemannian manifold

The choice of the proposal distribution in Section VI.3.2 and parameters updates in Section VI.3.3 make sense $\mathbb{M}$ is a convex open subset of the Euclidean space $\mathbb{R}^N$. However, if we no longer assume that $\mathbb{M}$ is open in $\mathbb{R}^N$, as it would the case for the sphere $\mathbb{S}^{n-1} \subset \mathbb{R}^n$, the multivariate Gaussian proposal distribution centered at the current state, used as proposal distribution in the Block MHwG sampler, would no longer make sense for the latent variable $\mathbf{p}_0$, constrained to remain on the sphere. The discussion below aims at proposing solutions to address this problem.

VI.3.6.1.1 Sampling step

In the block MHwG sampler (Algorithm 6), if $\mathbb{M} =]0, 1[$, the proposal distribution for $\mathbf{p}_0$ could be chosen to be a *logit-normal distribution* and if $\mathbb{M} = \mathbb{S}^2$, the proposal distribution could be the *Von Misses distribution*. These proposal distributions solutions are specific to these Riemannian manifolds. Even though considering these probability distributions would allow to sample on the Riemannian manifold, a more generic solution would be preferable. In the following, we consider a generalization of the Gaussian distribution to Riemannian manifolds.

Recall that for all $(\mathbf{p}, \mathbf{q}) \in \mathbb{M}$, the *Riemannian distance* on $\mathbb{M}$ is defined by: $d(\mathbf{p}, \mathbf{q}) = \|\mathrm{Log}_{\mathbf{p}}(\mathbf{q})\|_{\mathbf{p}}$, with $\mathrm{Log}_{\mathbf{p}}$, the Riemannian logarithm at $\mathbf{p} \in \mathbb{M}$. The *Riemannian Gaussian distribution* [Pennec, 2006], centered at $\overline{\mathbf{p}_0} \in \mathbb{M}$ and with variance σ^2, is the probability distribution on $\mathbb{M}$ whose density $q_{\mathbb{M}}(\cdot\,; \overline{\mathbf{p}_0}, \sigma^2)$ is given by:

$$\forall \mathbf{p} \in \mathbb{M},\ q_{\mathbb{M}}(\mathbf{p}\,; \overline{\mathbf{p}_0}, \sigma^2) \propto \exp\Big(-\frac{1}{2\sigma^2} d^2(\mathbf{p}, \overline{\mathbf{p}_0}) \Big). \qquad \text{[vi.65]}$$

As discussed in Section IV.3.5.2, the normalization constant of $q_{\mathbb{M}}(\cdot\,;\overline{\mathbf{p}_0},\sigma^2)$ depends, in general, on both the mean $\overline{\mathbf{p}_0} \in \mathbb{M}$ and the variance σ^2. However, in a Riemannian homogeneous space, the normalization constant in independent of the mean.

Note that the Riemannian manifolds considered in this thesis ($\mathbb{M} = \mathbb{R}^N$, $\mathbb{M} =]0,1[^N$ with $N \in \mathbb{N}^*$ or $\mathbb{M} = \mathrm{Spd}(n)$) are *Riemannian homogeneous spaces*. For each of these Riemannian manifolds, the group $\mathrm{Isom}(\mathbb{M})$ of the isometries (see Section III.1.2.2) of $\mathbb{M}$ acts transitively on $\mathbb{M}$. In other words, for any pair of distinct points on $\mathbb{M}$, there exist an isometry of the Riemannian manifold $\mathbb{M}$ which maps the first point onto the other. In the following section, we show that the Riemannian manifolds considered in this dissertation are Riemannian homogeneous spaces. It follows that, the Riemannian Gaussian distribution could be used to replace the Gaussian proposal for $\mathbf{p}_0$ in the sampling step of the MCMC-SAEM. Indeed, at the kth iteration of the MCMC-SAEM, a candidate $\mathbf{p}_0^*$ could be proposed as follows: $\mathbf{p}_0^* \sim q_{\mathbb{M}}(\cdot\,;\mathbf{p}_0^{(k-1)},\zeta^2_{\mathbf{p}_0})$. However, using this Riemannian Gaussian distribution as proposal distribution raises a difficulty: in general, this probability distribution does not belong to a known family of probability distributions. As a consequence, sampling from it may be difficult or impossible. To address this problem, a Metropolis-Hastings algorithm could be used to sample from this distribution.

In [Girolami and Calderhead, 2011], the authors proposed a Metropolis adjusted Langevin algorithm and Hamiltonian Monte Carlo algorithm to sample from probability distributions defined on Riemannian manifolds. However, these methods depend on the Fisher-Rao metric, which relies on the Hessian matrix of the target distribution. As a consequence, these methods could lead to heavy computations. In [Betancourt, 2013], the author proposes a new metric tensor for Riemannian Hamiltonian Monte Carlo methods, which is everywhere well-behaved and more practical to compute. Still, the proposed metric still relies on the Hessian matrix, which would be very costly to compute for the generic spatiotemporal model.

VI.3.6.1.2 The Riemannian manifolds $]0,1[$ and $\mathbb{S}^n$ are Riemannian homogeneous spaces

This section aims at proving that the Riemannian manifolds discussed above, namely the open interval $]0,1[$ equipped with the Riemannian metric defined in Eq. [iv.5] and the sphere $\mathbb{S}^n \subset \mathbb{R}^{n+1}$ equipped with the induced metric, are Riemannian homogeneous spaces.

Proposition VI.1. *The Riemannian manifold $\mathbb{M} =]0,1[$, equipped with the Riemannian metric defined in Eq. [iv.5], is a Riemannian homogeneous space.*

Proof. Let p, q be two points in $]0,1[$ with $p \neq q$. Consider the map $f_{p,q} :]0,1[\to]0,1[$

defined by:

$$\forall s \in \mathbb{M} =]0,1[,\ f_{p,q}(s) = \frac{1}{1+\exp\big(-\operatorname{logit}(q)+\operatorname{logit}(p)-\operatorname{logit}(s)\big)} = \Big(1+\frac{(1-p)q(1-s)}{p(1-q)s}\Big)^{-1}. \quad \text{[vi.66]}$$

For purposes of simplicity, we use the notation f instead of $f_{p,q}$. The map f is a diffeomorphism of $]0,1[$ onto itself. In order to prove that f is a local isometry of $\mathbb{M}$, we shall prove that (see Definition III.9):

$$\forall p_0 \in]0,1[,\ \forall (u,v) \in \mathrm{T}_{p_0}]0,1[\simeq \mathbb{R},\ \frac{uv\big(f'(p_0)\big)^2}{f(p_0)^2\big(1-f(p_0)\big)^2} = \frac{uv}{p_0^2(1-p_0)^2}. \quad \text{[vi.67]}$$

Writing $\forall s \in]0,1[,\ f(s) = 1/(1+g(s))$ with: $\forall s \in]0,1[,\ g(s) = \big((1-p)q(1-s)\big)/\big(p(1-q)s\big)$, Eq. [vi.67] is equivalent to proving that:

$$\forall p_0 \in]0,1[,\ \Big(\frac{g'(p_0)}{g(p_0)}\Big) = \frac{1}{p_0^2(1-p_0)^2} \quad \text{[vi.68]}$$

which follows easily from the definition of the function g. As a result, f is a local isometry. To complete the proof that f is, indeed, a global isometry of $]0,1[$, it remains to prove that:

$$\forall (p_0,q_0) \in]0,1[,\ d\big(f(p_0),f(q_0)\big) = d(p_0,q_0) \quad \text{[vi.69]}$$

where d denotes the Riemannian distance function, which is defined by:

$$\forall (p_0,q_0) \in]0,1[,\ d(p_0,q_0) = \Big|\ln\Big(\frac{q_0(1-p_0)}{p_0(1-q_0)}\Big)\Big|. \quad \text{[vi.70]}$$

Therefore, Eq. [vi.69] is equivalent to proving that:

$$\forall (p_0,q_0) \in]0,1[,\ \frac{g(p_0)}{g(q_0)} = \frac{q_0(1-p_0)}{p_0(1-q_0)} \quad \text{[vi.71]}$$

which follows directly from the definition of g. Finally, f is an isometry of $\mathbb{M}$. Since f is such that $f(p) = q$, the Riemannian manifold $]0,1[$ is a Riemannian homogeneous space. □

In [Lee, 2006], the author proves that the group of isometries of $\mathbb{S}^n$ is the orthogonal group $\mathcal{O}(n+1)$. Proposition 3.3 of this book proves that, given any pair of points $\mathbf{p}, \widetilde{\mathbf{p}}$ on $\mathbb{S}^n$ (with $\mathbf{p} \neq \widetilde{\mathbf{p}}$), there exist $\varphi \in \mathcal{O}(n+1)$ such that $\varphi(\mathbf{p}) = \widetilde{\mathbf{p}}$. In particular, this result proves that $\mathbb{S}^n$ is a Riemannian homogeneous space.

VI.3.6.1.3 Maximization step

In Section VI.3.3, we proposed parameters updates which do not take into account the Riemannian metric on $\mathbb{M}$. Even though, in practice, the parameter $\overline{\mathbf{p}_0}$ is always estimated on the Riemannian manifold $\mathbb{M}$, it is possible to derive an algorithm for the maximization step which ensures that the parameter $\overline{\mathbf{p}_0}$ always remain on the manifold $\mathbb{M}$. For all k, the function $\mathbf{Q}_k(\cdot)$ is defined on $\Theta = \mathbb{M} \times \mathbb{R}^N \times \mathbb{R}^{(N-1)N_s} \times]0, +\infty[^3$, *id est* the product of the manifold $\mathbb{M}$ and an open subset of the Euclidean space. Since $\mathbb{M}$ is a Riemannian manifold, the Riemannian metric on $\mathbb{M}$ changes the expression of the gradient of $\mathbf{Q}_k$. As mentioned in Section IV.3.3.1, the Riemannian metric $g^{\mathbb{M}}$ on $\mathbb{M}$ is of the form: $\forall \mathbf{p} \in \mathbb{M},\ \forall (\mathbf{u}, \mathbf{v}) \in \mathrm{T}_{\mathbf{p}}\mathbb{M},\ g^{\mathbb{M}}_{\mathbf{p}}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{G}(\mathbf{p})\mathbf{v}$. Therefore, it follows from the definitions in Section III.1.2.3 that the gradient at $\boldsymbol{\theta}$ of $\mathbf{Q}_k$ is defined by:

$$\operatorname{grad} \mathbf{Q}_k(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{G}(\overline{\mathbf{p}_0})^{-1} \dfrac{\partial \mathbf{Q}_k}{\partial \overline{\mathbf{p}_0}}(\boldsymbol{\theta}) \\ \dfrac{\partial \mathbf{Q}_k}{\partial \overline{\mathbf{v}_0}}(\boldsymbol{\theta}) \\ \vdots \\ \dfrac{\partial Q_k}{\partial \sigma}(\boldsymbol{\theta}) \end{bmatrix}. \tag{vi.72}$$

The gradient of the function can be used to maximize $\mathbf{Q}_k$ (or, equivalently, minimize -$\mathbf{Q}_k$) in a Steepest Descent (SD) algorithm. Indeed, gradient-based algorithm for the minimization of functionals defined on a Riemannian manifold were proposed in [Smith, 1994, Ring and Wirth, 2012].

In the case where $M =]0, 1[$ is equipped with the Riemannian metric given in Eq. [iv.5], the function G above is defined by: $\forall \overline{p_0} \in]0, 1[,\ G(\overline{p_0}) = \left(\overline{p_0}^2 (1 - \overline{p_0})^2\right)^{-1}$. Therefore, the coefficient $G(\overline{\mathbf{p}_0})^{-1}$ in Eq. [vi.72] ensures that the first coordinate of $\operatorname{grad} \mathbf{Q}_k$ vanishes as $\overline{\mathbf{p}_0}$ goes to the "ends" of of the manifold $\mathbb{M}$ and, in that case, the gradient descent moves less and less on the manifold, avoiding to converge to a point which is not on the manifold. The SD algorithm is described in 7. In this algorithm, the step size t_j is chosen such that the objective function $\mathbf{Q}_k$ decreases along the search direction given by the gradient.

Algorithm 7 Steepest descent on the Riemannian manifold $\mathbb{M}$ to maximize the function $\mathbf{Q}_k$.

Require: Initial guess $\boldsymbol{\theta}_0 \in \Theta$

Ensure: Set of parameters $\boldsymbol{\theta}^{(k+1)}$ solution of Eq. [vi.57]

1: Set $j = 0$, $\boldsymbol{\theta}_j = \boldsymbol{\theta}_0$ and $\overrightarrow{\mathbf{g}}_j = \operatorname{grad} \mathbf{Q}_k(\boldsymbol{\theta}_j)$
2: **repeat**
3: Compute the descent step size t_j using, for example, backtracking linesearch.
4: Set $\boldsymbol{\theta}_{j+1} = \operatorname{Exp}_{\boldsymbol{\theta}_j}(t_j \overrightarrow{\mathbf{g}}_j)$, $\overrightarrow{\mathbf{g}}_{j+1} = \operatorname{grad} \mathbf{Q}_k(\boldsymbol{\theta}_{j+1})$, $j = j + 1$
5: **until** convergence.
6: **Return**: $\boldsymbol{\theta}^{(k+1)} := \boldsymbol{\theta}_{j+1}$.

If we assume that the manifold $\mathbb{M}$ is no longer open in $\mathbb{R}^N$, the tangent space at a given point can no longer be identified with the ambient space. A way to maximize the function $\mathbf{Q}_k$ in the maximization step could be to consider that the function $\mathbf{Q}_k$ is defined on the product space $\mathrm{T}\mathbb{M} \times \mathbb{R}^{(N-1)N_s} \times]0, +\infty[$[3], where $\mathrm{T}\mathbb{M}$ denotes the tangent bundle of $\mathbb{M}$. Then, the function can be maximized using the steepest descent algorithm on a Riemannian manifold and taking into account the Riemannian manifold structure on the tangent bundle. As a matter of fact, $\mathrm{T}\mathbb{M}$ can be naturally equipped with a Riemannian metric called the *Sasaki metric* [Musso and Tricerri, 1988, Muralidharan and Fletcher, 2012]. Considering this Riemannian metric on the tangent bundle would ensure that the parameters $(\overline{\mathbf{p}_0}, \overline{\mathbf{v}_0})$ actually belong to the tangent bundle $\mathrm{T}\mathbb{M}$.

VI.4 Evaluation of the MCMC-SAEM

VI.4.1 Empirical validation on simulated data

VI.4.1.1 With the logistic curves propagation model

In order to validate the MCMC-SAEM, the logistic curves propagation model (Eq. [v.19]) is tested on a synthetic longitudinal dataset. Using the generative model (Eq. [v.19]), a longitudinal dataset with 248 individuals was generated, with an average of 7 time points per individual (min: 5, max: 9). For each individual, the observations were random perturbations of points in $]0, 1[$[4]. In addition to this, the number of independent components N_s was fixed to 2. The parameters used to generate this dataset are reported in Table 1. The MCMC-SAEM was run for a total of 1 020 000 iterations with a burn-in period of 600 000 iterations (which corresponds to 58% of the total number of iterations). The initial parameters (respectively estimated parameters) are reported in Table 2 (respectively Table 3).

The convergence of the parameters $\overline{\mathbf{p}_0}$, $\overline{t_0}$, $\overline{\mathbf{v}_0}$ is illustrated in Figure 12b. For a

$\overline{p_0}^*$	$\overline{t_0}^*$	$\overline{v_0}^*$	$\overline{\boldsymbol{\delta}}^*$	$\overline{\boldsymbol{\beta}}^*$	σ_ξ^*	σ_τ^*	σ^*
0.2	74	0.03	$\begin{bmatrix} 0 \\ -4 \\ -2 \\ -1 \end{bmatrix}$	$\begin{bmatrix} 0.27 \\ 0.176 \\ -0.02 \\ -0.12 \\ 0.17 \\ -0.16 \end{bmatrix}$	0.7	7	0.01

Table 1 – "True parameters": parameters used to generate the test dataset.

$\overline{p_0}$	$\overline{t_0}$	$\overline{v_0}$	$\overline{\boldsymbol{\delta}}$	$\overline{\boldsymbol{\beta}}$	σ_ξ	σ_τ	σ
0.3	65	0.01	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	0.5	5	1

Table 2 – Initial parameters used for the MCMC-SAEM.

matter of clarity, the other parameters are not displayed. Let $\boldsymbol{\theta}$ denote the parameters of the logistic curves propagation model. Let $\overline{p_0}^{(k)}$ (respectively $\overline{t_0}^{(k)}$, $\overline{v_0}^{(k)}$) denote the current estimate at the kth iteration of the MCMC-SAEM. The normalized errors plotted in Figure 12 (b) are defined as follows:

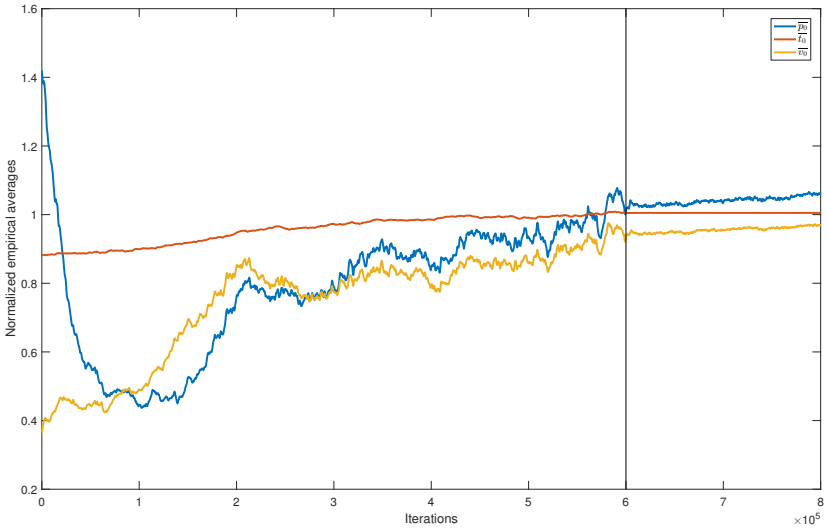
$$\frac{|\overline{p_0}^{(k)} - \overline{p_0}^*|}{\overline{p_0}^*} \left(\text{ respectively } \frac{|\overline{t_0}^{(k)} - \overline{t_0}^*|}{\overline{t_0}^*}, \frac{|\overline{v_0}^{(k)} - \overline{v_0}^*|}{\overline{v_0}^*} \right) \qquad \text{[vi.73]}$$

and the normalized empirical averages in Figure 12 (top) are defined as follows:

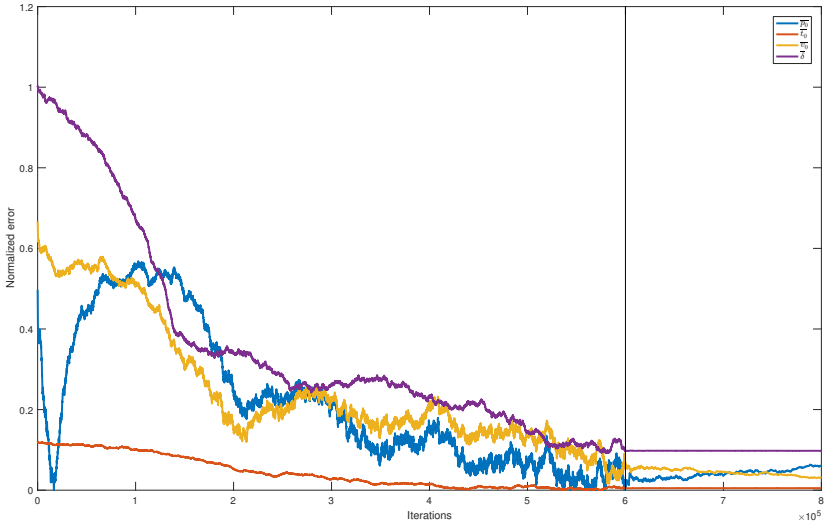
$$\frac{1}{k\overline{p_0}^*} \sum_{n=1}^{k} \overline{p_0}^{(k)} \left(\text{ respectively } \frac{1}{k\overline{t_0}^*} \sum_{n=1}^{k} \overline{t_0}^{(k)}, \frac{1}{k\overline{v_0}^*} \sum_{n=1}^{k} \overline{v_0}^{(k)} \right). \qquad \text{[vi.74]}$$

$\widehat{\overline{p_0}}$	$\widehat{\overline{t_0}}$	$\widehat{\overline{v_0}}$	$\widehat{\overline{\boldsymbol{\delta}}}$	$\widehat{\overline{\boldsymbol{\beta}}}$	$\widehat{\sigma_\xi}$	$\widehat{\sigma_\tau}$	$\widehat{\sigma}$
0.21	74.37	0.0291	$\begin{bmatrix} 0 \\ -4.29 \\ -1.96 \\ -0.8 \end{bmatrix}$	$\begin{bmatrix} 0.68 \\ 0.43 \\ -0.0545 \\ 0.09 \\ 0.0572 \\ -0.008 \end{bmatrix}$	0.65	6.69	0.01

Table 3 – Parameters estimated by the MCMC-SAEM.



(a)



(b)

Figure 12 – **Top**: Normalized errors for the parameters $\overline{p_0}$, $\overline{t_0}$ and $\overline{v_0}$. **Bottom**: Normalized empirical averages for the parameters $\overline{p_0}$, $\overline{t_0}$ and $\overline{v_0}$.

As discussed in Section VI.3.5.2, the normalized empirical averages are computed using a moving average of 1000 iterations.

VI.4.1.2 With the SPD matrices model

In this section, we provide results regarding the validation of the SPD matrices model on a synthetic dataset. Using the model described in Eq. [v.8], a longitudinal dataset of 3×3 symmetric positive definite matrices (covariance matrices) was generated with $p = 250$ individuals and an average of 6 time points per individual (a minimum of 5

and a maximum of 8). The parameters used to generate the test dataset are reported in the first column of Table 4. The MCMC-SAEM was run for a total of 2 900 000 iterations and with a burn-in period of 2 500 000 iterations, which corresponds to 83% of the total number of iterations. The initial parameters are reported in Table 5. Note that, in Table 6, the sign of $\widehat{\overline{\boldsymbol{\beta}}}$ is not correct. However, this is not surprising as, given the Independent Component Analysis (ICA) model on the space shifts $(\mathbf{w}_i)_{1\leq i\leq p}$, the vector $\overline{\boldsymbol{\beta}}$ can only be estimated up to a sign change. Similarly to the previous

$\overline{\mathbf{P}_0}^*$	$\overline{t_0}^*$	$\overline{\mathbf{V}_0}^*$	$\overline{\boldsymbol{\beta}}^*$	σ_ξ^*	σ_τ^*	σ^*
$\begin{bmatrix} 5.2 & 1.9 & 2.2 \\ 1.9 & 11.4 & 4.6 \\ 2.2 & 4.6 & 6.3 \end{bmatrix}$	60	$\begin{bmatrix} -0.13 & -0.17 & -0.16 \\ -0.17 & -0.66 & -0.37 \\ -0.16 & -0.37 & -0.25 \end{bmatrix}$	$\begin{bmatrix} 0.03 \\ 0.49 \\ 0.15 \\ 0.32 \\ -0.493 \end{bmatrix}$	0.4	3	0.20

Table 4 – "True parameters": parameters used to generate the test dataset.

$\overline{\mathbf{P}_0}$	$\overline{t_0}$	$\overline{\mathbf{V}_0}$	$\overline{\boldsymbol{\beta}}$	σ_ξ	σ_τ	σ
$\begin{bmatrix} 4 & 0 & 1 \\ 0 & 9 & 1 \\ 1 & 1 & 3 \end{bmatrix}$	70	$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	1	5	1

Table 5 – Initial parameters used for the MCMC-SAEM.

$\widehat{\overline{\mathbf{P}_0}}$	$\widehat{\overline{t_0}}$	$\widehat{\overline{\mathbf{V}_0}}$	$\widehat{\overline{\boldsymbol{\beta}}}$	$\widehat{\sigma_\xi}$	$\widehat{\sigma_\tau}$	$\widehat{\sigma}$
$\begin{bmatrix} 5.14 & 1.89 & 2.16 \\ 1.89 & 11.28 & 4.52 \\ 2.16 & 4.52 & 6.19 \end{bmatrix}$	60.74	$\begin{bmatrix} -0.13 & -0.16 & -0.15 \\ -0.16 & -0.63 & -0.39 \\ -0.15 & -0.39 & -0.24 \end{bmatrix}$	$\begin{bmatrix} -0.035 \\ -0.501 \\ -0.159 \\ -0.336 \\ 0.5105 \end{bmatrix}$	0.39	2.92	0.2011

Table 6 – Parameters estimated with the MCMC-SAEM and the SPD matrices model.

experiment on simulated data, the convergence of the MCMC-SAEM is illustrated with the normalized error of parameters $\overline{\mathbf{P}_0}$, $\overline{t_0}$ and $\overline{\mathbf{V}_0}$. As above, the normalized error for $\overline{\mathbf{P}_0}$ (respectively $\overline{t_0}$, $\overline{\mathbf{V}_0}$) are defined by:

$$\frac{\|\overline{\mathbf{P}_0}^{(k)} - \overline{\mathbf{P}_0}^*\|_F}{\|\overline{\mathbf{P}_0}^*\|_F} \left(\text{respectively } \frac{|\overline{t_0}^{(k)} - \overline{t_0}^*|}{\overline{t_0}^*}, \frac{\|\overline{\mathbf{V}_0}^{(k)} - \overline{\mathbf{V}_0}^*\|}{\overline{\mathbf{V}_0}^*} \right) \qquad \text{[vi.75]}$$

This experiment on a synthetic dataset shows that the MCMC-SAEM succeeds in

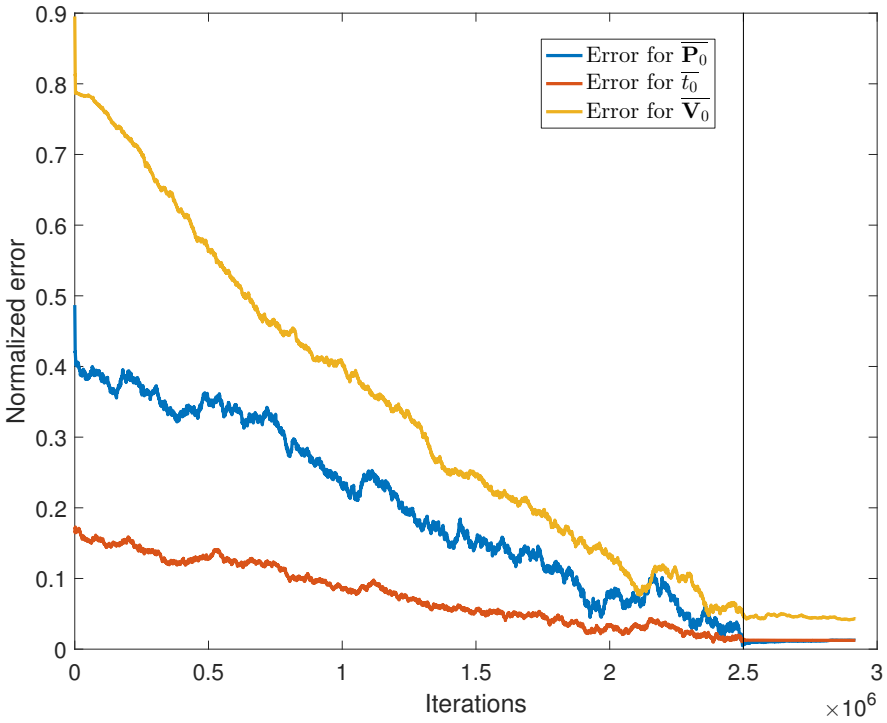


Figure 13 – Normalized errors for the parameters $\overline{\mathbf{P}_0}$, $\overline{t_0}$ and $\overline{\mathbf{V}_0}$.

providing estimates of the parameters of the model which are quite close to the ones used to generate the data, even though a large number of iterations was necessary to reach convergence.

VI.4.1.3 Runtime

In the previous sections, the MCMC-SAEM is used with the logistic curves propagation model and with the SPD matrices model. This section provides informations on the runtime of the MCMC-SAEM for these two examples. The two experiments on simulated data were run using MATLAB®, on a computer with 4 Intel®Xeon(R) CPU at 3.20GHz. The timings given below were measured using the `cputime` MATLAB function. The runtimes below are given for 1000 iterations.

VI.4.1.3.1 For the logistic curves propagation model

- Overall MCMC-SAEM algorithm: $333\,s$ (5.5 minutes ; $\sim 0.3\,s$ per iteration)

- Sampling step of the MCMC-SAEM: $300\,s$ (5 minutes)

- Computation of an orthonormal basis $(\mathcal{B}_k)_{1\leq k\leq (N-1)N_s}$: $3.05\,s$ for 7992 calls ($\sim 3.8\times 10^{-4}\,s$ per call).

VI.4.1.3.2 For the SPD matrices model

The SPD matrices model, without approximation of the parallel transport, requires to compute multiple exponentials, logarithms and square roots of matrices. Using MATLAB function to perform these operations leads to a costly algorithm. Indeed, for 1000 iterations, the MATLAB code (not optimized) of the overall MCMC-SAEM runs in $23\,714\,s$ (6.58 hours). By using MATLAB Executable Files (MEX files), the overall runtime was divided by 70.

- Overall MCMC-SAEM algorithm: $339\,s$ (5.6 minutes ; $\sim 0.3\,s$ per iteration)

- Sampling step of the MCMC-SAEM: $316\,s$ (~ 5 minutes)

- Computation of an orthonormal basis $(\mathcal{B}_k)_{1\leq k\leq (N-1)N_s}$: $9\,s$ for 8991 calls ($\sim 3.8\times 10^{-4}\,s$ per call).

VI.4.1.3.3 Discussion

These results show that the sampling step is the most expensive step of the MCMC-SAEM. The computations which are not included in the timing of the sampling step include the computation of the sufficient statistics, the update of the parameters. For both experiments, approximately 65% of the runtime of the sampling step is spent in the computation of the model likelihood $q(\mathbf{y}\,|\,\mathbf{z},\boldsymbol{\theta})$. The computation of an orthonormal basis of $\mathrm{Span}\big(\dot{\boldsymbol{\gamma}}_0(t_0)\big)$ is definitely not costly for the low dimensional examples considered here.

VI.4.2 Comparison with standard methods and algorithms

VI.4.2.1 Comparison between the "logistic curves model" and a LME model

In Section V.1.3.1, it is mentioned that the logistic curves model (abridged LC model ; see Eq. [v.7]) is not equivalent to a LME model on observations transformed with the logit function. In this section, we provide a numerical comparison of both methods, which completes the discussion of Section V.1.3.1. To this end, we consider a longitudinal dataset which is analyzed later in Section VII.3.1. This longitudinal dataset consists in normalized scalar observations for $p = 1393$ individuals, with a minimum

(respectively maximum, average) of 3 (respectively 11, 5) time points per individual. Let $(y_{i,j}, t_{i,j})$ denote this longitudinal dataset.

Recall that the logistic curves model writes:

$$y_{i,j} = \left(1 + \left(\frac{1}{p_0} - 1\right) \exp\left(-\frac{v_0 \alpha_i (t_{i,j} - t_0 - \tau_i)}{p_0(1-p_0)}\right)\right)^{-1} + \varepsilon_{i,j}. \qquad \text{[vi.76]}$$

In this section, we assume that $\varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\text{LC}})$. The other random effects are of the model are distributed as in Section IV.3.4. We propose to compare this nonlinear mixed-effects model to a "random slope and intercept" model on transformed observations. Indeed, it is quite common, with observations in $]0,1[$, to map these observations to the real line using the logit transform (defined in Eq. [iv.7]) and then perform a linear analysis. As a consequence, the logistic curves model is to be compared to the following LME model:

$$\text{logit}(y_{i,j}) = (\overline{A} + A_i) t_{i,j} + \overline{B} + B_i + \widetilde{\varepsilon}_{i,j}. \qquad \text{[vi.77]}$$

with:

$$\begin{bmatrix} A_i \\ B_i \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_A & 0 \\ 0 & \sigma^2_B \end{bmatrix}\right) \text{ and } \widetilde{\varepsilon}_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_{\text{LME}}). \qquad \text{[vi.78]}$$

Since the logistic curves model is used with observations in $]0,1[$ and the LME model is used with observations in $\mathbb{R}$, the variance of the noise terms $(\varepsilon_{i,j})$ and $(\widetilde{\varepsilon}_{i,j})$ cannot be directly compared. To address this problem, we propose to compare the percentage of variance explained by both models.

For the logistic curves (LC) model, let $\widehat{\boldsymbol{\theta}}_{\text{LC}} = \left(\widehat{p_0}, \widehat{t_0}, \widehat{v_0}, \widehat{\sigma^2_\xi}, \widehat{\sigma^2_\tau}, \widehat{\sigma^2_{\text{LC}}}\right)$ be the parameters estimated with the MCMC-SAEM. For each individual, the MAP (Maximum A Posteriori) estimates $(\widehat{\alpha}_i, \widehat{\tau}_i)$ of the individual random effects are obtained by maximizing the joint conditional distribution $q(\alpha_i, \tau_i \mid \mathbf{y}_i, \widehat{\boldsymbol{\theta}}_{\text{LC}})$. The residuals $(\widehat{r}_{i,j})_{i,j}$ of the model are therefore computed as follows:

$$\forall i, j,\ \widehat{r}_{i,j} = y_{i,j} - \left(1 + \left(\frac{1}{\widehat{p_0}} - 1\right) \exp\left(-\frac{\widehat{v_0}\widehat{\alpha}_i (t_{i,j} - \widehat{t_0} - \widehat{\tau}_i)}{\widehat{p_0}(1-\widehat{p_0})}\right)\right)^{-1}. \qquad \text{[vi.79]}$$

and the percentage of total variance explained by the LC model is defined by:

$$R^2_{\text{LC}} = 1 - \frac{\text{var}(\widehat{r}_{i,j})}{\text{var}(y_{i,j})}. \qquad \text{[vi.80]}$$

For the LME model (see Eq. [vi.77]), let $\widehat{\boldsymbol{\theta}}_{\text{LME}} = \left(\widehat{\overline{A}}, \widehat{\overline{B}}, \widehat{\sigma^2_A}, \widehat{\sigma^2_B}, \widehat{\sigma^2_{\text{LME}}}\right)$ denote the parameters estimated using the `fitlmematrix` function of the MATLAB software.

The MAP estimates $(\widehat{A}_i, \widehat{B}_i)_{1\leq i\leq p}$ of the individual random effects are obtained using the BLUP, given in Eq. [viii.8]. The residuals $\left(\widehat{\widetilde{r}}_{i,j}\right)$ of this model are defined by:

$$\forall i,j,\ \widehat{\widetilde{r}}_{i,j} = y_{i,j} - \frac{1}{1+\exp\left(-(\widehat{\widetilde{A}}+\widehat{A}_i)t_{i,j} - (\widehat{\widetilde{B}}+\widehat{B}_i)\right)}. \qquad \text{[vi.81]}$$

The percentage of total variance explained by the LME model is defined by:

$$R^2_{\text{LME}} = 1 - \frac{\text{var}\left(\widehat{\widetilde{r}}_{i,j}\right)}{\text{var}(y_{i,j})}. \qquad \text{[vi.82]}$$

Finally, we obtained: $R^2_{\text{LME}} = 82.2\%$ and $R^2_{\text{LC}} = 93.3\%$. Therefore, the logistic curves model fits the observations better than the LME model. These results illustrates the idea that the logistic curves model is not equivalent to a LME model. In the LME model (see Eq. [vi.77]), if the term $t_{i,j}$ is replaced by $(t_{i,j} - \widehat{t_0})$, where $\widehat{t_0} = 70.3$ years is the parameter estimated with the MCMC-SAEM, the percentage of total variance explained by the LME model raises to $90,7\%$. Even though the result improved, it remains lower than with the one obtained with the logistic curves model. In addition to this, if the logistic curves model had not been used before, it would have been impossible to estimate $\widehat{t_0}$ with the other parameters of the LME model.

VI.4.2.2 Comparison between the MCMC-SAEM and the Laplacian Approximation

In this section, two algorithms are considered for the estimation of the parameters of the univariate logistic curves model (seeEq. [v.7]). The longitudinal dataset considered here is the same as the one introduced in the previous section.

The first algorithm considered is the Laplacian Approximation. The `nlmixed` procedure of the SAS Software implements, among other likelihood approximation methods, the Adaptive Gaussian quadrature method presented in [Pinheiro and Bates, 1996]. With a single quadrature point, this method is equivalent to the Laplacian Approximation presented in Eq. [vi.28]. Note that no priors on the parameters were used in this experiment. At the beginning of the algorithm, the parameters were initialized to the value of the hyperparameters proposed in Section VI.3.4. The estimates obtained with the SAS Software are reported in Table 7. The MCMC-SAEM was also

$\overline{p_0}$	$\overline{t_0}$	$\overline{v_0}$	σ_ξ	σ_τ	σ
0.22	77.17	0.0139	0.64	7.42	**0.13**

Table 7 – Parameters estimated with the Laplacian Approximation (`nlmixed` procedure) of the SAS Software.

used to estimate the parameters of the univariate logistic curves model on this dataset.

The hyperparameters were chosen as discussed in Section VI.3.4. As for the Laplacian Approximation, the parameters were initialized to the same value than the hyperparameters. The estimates obtained with the MCMC-SAEM are reported in Table 8. The parameters estimated by both methods are not really similar and this is coherent

$\overline{p_0}$	$\overline{t_0}$	$\overline{v_0}$	σ_ξ	σ_τ	σ
0.1148	69.68	0.0076	0.94	10.42	**0.0384**

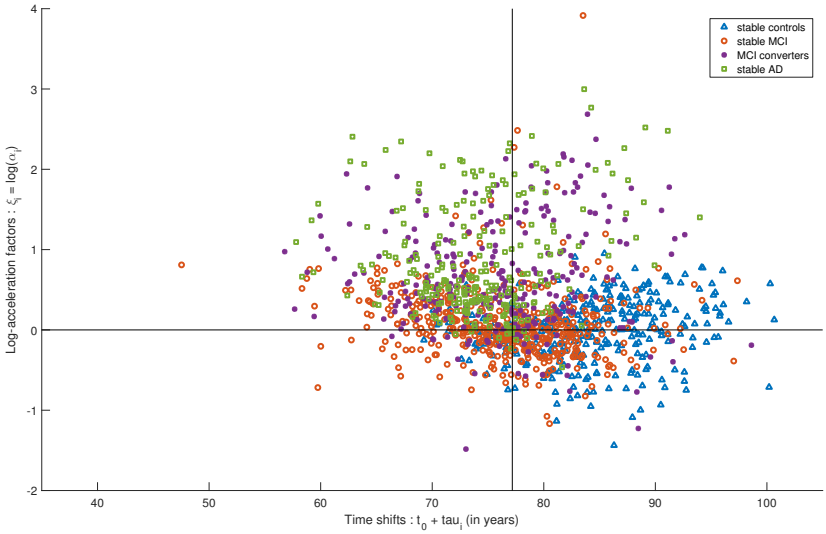
Table 8 – Parameters estimated with the MCMC-SAEM.

with the fact that the standard deviation of the noise is much smaller with the MCMC-SAEM than with the Laplacian Approximation. Hence, the MCMC-SAEM explained more variance than the Laplacian approximation. Still, the MCMC-SAEM converges more slowly due to the high number of iterations required to observe the convergence of the empirical averages of the parameters $\overline{p_0}$, $\overline{t_0}$ and $\overline{v_0}$. In fact, the MCMC-SAEM was run for a total of 700 000 iterations with a burn-in period of 500 000 iterations. The difference we observe between the two methods may be explained by the use of priors with the MCMC-SAEM. Indeed, the priors act a regularization terms on the likelihood.

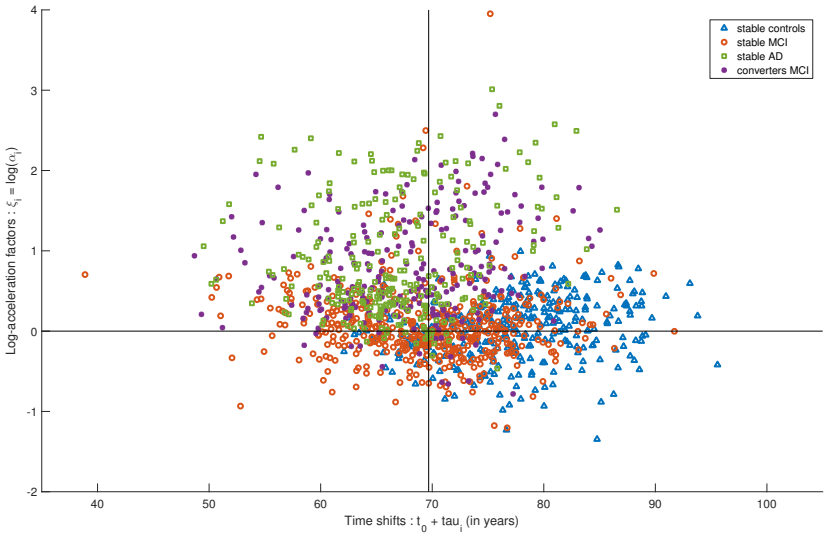
We propose to show that, even though the parameters estimated with both methods are quite different, the information provided by the MAP (maximum a posteriori) estimates of the individual random effects are quite similar. For each method, let $\boldsymbol{\theta}^*$ denote the estimated parameters. MAP estimates of the individual random effects $\mathbf{z}_i = (\xi_i, \tau_i)_{1 \leq i \leq p}$ were obtained by maximizing the conditional distribution $q(\xi_i, \tau_i \mid \mathbf{y}_i, \boldsymbol{\theta}^*)$ of (ξ_i, τ_i) given the observations $\mathbf{y}_i$ of the ith individual and the estimated parameters $\boldsymbol{\theta}^*$. The MAP estimates obtained from both methods are plotted in Figure 14. The 1393 individuals are grouped into 4 groups: "stable controls", "stable Mild Cognitive Impairment (MCI)", "stable Alzheimer's Disease (AD)" and "converters MCI". Even though the MCMC-SAEM allowed to obtain a better residual noise for the same longitudinal dataset, we can observe that the plots of the individual random effects are not dramatically different. In particular, they both provide similar informations in terms of disease progression among the studied population. Figure 14 (b) is interpreted and discussed in Section VII.3.

VI.4.2.3 Comparison of our MCMC-SAEM algorithm with STAN and MONOLIX

This section aims at comparing our implementation of the MCMC-SAEM algorithm with other state-of-the-art algorithms for the inference in nonlinear mixed-effects models and implementations: STAN and MONOLIX. Recall that STAN is a library, available in R or C++, which implements an adaptive Hamiltonian Monte Carlo sampler called the *No U-Turn Sampler* (NUTS, [Hoffman and Gelman, 2014]). MONOLIX is



(a)



(b)

Figure 14 – Plot of (the MAP estimates of) the individual log-acceleration factor ξ_i against the time shifts $t_0 + \tau_i$. Figure (a) is the result of the Laplacian approximation and Figure (b) is the result of the MCMC-SAEM. Each point corresponds to an individual. On both plots, a vertical line was drawn at the estimated value of t_0. Figure (a) is taken from [Schiratti et al., 2015d].

a software, developed by Marc Lavielle [Lavielle and Mentré, 2007] and promoted by the Lixoft company, which implements the MCMC-SAEM algorithm.

In order to compare these algorithms, we considered a synthetic longitudinal dataset of observations in $]0, 1[$. This dataset was generated for $p = 250$ individuals, with an average of 5 time points per individual, using the logistic curves model (see Eq. [v.7])

with parameters reported in Table 9. The three algorithms (our MCMC-SAEM described in this dissertation, STAN and MONOLIX) were used with the logistic curves model (Eq. [v.7]) and with initial parameters reported in Table 10. Our MCMC-SAEM

p_0	t_0	v_0	σ_ξ	σ_τ	σ
0.24	70	0.034	0.5	7	0.01

Table 9 – Parameters used to generate the longitudinal dataset.

p_0	t_0	v_0	σ_ξ	σ_τ	σ
0.6	60	0.05	1	1	1

Table 10 – Initial parameters: the same initial parameters were used with each algorithm.

algorithm and STAN were run on a PC with 4 Intel Xeon(R) CPU at 3.20 GHz, even though the computations were not done in parallel. The MONOLIX algorithm was run on a MacBook Pro, with Intel Core $i7$ CPU at 2.5 GHz. Since the algorithms were not run on the same platform, with the same resources, the runtimes presented below should be read with caution. The results obtained with each algorithm are given in the following sections.

VI.4.2.3.1 Results obtained with our MCMC-SAEM algorithm

Our MCMC-SAEM was run for 230 000 iterations, with a burn-in period of 175 000 iterations. The parameters estimated with our MCMC-SAEM are reported in Table 11. The number of iterations and length of the burn-in period were determined by running several times the experiment since the algorithm does not have a generic stopping rule. Figure 15 represents the evolution of the normalized empirical averages for the parameters $\overline{p_0}$, $\overline{t_0}$ and $\overline{v_0}$, computed on a moving window of length 1000. This figure shows that, after 230 000 iterations, the MCMC-SAEM had converged to values close to the ones used to generate the data. Regarding the runtime of our MCMC-SAEM,

p_0	t_0	v_0	σ_ξ	σ_τ	σ
0.23	69.93	0.0317	0.52	6.75	0.01

Table 11 – Parameters estimated with our MCMC-SAEM

1 000 iterations of the algorithm (run in MATLAB) took, on average, 6.05 seconds. The runtime is actually quite fast because the MCMC-SAEM uses, as discussed in the previous sections, a block MHwG sampler, which allows to reduce the number

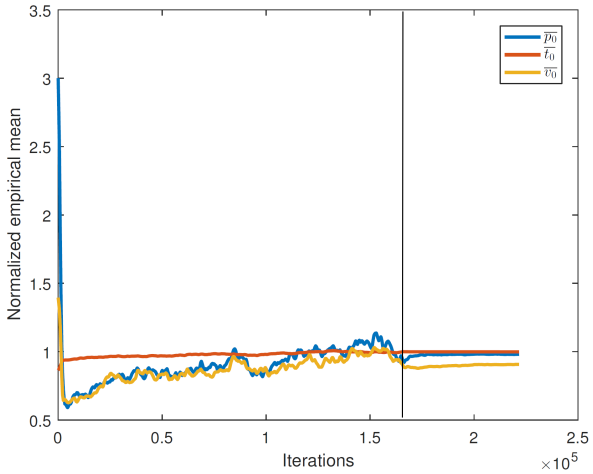


Figure 15 – Convergence of the MCMC-SAEM: plots of normalized empirical averages, computed on a moving window of length 1000 iterations.

of computations of the model likelihood. Moreover, no parallel transport had to be computed.

VI.4.2.3.2 Results obtained with STAN

As discussed above, the STAN software does not implement a variant of the MCMC-SAEM. It implements the "NO U-Turns Sampler" (NUTS, [Hoffman and Gelman, 2014]), which is an adaptive Hamiltonian Monte Carlo sampler. As for our MCMC-SAEM, no automatic stopping rule is implemented in STAN. In order to assess the convergence of the sampler, STAN computes a coefficient, for each Markov chain, called `Rhat` (potential scale reduction factor). This coefficient is also known as the "Gelman and Rubin's convergence diagnostic" [Gelman and Rubin, 1992]. This convergence diagnostic is quite popular to determine whether a Markov chain may have converged. When the value of a `Rhat` coefficient is below 1.01, as well as its corresponding upper confidence limit, we can assume that the chain has converged. A review of other convergence diagnostics for MCMC methods can be found in [Cowles and Carlin, 1996].

In a first experiment, the STAN software was run for 2 000 iterations. The parameters estimates obtained after 2 000 iterations are reported in Table 12. Only the variance parameters were correctly estimated. For the other parameters, the associated Markov chains had not converged yet since their `Rhat` coefficient was above 1.3. A total of 15 000 iterations were necessary, in a second experiment, to reach convergence

and obtain a `Rhat` coefficient below 1.01 for each Markov chain. The results obtained after 15 000 iterations are presented in Table 13. Regarding the runtime of the STAN

p_0	t_0	v_0	σ_ξ	σ_τ	σ
0.20	68.35	0.029	0.52	6.93	0.0998

Table 12 – Parameters estimated with STAN after 2 000 iterations.

p_0	t_0	v_0	σ_ξ	σ_τ	σ
0.218	68.66	0.0305	0.53	6.73	0.098

Table 13 – Parameters estimated with STAN after 15 000 iterations.

software, 1 000 iterations took, on average, 1 500 seconds (25 minutes). The "warm-up" phase of the algorithm is the most costly part of the algorithm since it takes up to 75% of the overall runtime. The "sampling" part of the algorithm is usually quite fast. This large runtime may be explained by the fact that, in opposition to the block MHwG sampler, the NUTS sampler require numerous computations of the gradient of the model likelihood.

VI.4.2.3.3 Results obtained with MONOLIX

The MONOLIX software was run on the simulated dataset and stopped automatically after 500 iterations. Indeed, the MCMC-SAEM implemented in MONOLIX uses, by default, an automatic rule which sets the burn-in period and the total number of iterations. The software was run several times and, each time, after 500 iterations (burn-in period stopped at 300 iterations), the algorithm had not converged to the parameters used to generate the data. As a matter of fact, the results obtained with MONOLIX after 500 iterations are reported in Table 14. Only the standard deviation σ of the noise and the standard deviation of the log-acceleration factors $(\xi_i)_{1\leq i\leq p}$ are correctly estimated. In another experiment, the MONOLIX software was forced to run for a large number of iterations (with a long burn-in period). After approximately 20 000 iterations, the parameters estimates almost did not change anymore. The values estimated after 20 000 iterations are reported in Table 15. In addition to this, the evolution of the parameters is represented in Figure 16.

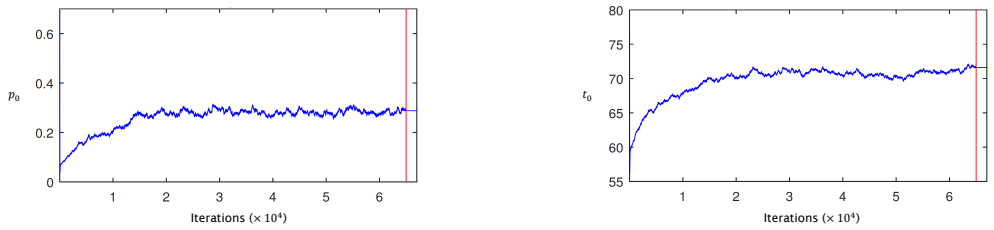
After 20 000 iterations, all the variance parameters were correctly estimated. Still, the estimates of the fixed effects remain quite different from the ones used to generate the data. Regarding the runtime of the MONOLIX software, 1 000 iterations took, on average, 3.5 minutes. In contrast to the STAN software, the MCMC-SAEM implemented in the MONOLIX software does not require to compute derivatives of the model likelihood. Moreover, its C++ implementation makes it very efficient. Moreover,

p_0	t_0	v_0	σ_ξ	σ_τ	σ
0.078	60.7	0.010	0.55	8.11	0.099

Table 14 – Parameters estimated with MONOLIX after 500 iterations.

p_0	t_0	v_0	σ_ξ	σ_τ	σ
0.37	71.6	0.0406	0.52	6.8	0.01

Table 15 – Parameters estimated with MONOLIX after 20 000 iterations.



(a) (b)

Figure 16 – **Figure (a)** (respectively **(b)**): evolution of the parameter p_0 (respectively t_0).

the MCMC-SAEM implemented in MONOLIX uses a Simulated Annealing algorithm to avoid getting stuck in local maximums. This allows for a better and faster exploration of the parameters space. However, the MONOLIX software currently does not implements vectors or matrices, which prevented its use with more complex models.

VI.4.2.3.4 Comments

The results presented above show that, even with optimized algorithms (like STAN or MONOLIX) large number of iterations are required to reach convergence. A possible explanation would be that the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$, which is maximized by these algorithms, is probably quite flat around its global maximum. Therefore, exploring the parameters space to find this maximum is a difficult optimization problem. In higher dimension, adding the notion of parallel variation into the model, as well as many latent variables, increases the difficulty of this optimization problem. This could explain why the experiments presented in Section VI.4.1 required many iterations to converge.

VI.4.3 Detecting errors in the sampling step

This section aims at proposing a method to help detect coding errors in MCMC samplers. The method we propose is adapted from the work of [Geweke, 2004]. Originally introduced in a Bayesian framework, the generic method is described in Section VI.4.3.1. Then, this method is derived for our generic spatiotemporal model in Section VI.4.3.2.

VI.4.3.1 The generic method

We consider a statistical model which specifies a probability distribution for some observations $\mathbf{y} \in \mathbb{R}^T$, conditionally on unobserved parameters $\boldsymbol{\theta} \in \widetilde{\Theta} \subset \mathbb{R}^L$. The model is characterized by the density function $q(\mathbf{y} \mid \boldsymbol{\theta})$. We also assume that the model specifies a proper prior distribution for the parameters $\boldsymbol{\theta}$, characterized by the density function $q(\boldsymbol{\theta})$. As discussed in Chapter III, "Fully Bayesian" methods aim at learning the posterior distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$ of the unobserved parameters, given some observations $\mathbf{y}$. When sampling from the posterior is not directly possible (because it is known up to an intractable normalizing constant, for example), the sampling can be done by using MCMC samplers. These samplers construct an ergodic Markov chain with stationary distribution $q(\boldsymbol{\theta} \mid \mathbf{y})$. For complex models, writing a MCMC sampler to sample from $q(\boldsymbol{\theta} \mid \mathbf{y})$ may be complicated and could lead to coding error. We propose an iterative method which could help determine whether a MCMC sampler is "error-free". To do this, we use the MCMC sampler to construct an ergodic Markov chain $\left(\mathbf{y}^{(k)}, \boldsymbol{\theta}^{(k)}\right)_{k \geq 0}$ whose stationary distribution is the joint distribution $q(\mathbf{y}, \boldsymbol{\theta})$. At each iteration, the sample $\boldsymbol{\theta}^{(k)}$ is saved. After a sufficiently large number of iterations (necessary for the chain to converge to its stationary distribution), if the sampler is "error-free", the samples $\left(\mathbf{y}^{(k)}, \boldsymbol{\theta}^{(k)}\right)$ should be (approximately) distributed as $q(\mathbf{y}, \boldsymbol{\theta})$. Therefore, comparing an histogram of the samples $\left(\boldsymbol{\theta}^{(k)}\right)_{k \geq 0}$ to the density function of the prior distribution allows to determine if there could be a problem with the sampler. This method is implemented in Algorithm 8.

Algorithm 8 Posterior sampler test

Require: $N_{\max}$: number of iterations.

Ensure: Samples $\left(\boldsymbol{\theta}^{(k)}\right)_{k\geq 0}$

1: **Initialization**: $\boldsymbol{\theta}^{(0)} \sim q(\boldsymbol{\theta})$ and $\mathbf{y}^{(0)} \sim q(\mathbf{y} \mid \boldsymbol{\theta}^{(0)})$.
2: **for** $k = 1$ to $N_{\max}$ **do**
3: $\quad \boldsymbol{\theta}^{(k)} \sim \boldsymbol{\pi}_{\mathbf{y}^{(k-1)}}(\boldsymbol{\theta}^{(k-1)}, \cdot)$
4: $\quad$ Save $\boldsymbol{\theta}^{(k)}$
5: $\quad \mathbf{y}^{(k)} \sim q(\mathbf{y} \mid \boldsymbol{\theta}^{(k)})$
6: **end for**
7: **Return**: samples $\left(\boldsymbol{\theta}^{(k)}\right)_{k\geq 0}$

In Algorithm 8, $\boldsymbol{\pi}_{\mathbf{y}^{(k-1)}}(\boldsymbol{\theta}^{(k-1)}, \cdot)$ denotes the transition kernel of an ergodic Markov chain whose stationary distribution is $q(\boldsymbol{\theta} \mid \mathbf{y}^{(k-1)})$.

Proposition VI.2. *Algorithm 8 generates an ergodic Markov chain* $(\mathbf{y}^{(k)}, \boldsymbol{\theta}^{(k)})_{k\in\mathbb{N}}$ *whose stationary distribution is the joint distribution* $q(\mathbf{y}, \boldsymbol{\theta})$.

An outline of the proof of Proposition VI.2 is given in [Geweke, 2004].

VI.4.3.2 Application to the generic spatiotemporal model

In order to apply this method to the generic spatiotemporal model, the method proposed in the previous section should be adapted to deal with latent variables. To this end, we propose to consider the parameters $\boldsymbol{\theta}$ of the generic spatiotemporal model *fixed* and use Algorithm 8 to construct an ergodic Markov chain whose stationary distribution is the joint distribution $q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$, where $\mathbf{y}$ denote some observations and $\mathbf{z}$ the latent variables of the model. The proposed method for the generic spatiotemporal model is given in Algorithm 9. Recall form Section IV.3.4 that $\mathbf{z} = \left(\mathbf{z}_{\text{pop}}, (\mathbf{z}_i)_{1\leq i\leq p}\right)$ with $\mathbf{z}_{\text{pop}} = \left(\mathbf{p}_0, t_0, \mathbf{v}_0, (\beta_{l,k})_{l,k}\right)$ and, for all $i \in \{1, \ldots, p\}$, $\mathbf{z}_i = \left(\xi_i, \tau_i, (s_{l,i})_l\right)$. If, at each iteration of Algorithm 9, the latent variable $t_0^{(k)}$ is saved into the vector $\mathbf{R}$, after a large number of iterations $N_{\max}$, the samples $(t_0^{(k)})_k$ should be approximately distributed as its prior distribution, if step `3:` of Algorithm 8 is "error-free". Indeed, the marginal distribution of t_0 in $q(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta})$ is the prior t_0, *id est* $t_0 \sim \mathcal{N}(\overline{t_0}, \sigma_{t_0}^2)$.

Algorithm 9 Posterior sampler test derived for the generic spatiotemporal model

Require: Observations $\mathbf{y}$ generated using Eq. [iv.66], a set of parameters $\boldsymbol{\theta}$, $N_{\max}$: number of iterations.

Ensure: Samples $\left(\mathbf{y}^{(k)}, \mathbf{z}^{(k)}\right)_{k\geq 0}$

1: **Initialization**: $\mathbf{z}^{(0)} \sim q(\mathbf{z} \mid \boldsymbol{\theta})$ and $\mathbf{y}^{(0)} \sim q(\mathbf{y} \mid \mathbf{z}^{(0)}, \boldsymbol{\theta})$.
2: **for** $k = 1$ to $N_{\max}$ **do**
3: $\quad \mathbf{z}^{(k)} \sim \boldsymbol{\pi}_{\mathbf{y}^{(k-1)},\boldsymbol{\theta}}(\mathbf{z}^{(k-1)}, \cdot)$
4: $\quad \mathbf{R}[k] \leftarrow f\left(\mathbf{z}^{(k)}\right)$
5: $\quad \mathbf{y}^{(k)} \sim q(\mathbf{y} \mid \mathbf{z}^{(k)}, \boldsymbol{\theta})$
6: **end for**
7: **Return**: the vector $\mathbf{R}$.

VI.4.3.2.1 Numerical examples and discussion

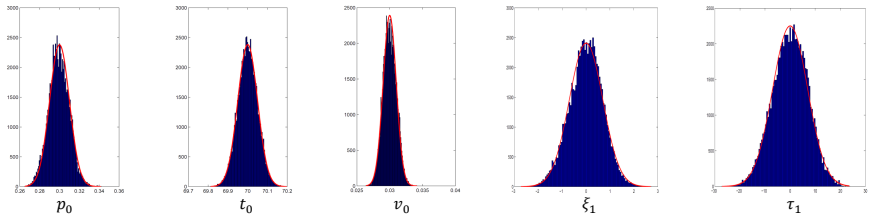
The posterior sampler test described in Algorithm 8 was tested in two different situations:

(i) In the first experiment, the algorithm was tested with the univariate logistic curves model (see Eq. [v.7]). Therefore, the observations $\mathbf{y}^{(k)}$ generated at each step of the algorithm consist, for each individual, in perturbations of points in the one-dimensional Riemannian manifold $\mathbb{M} =]0, 1[$. The parameters used for this experiment are given by: $\overline{p_0} = 0.3$, $\overline{t_0} = 70$ years, $\overline{v_0} = 0.03$, $\sigma_\xi = 0.7$, $\sigma_\tau = 7$ and $\sigma = 0.1$. The number of individuals was chosen to be $p = 250$.

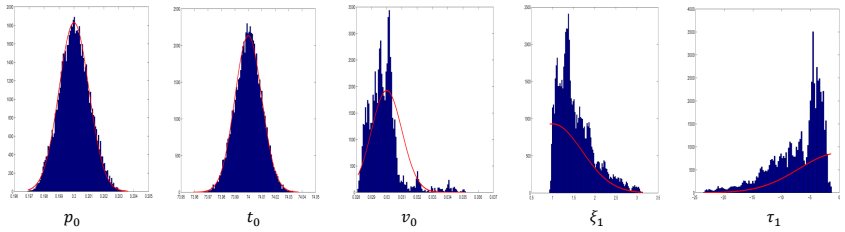
(ii) For the second experiment, we used the logistic curves propagation model (see Eq. [v.19]). This model was used with $N = 4$. Therefore, at each step of Algorithm 9, the observations $\mathbf{y}^{(k)}$ were, for each individual, a random perturbation of a point in $\mathbb{M} =]0, 1[^4$. The parameters used for this experiment are: $\overline{p_0} = 0.2$, $\overline{t_0} = 74$ years, $\overline{v_0} = 0.03$, $\overline{\delta} = \begin{bmatrix} 0 & -4 & -2 & -1 \end{bmatrix}^\top$ (in years), $\sigma_\xi = 0.7$, $\sigma_\tau = 7$ years, $\sigma = 0.01$ and $\overline{\boldsymbol{\beta}} = \begin{bmatrix} 0.27 & 0.1768 & -0.02 & -0.12 & 0.1718 & -0.16 \end{bmatrix}^\top$. The number of individuals was chosen to be $p = 250$ and the number of independent components N_s was chosen equal to $N_s = 2$.

For the first experiment, the algorithm was run for 50 000 iterations. At each iteration, the latent variables $(p_0, t_0, v_0, \xi_1, \tau_1)$ were saved. At the end of the algorithm, for each latent variable, a normalized histogram of the saved samples was plotted and the density of the probability distribution assumed for this variable was superimposed to the histogram. The results are given in Figure 17 (a). These results show that, for each latent variable, the density function is well superimposed to the histogram. This means that the validation of the sampler was successful.

Similarly, for the second experiment, the algorithm was run for 150 000 iterations and the latent variables $(p_0, t_0, v_0, \xi_1, \tau_1)$ were saved at each iteration. The results obtained in the second experiment are given in Figure 17 (b). The results obtained with this experiment are less satisfying. Indeed, the density function is well superimposed to the histogram for p_0 and t_0. However, it is not the case for the other latent variables. The reason why the results are different in dimension $N = 4$ remains an open methodological question. A possible answer would be as follows. In the first experiment, the dimension of the space of the latent variables equals $2p + 3 = 503$, whereas in the second experiment, it equals: $4p + 12 = 1012$. The block MHwG sampler used in the MCMC-SAEM may have difficulties sampling efficiently in this very high-dimensional setting. Even though the results are less satisfying in dimension $N = 4$ than in di-



(a)



(b)

Figure 17 – **Figure (a)** (respectively **Figure (b)**): Normalized histograms of the samples of the latent variables $(p_0, t_0, v_0, \xi_1, \tau_1)$ for the first experiment (respectively second experiment).

mension $N = 1$, the experimental results presented in Chapter VII show that the MCMC-SAEM, used with the block MHwG sampler, allows to obtain meaningful parameters estimates. Moreover, the maximum a posteriori estimates of the individual random effects provide informations which are consistent with the knowledge on the progression of neurodegenerative diseases (see Section VII.3).

VI.4.4 Numerical schemes for parallel transport and construction of an orthonormal basis

VI.4.4.1 The Schild's Ladder algorithm

As mentioned in Section IV.2.3, the parallel transport is not always available in closed-form. In such a situation, numerical schemes can be used to approximate it. The first approach to address this problem consists in solving the set of differential equation which define the parallel transport. If the Riemannian exponential and Riemannian logarithm are available, another possible solution is a numerical scheme called *Schild's ladder*. The Schild's ladder was introduced in the 1970's by Alfred Schild, in the context of general relativity. Since then, it has been used in various fields, such as medical imaging [Lorenzi et al., 2011, Ng et al., 2014], general relativity or computer vision [Rumpf and Wirth, 2012]. In [Kheyfets et al., 2000], the authors describe the Schild's ladder for an arbitrary affine connection and prove its convergence. However, they do not give the order of convergence of this numerical scheme.

The Schild's Ladder is an iterative algorithm which approximates the parallel transport of a tangent vector along a curve or a geodesic by repeating the procedure described below. Consider a curve c drawn on a smooth manifold $\mathbb{M}$. Let $\mathbf{P}_0$, $\mathbf{P}_1$ be points on c. Let $\mathbf{v}$ be a tangent vector to $\mathbb{M}$ at $\mathbf{P}_0$. One step of the Schild's Ladder transports the vector $\mathbf{v}$ from $\mathbf{P}_0$ to $\mathbf{P}_1$ as follows:

(i) Let $\mathbf{P}'_0 = \mathrm{Exp}_{\mathbf{P}_0}(\mathbf{v})$.

(ii) Let $\mathbf{P}_2 = \mathrm{Exp}_{\mathbf{P}'_0}\left(\frac{1}{2}\mathrm{Log}_{\mathbf{P}'_0}(\mathbf{P}_1)\right)$. The point $\mathbf{P}_2$ correspond to the midpoint of the geodesic from $\mathbf{P}'_0$ to $\mathbf{P}_1$.

(iii) Let $\mathbf{P}'_1 = \mathrm{Exp}_{\mathbf{P}_0}\big(2\mathrm{Log}_{\mathbf{P}_0}(\mathbf{P}_2)\big)$. The point $\mathbf{P}'_1$ corresponds to the endpoint of the geodesic starting from $\mathbf{P}_0$, whose midpoint is $\mathbf{P}_2$.

(iv) The parallel transport along c of the vector $\mathbf{v}$, from $\mathbf{P}_0$ to $\mathbf{P}_1$, is $\mathbf{v}' = \mathrm{Log}_{\mathbf{P}_1}(\mathbf{P}'_1)$.

This procedure is illustrated in Figure 18. For an arbitrary smooth manifold $\mathbb{M}$, the points $\mathbf{P}_0$ and $\mathbf{P}_1$ are chosen close enough that this numerical scheme takes place within a single coordinate chart.

Algorithm 10 Approximation of the parallel transport based on the Schild's Ladder procedure

Require: Step size $\varepsilon > 0$, tangent vector $\mathbf{w} \in \mathrm{T}_{\boldsymbol{\gamma}(t_0))}\mathbb{M}$, time $t \in \mathbb{R}$

Ensure: Approximation of $\mathrm{P}_{\boldsymbol{\gamma},t_0,t}(\mathbf{w})$

if $|t - t_0| < \varepsilon$ **then**
  $N_{\text{steps}} \leftarrow 1$
else
  $N_{\text{steps}} \leftarrow \left\lfloor \frac{|t-t_0|}{\varepsilon} \right\rfloor$
end if
for $k = 0$ to $N_{\text{steps}} - 1$ **do**
  $t_k \leftarrow t_0 + \dfrac{k}{N_{\text{steps}}}(t - t_0)$
  $t_{k+1} \leftarrow t_0 + \dfrac{k+1}{N_{\text{steps}}}(t - t_0)$
  $\mathbf{P}_k \leftarrow \boldsymbol{\gamma}(t_k)$
  $\mathbf{P}_{k+1} \leftarrow \boldsymbol{\gamma}(t_{k+1})$
  procedure SCHILD'S LADDER($\mathbf{P}_k$, $\mathbf{P}_{k+1}$, $\mathbf{w}$):
    $\mathbf{P}'_0 \leftarrow \mathrm{Exp}(\mathbf{P}_k, \mathbf{w})$
    $\mathbf{P}_2 \leftarrow \mathrm{Exp}\big(\mathbf{P}'_0, 0.5\mathrm{Log}(\mathbf{P}'_0, \mathbf{P}_{k+1})\big)$
    $\mathbf{P}'_1 \leftarrow \mathrm{Exp}\big(\mathbf{P}_k, 2\mathrm{Log}(\mathbf{P}_k, \mathbf{P}_2)\big)$
    $\mathbf{w} \leftarrow \mathrm{Log}(\mathbf{P}_{k+1}, \mathbf{P}'_1)$
  end procedure
end for
Return: $\mathbf{w}$

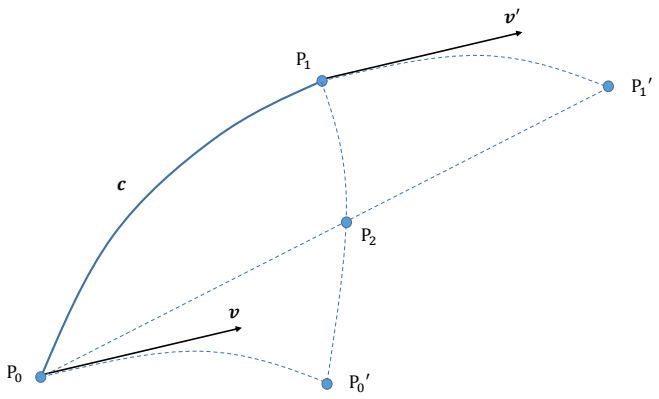


Figure 18 – Schild's Ladder: construction of one "geodesic parallelogram" to transport the tangent vector $\mathbf{v} \in \mathrm{T}_{\mathbf{P}_0}\mathbb{M}$ to P_1 along $\boldsymbol{\gamma}$.

Since the step size ε is ought to be small, the procedure described above is to be repeated several times to transport a tangent vector from a point on c to another.

VI.4.4.1.1 Influence on the runtime of the MCMC-SAEM

The Schild's Ladder can be used to replace the exact computation of the parallel transport on the Riemannian manifold $\mathrm{Spd}(n)$. We used this numerical scheme to approximate the parallel transport in the SPD matrices model. This model was then used to analyze the longitudinal dataset of symmetric positive definite matrices discussed in Section VI.4.1.2. For the Schild's ladder, we chose a step size $\varepsilon = 0.1$. As a result, one iteration of the MCMC-SAEM takes, on average, $53\,s$. In contrast to the runtimes presented in Section VI.4.1.3, we can observe that the MCMC-SAEM is, on average, 176 times slower with this numerical scheme. Note that the multiplicative factor 176 could be even larger if the step size ε of the Schild's Ladder was chosen smaller.

VI.4.4.2 Algorithms for the construction of an orthonormal basis

In Section IV.3.3.1, the Householder method and Gram-Schmidt algorithm were introduced as possible solution to compute an orthonormal basis of the vector space $\mathrm{Span}\big(\mathbf{G}(\mathbf{p}_0)\dot{\boldsymbol{\gamma}}_0(t_0)\big)^{\perp}$. Such an orthonormal basis is then used to define the space shifts and independent components according to Eq. [iv.55] and Eq. [iv.56]. In this section, we exploit these two methods to derive algorithms to compute the space shift of a given individual. Two algorithms are proposed, one based on the Householder method and the other on the Gram-Schmidt. This section ends with a comparison of both methods and a discussion regarding their computational cost.

Following the description of the Householder method in Section IV.3.3.1 and using the same notations, let $\mathbf{a}$ be the vector in $\mathbb{R}^N$ defined by:

$$\mathbf{a} = \mathbf{S}_0 + \mathrm{sgn}(\mathbf{S}_{0,1})\|\mathbf{S}_0\|\mathbf{e}_1 \qquad \text{[vi.83]}$$

where $\mathbf{S}_0 = \mathbf{G}(\mathbf{p}_0)\dot{\boldsymbol{\gamma}}_0(t_0)$ and $(\mathbf{e}_1, \ldots, \mathbf{e}_N)$ denotes the canonical basis of $\mathbb{R}^N$. The following algorithm is based on the Householder method and allows to compute the space shifts $(\mathbf{w}_i)_{1 \leq i \leq p}$:

Algorithm 11 Computation of the individual space shift $\mathbf{w}_i$ based on the Householder method

Require: N_s, $\mathbf{a}$ as in Eq. [vi.83], $(s_{l,i})_{1\leq l\leq N_s}$ as in Eq. [iv.55], $(\beta_{l,k})_{1\leq l\leq N_s, 1\leq k\leq N-1}$ as in Eq. [iv.56]

Ensure: Individual space shift $\mathbf{w}_i$

$\tilde{\mathbf{e}} \leftarrow \mathbf{0}$
$\mathbf{w}_i \leftarrow \mathbf{0}$
for $l = 1$ to N_s **do**
  for $k = 1$ to $(N-1)$ **do**
    $\tilde{\mathbf{e}} \leftarrow \mathbf{e}_{k+1} - 2\frac{(\mathbf{a}^\top \mathbf{e}_{k+1})}{\mathbf{a}^\top \mathbf{a}}\mathbf{e}_{k+1}$
    $\mathbf{w}_i \leftarrow \beta_{l,k}\tilde{\mathbf{e}} + \mathbf{w}_i$
  end for
  $\mathbf{w}_i \leftarrow s_{l,i}\mathbf{w}_i$
end for
Return: $\mathbf{w}_i$.

A second algorithm can be proposed using the Gram-Schmidt algorithm described in Section IV.3.3.1:

Algorithm 12 Computation of the individual space shift $\mathbf{w}_i$ based on the Gram-Schmidt algorithm

Require: N_s, $(\mathbf{v}_1, \ldots, \mathbf{v}_{N-1})$ a basis of $\mathrm{Span}\big(\mathbf{G}(\mathbf{p}_0)\dot{\boldsymbol{\gamma}}_0(t_0)\big)^\perp$, $(s_{l,i})_{1\leq l\leq N_s}$ as in Eq. [iv.55], $(\beta_{l,k})_{1\leq l\leq N_s, 1\leq k\leq (N-1)}$ as in Eq. [iv.56]

Ensure: Individual space shift $\mathbf{w}_i$

$\mathbf{w}_i \leftarrow \mathbf{0}$
for $l = 1$ to N_s **do**
  for $k = 1$ to $(N-1)$ **do**
    $\tilde{\mathbf{Q}}_k \leftarrow \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|}$
    $\mathbf{w}_i \leftarrow \beta_{l,k}\tilde{\mathbf{Q}}_k + \mathbf{w}_i$
    for $s = (k+1)$ to N **do**
      $\mathbf{v}_s \leftarrow \mathbf{v}_s - (\mathbf{v}_s^\top \tilde{\mathbf{Q}}_k)\tilde{\mathbf{Q}}_k$
    end for
  end for
  $\mathbf{w}_i \leftarrow s_{l,i}\mathbf{w}_i$
end for
Return: $\mathbf{w}_i$.

Note that the algorithm 12 uses the "Modified Gram-Schmidt" algorithm. This version of the Gram-Schmidt algorithm does the same computations as the usual Gram-

Schmidt algorithm but is known to be numerically more stable and less sensible to rounding errors. Still, the Householder method is usually numerically more stable than the Gram-Schmidt algorithm. In addition to numerical stability, these two algorithms for the computation of space shifts do not have the same cost.

VI.4.4.2.1 Comparison of the Householder method and Gram Schmidt algorithm

The most computationally expensive step of Algorithm 11 is the step `5:`.
Indeed, it can be decomposed as follows:

- $\mathbf{a}^\top \mathbf{e}_{k+1}$ (respectively $\mathbf{a}^\top \mathbf{a}$)): N multiplications, $(N-1)$ additions.
- Given $\mathbf{a}^\top \mathbf{e}_{k+1}$ and $\mathbf{a}^\top \mathbf{a}$, $2\dfrac{(\mathbf{a}^\top \mathbf{e}_{k+1})}{\mathbf{a}^\top \mathbf{a}}$ requires 2 multiplications.
- Given $2\dfrac{(\mathbf{a}^\top \mathbf{e}_{k+1})}{\mathbf{a}^\top \mathbf{a}}$, $2\dfrac{(\mathbf{a}^\top \mathbf{e}_{k+1})}{\mathbf{a}^\top \mathbf{a}}\mathbf{e}_{k+1}$ requires N multiplications.
- Finally, $e_{k+1} - 2\dfrac{(\mathbf{a}^\top \mathbf{e}_{k+1})}{\mathbf{a}^\top \mathbf{a}}\mathbf{e}_{k+1}$ requires N additions.

As a consequence, step `5:` of Algorithm 11 requires $2N+2(N-1)+2+2N = 6N$ floating-point operations. Step `6:` (respectively `8:`) requires $2N$ (respectively N) floating-point operations. Therefore, the innermost loop of the first algorithm requires $8N$ floating-point operations and steps `3:` to `9:` require: $N_s\big(8N(N-1)+N\big) = 8N^2N_s - 7NN_s$ floating point operations.

Similarly, step `7:` of Algorithm 12 requires $4N-1$ floating-point operations. Indeed, it is decomposed as follows:

- $\mathbf{v}_s^\top \tilde{\mathbf{Q}}_k$: N multiplications, $(N-1)$ additions.
- Given $\mathbf{v}_s^\top \tilde{\mathbf{Q}}_k$, $(\mathbf{v}_s^\top \tilde{\mathbf{Q}}_k)\tilde{\mathbf{Q}}_k$ requires N multiplications.
- Given $(\mathbf{v}_s^\top \tilde{\mathbf{Q}}_k)\tilde{\mathbf{Q}}_k$, $\mathbf{v}_s - (\mathbf{v}_s^\top \tilde{\mathbf{Q}}_k)\tilde{\mathbf{Q}}_k$ requires N additions.

As a consequence, step `7:` requires $4N-1$ floating-point operations. It follows that steps `6:` to `8:` require: $\sum_{s=k+1}^{N}(4N-1) = (4N-1)(N-k)$ floating-point operations.

Steps `4:` (respectively `5:`) require N (respectively $2N$) elementary operations. Therefore, steps `2:` to `11:` require

$$\begin{aligned} & N_s\left[\sum_{k=1}^{N-1}\Big(\sum_{s=k+1}^{N}(4N-1)+3N\Big)+N\right] \\ &= N_s\sum_{k=1}^{N}\big((4N-1)(N-k)+3N\big)+NN_s \\ &= N_s\sum_{k=1}^{N-1}\big(4N^2+2N+k-4Nk\big)+NN_s \\ &= N_s\Big[2N^3+\frac{N(N-1)}{2}-N\Big] \end{aligned} \qquad \text{[vi.84]}$$

floating-point operations.

In the end, the cost (in time) of the Algorithm 11 is $O(N^2)$ whereas it is $O(N^3)$ for the Algorithm 12. In addition to this, Algorithm 11, based on the Householder method, only requires to store $\tilde{\mathbf{e}}$ and $\mathbf{a}$, which are two vectors in $\mathbb{R}^N$. Algorithm 12, based on the modified Gram-Schmidt algorithm, requires to store the set $(\mathbf{v}_1,\ldots,\mathbf{v}_{N-1})$ which consists in $(N-1)$ vectors in $\mathbb{R}^N$. It follows that the memory footprint of the first algorithm is $O(N)$ whereas it is $O(N^2)$ for the second algorithm. Note that the first algorithm repeats, each time it is called, the same computations in step 5 :. This is because the orthonormal basis of $\mathrm{Span}\big(\mathbf{G}(\mathbf{p}_0)\dot{\boldsymbol{\gamma}}_0(t_0)\big)$ is not stored. Judging by the runtime and memory footprint of both algorithms, the first algorithm seems to be a better choice, especially when dealing with high-dimensional observations.

Part VII

Estimation of digital models of progression from health data

Summary

VII.1 Motivation

This chapter proposes experimental results obtained with the particular cases of the generic spatiotemporal model. These experiments on longitudinal datasets of health data aim at showing that the generic spatiotemporal model allows to estimate propagation models which provide insightful informations on the evolution of a specific biological phenomenon. In Section VII.2, we propose a method to assess how well the generic spatiotemporal model allows to put into correspondence individual trajectories of progression. This evaluation method is used in the different experiments presented below. We show that, in various situations, the time reparametrization estimated by the MCMC-SAEM provides a good correspondence between individuals.

In Section VII.3, we analyze longitudinal datasets of neuropsychological test scores. These test scores provide a measure of the impairment of cognitive functions among a group of healthy individuals and individuals diagnosed with Alzheimer's disease. By analyzing these neuropsychological test scores with the logistic curves models, we derive normative data-driven scenarios of the impairment of cognitive functions during the course of Alzheimer's disease. These normative scenarios provide an ordering in which the cognitive functions are impaired, as well as the relative timing between these impairments. We also show that the individual acceleration factors and time shifts, estimated a posteriori from the individual data, successfully capture the temporal variability in the measurements.

Section VII.4 proposes an analysis of a longitudinal dataset of cortical thickness measurements. With this dataset, we were able to find anatomical regions where the cortical atrophy progresses faster than others. These findings are consistent with previous knowledge on Alzheimer's disease pathophysiology.

In Section VII.5, we propose the analysis of a longitudinal dataset of body fat measurements, collected from a population of pre-menarcheal young girls. It is known that the percentage of body fat increases after menarche. In this dataset, the time of menarche is known for each individual. The menarche is a biological event which leads to changes in the metabolism, and in the body fat measurements. This could be considered as the equivalent of the onset of Alzheimer's disease. Therefore, the body fat dataset offers the advantage that the time of occurrence of this specific event is known for each individual, unlike the age at onset in Alzheimer's disease.

Finally, Section VII.6 proposes the analysis of a synthetic dataset of 3×3 symmetric positive definite matrices. This dataset was not obtained from a clinical study, but is intended to represent longitudinal datasets which could be obtained in Diffusion Tensor Imaging (DTI). DTI is a medical imaging technique which aims at characterizing the structure of biological tissues organized in fibers (like the white matter in the brain or cardiac fibers) by estimating the directions in which water molecules diffuse. Moreover, analyzing this longitudinal dataset with the SPD matrices model allows to validate the model and the algorithm in a Riemannian manifold of negative sectional curvature.

VII.2 Experimental setup and evaluation criteria

In this section, we propose a method to evaluate how well the individual time reparametrizations, introduced in Section IV.3.1, allow to put in correspondence the evolution of the individuals. In order to do this, we will use an additional information which is not used in the model: the time at which a particular event occurs in the life of an individual. For example, the event could be the time of onset of a disease, if this information is known, or a time at which a change occurs in the metabolism. In sequences of images of people smiling, the event could be the time at which the muscles of the face relax and the person stops smiling. The event occurs at a different time point for each individual. The individual time reparametrization ψ_i aims at putting

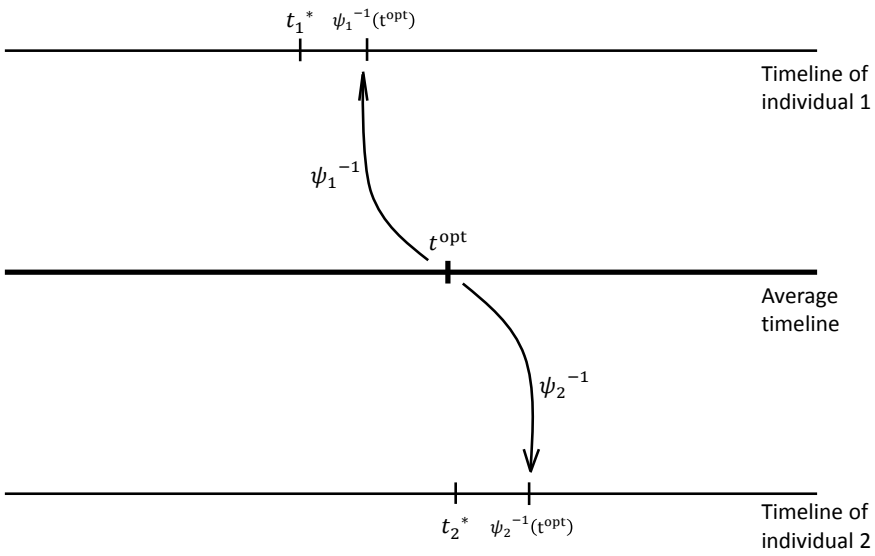


Figure 19 – The average time of event t^{opt} is mapped to the individual timelines using ψ_i^{-1}.

in correspondence the progression of the different individuals. For the ith individual, ψ_i maps the timeline of this individual to the "average timeline", namely the timeline of the average trajectory. Let t_i^* be the time point at which the event occurs in the timeline of the ith individual. Given this information, we define t^{opt} as the time point, in the average trajectory γ_0, corresponding to the occurrence of the event. This time point t^{opt} is obtained by minimizing the sum of errors $E : t \mapsto \sum_i |t_i^* - \psi_i^{-1}(t)|$. Note that t^{opt} can be interpreted as a median of the normalized ages $(\psi_i(t_i^*))_i$, and could therefore not be unique. Then given t^{opt} which minimizes the sum of errors E, this time point is mapped from the average timeline to the individual timelines by the mappings ψ_i^{-1}, as illustrated in Figure 19. If the time-reparametrization ψ_i allowed for an

exact correspondence between the average timeline and the timeline of the ith individual, the corresponding age $\psi_i^{-1}(t^{\text{opt}})$ would actually be t_i^*. In practice, the difference $|t_i^* - \psi_i^{-1}(t^{\text{opt}})|$ allows to quantify how well the timeline of the ith individual and the average timeline have been put into correspondence.

In the experiments considered in the following sections, the median t^{opt} of $\left(\psi_i(t_i^*)\right)_{1\leq i\leq p}$ was computed unambiguously. To assess how well the individual trajectories and the average trajectory have been put into correspondence, we proposed an histogram of the errors $\left(|t_i^* - \psi_i^{-1}(t^{\text{opt}})|\right)_{1\leq i\leq p}$.

VII.3 Neuropsychological test scores

VII.3.1 The datasets

In this section, three longitudinal datasets are considered. All these datasets consist in normalized neuropsychological test scores collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The neuropsychological test used is the modified "ADAS-Cog" test [Mohs et al., 1997]. Each of the 13 items of the test measures the impairment of either memory, language, concentration or praxis. The sum of the 13 items is marked out of 85. The higher the score, the more impaired the cognitive function.

(i) For the first longitudinal dataset, the scores to the 13 items were summed and normalized by the maximum possible value, 85. Therefore, the observations are points in $]0,1[$ (no individual had a score equal to 0 or to 85). These observations were collected for 1393 individuals from the ADNI1, ADNI2 and ADNIGo cohorts of the ADNI database, with an average of 4 time points per individual (min: 3 ; max: 11). Results obtained with this dataset are presented in Section VII.3.2. Elements of this section are taken from [Schiratti et al., 2015d].

(ii) The second longitudinal dataset was obtained by grouping the items by cognitive function (memory, language, praxis and concentration). For each cognitive function, the sum of the scores was normalized by the maximum possible value. Therefore, the observations are points in $]0,1[^4$. These observations were collected for individuals, from the ADNI1 cohort, with mild cognitive impairment (MCI) who converted to Alzheimer's disease (AD) during the observation period. A total of 248 individuals were included in this dataset. Results with this dataset are presented in Section VII.3.3. Elements of this sections are taken from [Schiratti et al., 2015a].

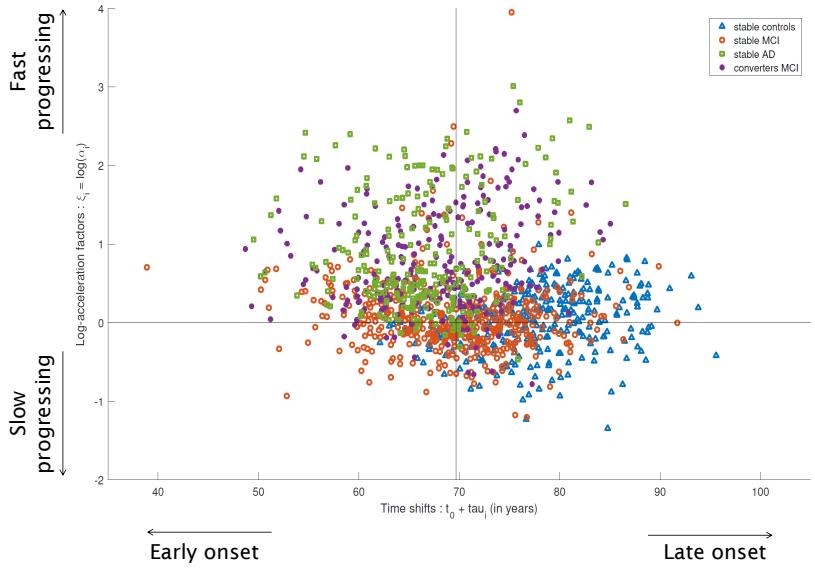
(iii) The third longitudinal dataset consists in the same population as for the second one. However, instead of grouping the scores by cognitive functions, each item

score was normalized by its maximum possible value, resulting in observations in $]0,1[$[13]. The results obtained with this dataset are presented in Section VII.3.4. Elements of this section are taken from [Schiratti et al., 2015b].

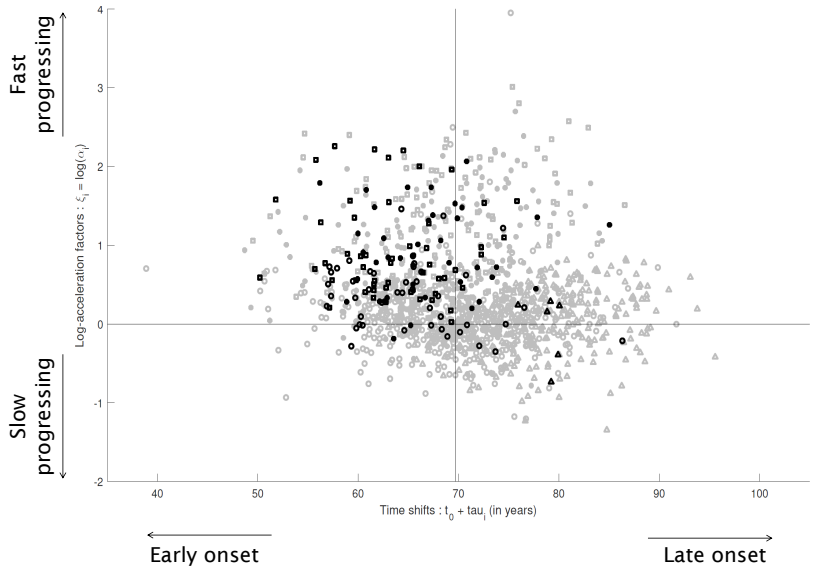
Since the observations of the first longitudinal dataset are normalized *univariate* measurements, this dataset was analyzed using the univariate logistic curves model (see Eq. [v.7]). The two other datasets, which consist in *multivariate* normalized observations, were analyzed using the unstructured logistic curves propagation model (see Eq. [v.19]). The parameters of each model was estimated using the MCMC-SAEM algorithm.

VII.3.2 Results with observations in $]0,1[$

At each visit, a diagnosis (healthy, mild cognitive impairment, Alzheimer's disease) was given to each individual by a clinician. The evolution of the diagnosis sequence with time allowed to group the 1393 individuals into 4 groups of interest: "stable controls", "stable mild cognitive impairment (MCI)", "stable Alzheimer's Disease (AD)" and "MCI converters". The group "stable controls" (resp. "stable MCI", "stable AD") consists in individuals who were diagnosed *healthy* (resp. *mild cognitive impairment*, *Alzheimer's Disease*) at each visit. These MCI individuals are not considered as healthy, nor as Alzheimer patients. MCI might be considered as a transition state between a healthy state and Alzheimer's disease (AD). Finally, "MCI converters", consists in individuals who were diagnosed as MCI at their first visit and converted to AD by the end of the observation period. Among the 1393 individuals of this longitudinal dataset, 329 individuals are stable controls, 472 are stable MCI, 248 are stable *AD* and 248 are MCI converters. The 96 remaining are individuals who converted from control to MCI (54 out of 96), who converted from control to AD (3 out of 96), who reverted from AD to MCI (3 out of 96) or who reverted from MCI to control (36 out of 96). These 96 individuals were included in the estimation of the parameters of the model but not in the plot of the estimates of the individual random effects, in Figure 20. The MCMC-SAEM, used with the logistic curves model, allowed to estimate an average trajectory characterized as follows: it is the logistic curve which goes through the point $p_0 = 0.11$ at time $t_0 = 69.68$ years with velocity $v_0 = 0.0076$ units per year. The estimated variance parameters $\sigma_\xi = 0.94$ and $\sigma_\tau = 10.42$ years inform on the temporal variability of the progression among the population. To illustrate this temporal variability, Figure 20 (a) presents a plot of the maximum a posteriori (MAP) estimates of the individual random effects (ξ_i, τ_i) across the four groups of interest. We can observe that stable controls have larger time shifts than other groups. On the one hand, the stable controls are mainly late-onset individuals who are not, on average, evolving faster than the average disease progression trajectory. On the other hand, stable Alzheimer patients and MCI individuals tend to have smaller time shifts than stable controls. A portion of the stable MCI and most of the stable Alzheimer patients can be considered as early-onset individuals. It appears clearly that stable Alzheimer patients and



(a)



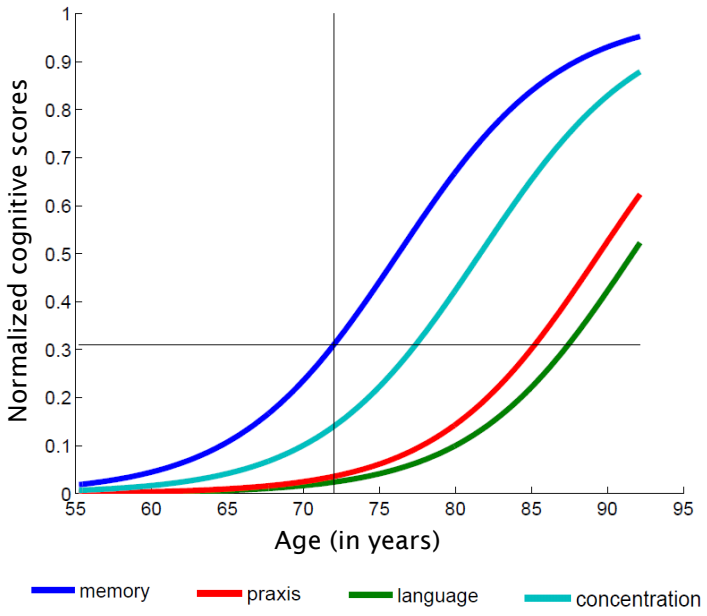
(b)

Figure 20 – **Figure (a)**: plot of (MAP estimates of) the individual log-acceleration factor $\xi_i = \log(\alpha_i)$ against the time shifts $t_0 + \tau_i$. Each point corresponds to an individual. The parameter $t_0 = 69.68$ years was estimated with the MCMC-SAEM. **Figure (b)**: the points of Figure (a) are colored in black if the individual has 4/4 APOE genotype. Other genotypes are colored in grey.

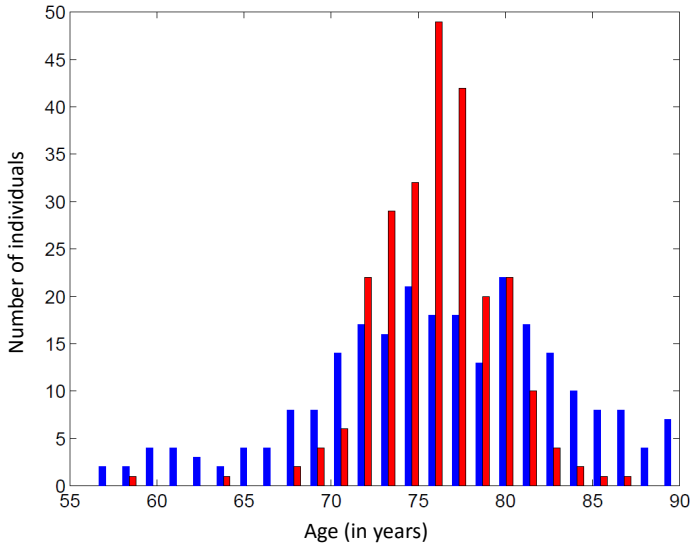
MCI converters are fast progressing individuals even though a small number of MCI converters are slow-progressers. On average, stable MCI and stable controls are not progressing faster than the average disease progression scenario. These observations are coherent with the diagnoses of the individuals and the subject-specific random effects allow to distinguish groups of individuals. The disease progression for Alzheimer patients and converters MCI is faster and started earlier than for stable controls. The Alzheimer patients are quite clearly separated from the stable controls. In addition to the information provided by the individual random effects, genetic information was considered to obtain the results presented in Figure 20 (b). Apolipoprotein E (APOE) locus on chromosome 19 is a gene which is known to be a strong risk factor for AD. This gene has three alleles: APOE-$\varepsilon 2$, APOE-$\varepsilon 3$ and APOE-$\varepsilon 4$. In [Corder et al., 1993], the authors find that individuals with an $\varepsilon 4 - \varepsilon 4$ genotype are eight times more likely to be affected by AD than individuals with $\varepsilon 2 - \varepsilon 3$ or $\varepsilon 3 - \varepsilon 3$ genotypes. APOE is also a strong risk factor in familial forms of AD. In Figure 20 (b), each point (or individual) is colored according to its genotype. Only the individuals with $\varepsilon 4 - \varepsilon 4$ genotype were colored in black. This result shows that, on average, the $\varepsilon 4$ homozygotes are fast progressers who are evolving ahead of the average trajectory. Indeed, even though some of the $\varepsilon 4 - \varepsilon 4$ individuals appear as slow progressers who evolve behind of the average trajectory (bottom right quadrant), most of these individuals are fast progressers and in the left quadrants (fast progressers). These results are coherent with the fact that an $\varepsilon 4 - \varepsilon 4$ genotype is a risk factor for AD.

VII.3.3 Results with observations in $]0,1[^4$

As mentioned above, this longitudinal dataset of multivariate observations was analyzed with the logistic curves propagation model (see Eq. [v.19]). In contrast to the model used in the previous section, this model includes space shifts which aim a capturing and estimating the distribution of the directions of the trajectories on the manifold. Recall that, before using this model, a number N_s of independent components must be chosen (see Section IV.3.3). Given that $\mathbb{M} =]0,1[^4$ is a four-dimensional Riemannian manifold, the number N_s of independent components could have been either $1, 2$ or 3. The model with two and tree independent sources allowed to better explain the total variance and reduce the residual noise. Indeed, the model with one independent component estimated a residual noise variance $\sigma^2 = 0.012$ and explained 79% of the total variance, whereas the model with two (respectively three) independent components estimated a noise variance $\sigma^2 = 0.008$ (respectively $\sigma^2 = 0.0084$) and explained 84% (respectively 85%) of the total variance. Because the results obtained with three independent components are similar to the results obtained with two independent components, we choose, for the sake of clarity, to report the results obtained with two components ($N_s = 2$). The average trajectory estimated by logistic curves propagation model, plotted in Figure 21 (a), is characterized by the fixed effects $p_0 = 0.3$, $t_0 = 72$ years old, $v_0 = 0.04$ units per year and $\boldsymbol{\delta} = [0; -15; -13; -5]$ years. The first

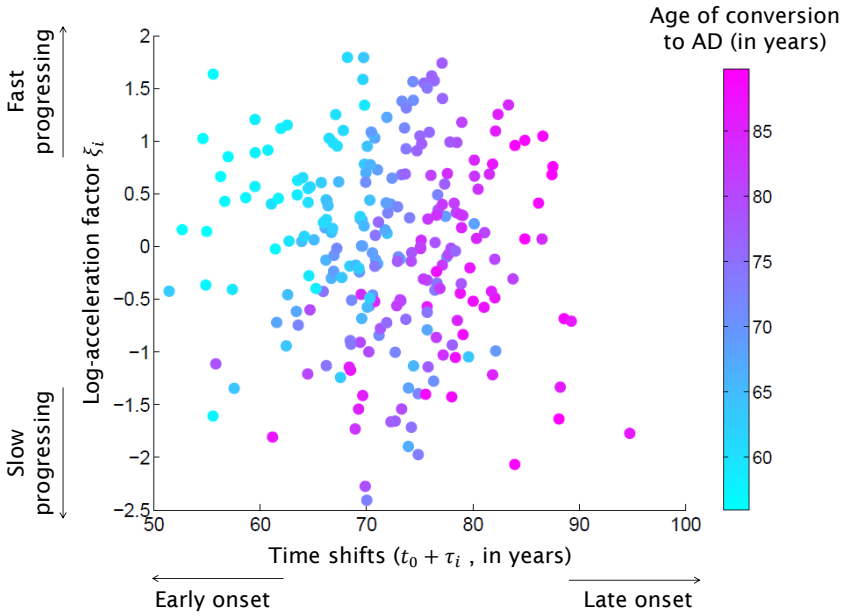


(a)

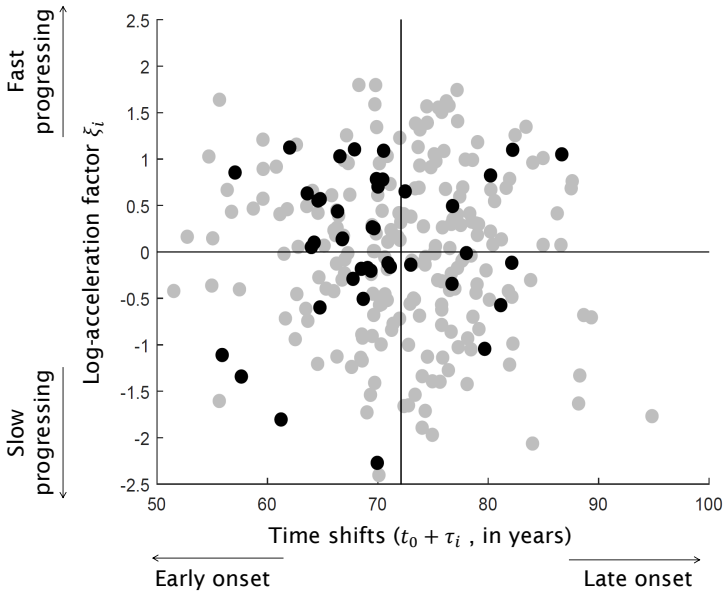


(b)

Figure 21 – **Figure (a)**: the average trajectory estimated with the MCMC-SAEM. The estimated parameters p_0 (resp. t_0) are represented by an horizontal (resp. vertical line) at $p_0 = 0.3$ (resp. $t_0 = 72$ years). **Figure (b)**: histogram of the ages $(t_i^{\text{diag}})_{1 \leq i \leq p}$ at which individuals converted to AD (in blue) in blue ; Histogram of the normalized ages $\left(\psi_i(t_i^{\text{diag}})\right)_{1 \leq i \leq p}$ in red.



(a)



(b)

Figure 22 – **Figure (a)**: Plot of $t_0 + \tau_i$ with respect to the log-acceleration factor ξ_i. Each point is colored with respect to the estimated age of conversion to AD. **Figure (b)**: the points of Figure (a) are colored in black if the individual has 4/4 APOE genotype. Other genotypes are colored in grey.

cognitive function, *memory*, reaches the value $p_0 = 0.3$ at 72 years, on average. The second cognitive function to reach the same value is *concentration*, at $t_0 + 5 = 77$ years on average. The progression of these two cognitive functions is followed by *praxis* and

language. The fixed effect of the model provide an ordering of the cognitive functions and the relative delay between two given cognitive functions. The random effects of the model characterize the spatiotemporal variability of the average trajectory among the population. Indeed, the time-shifts allow to determine whether an individual is evolving ahead or behind the average trajectory and account for the variability in age at disease onset. To illustrate the role played by the time shifts and acceleration factors, maximum a posteriori estimates of the individual time-shits and log-acceleration factors are plotted in Figure 22 (a). This figure shows a clear correspondence between the time shifts and the estimated age of conversion to AD. Indeed, the individuals with a negative (resp. positive) time-shift, the individuals evolving ahead (resp. behind) the average trajectory, are the ones who convert young (resp. late) to AD. In other words, the normalized age $\psi_i(t)$ is, as opposed to age, a good indicator of disease progression. In order to show that the individual time reparametrizations, defined by the estimated

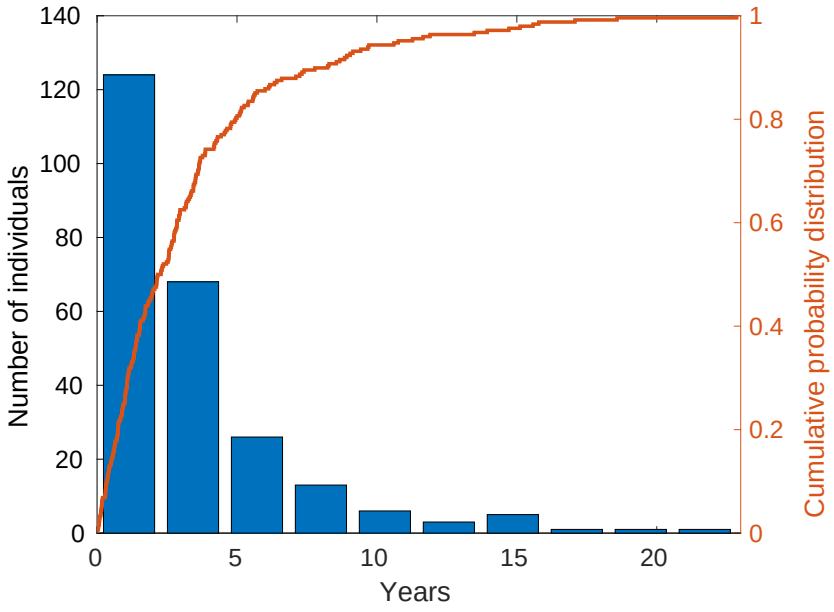


Figure 23 – Histogram of the errors $\left(|t_i^{\text{diag}} - \psi_i^{-1}(t^{\text{opt}})|\right)_{1\leq i\leq 248}$ superimposed with the cumulative distribution of these errors, with $t^{\text{opt}} = 74.30$ years.

time shifts and acceleration factors, actually put into correspondence individuals, we used the evaluation criteria described in Section VII.2. For each individual, we derived an age of conversion to Alzheimer's disease. This age, denoted by t_i^{diag} for the ith individual, is defined as the mean between his age at the last time point where he was MCI and his age at the first time point he was diagnosed with Alzheimer's disease. The histogram given in Figure 23 shows that the estimated age of conversion $\psi_i^{-1}(t^{\text{opt}})$, with $t^{\text{opt}} = 74.30$ years, is a good estimation of t_i^{diag}. Indeed, this error is less than 3 years for 62% of the individuals. In addition to this, the effect of the time reparametrizations is illustrated in Figure 21 (b). This figure shows (in blue) an his-

togram of the ages of conversion $(t_i^{\text{diag}})_{1\leq i\leq p}$. The histogram (in red) of the "normalized ages" $\left(\psi_i(t_i^{\text{diag}})\right)_{1\leq i\leq p}$ was superimposed to the previous one. We can observe that the histogram of the normalized ages is peaked around 77 years, with a smaller variance compared to the distribution of the ages of conversion.

In Figure 22 (b), each point (or individual) is colored in black if the corresponding individual has a $\varepsilon 4$-$\varepsilon 4$ genotype. In the considered population of 248 converters MCI, 41 have an $\varepsilon 4$-$\varepsilon 4$ genotype. If the quadrants are numbered from one to four in a trigonometry fashion), 43% of the $\varepsilon 4$-$\varepsilon 4$ individuals are in the second quadrant (early onset and fast progressers) and 29% of these individuals are in the third quadrant (early onset and slow progressers). Among the 41 individuals, 73% are early-onset individual. By comparing the black dots in Figure 22 (b) with the colored dots of Figure (a), one can observe that some $\varepsilon 4$-$\varepsilon 4$ individuals actually convert quite late to AD (purple dots). However, the logistic curves propagation model estimated that these individuals tend to evolve ahead of the average individual.

Figure 24 illustrates the spatiotemporal variability of the average trajectory among the population. The temporal variability is characterized by the time shifts and acceleration factors. For this longitudinal dataset, the estimated standard deviation of the time shifts equals $\sigma_\tau = 7.5$ years. Therefore, the average trajectory is shifted by ± 7.5 years for 95% of the population (second row of Figure 24). The estimated standard deviation of the log-acceleration factors $(\xi_i)_{1\leq i\leq p}$ is $\sigma_\xi = 0.9$. As a consequence, most of the individuals are progressing between $e^{\sigma_\xi} \simeq 2.4$ times faster or $e^{-\sigma_\xi} \simeq 0.4$ times slower than the average trajectory (first row of Figure 24). The spatial variability is characterized by the random variations of the space-shifts. Let $\mathbf{A}_1$ (respectively $\mathbf{A}_2$) denote the first (respectively second) estimated independent component. Third row of Figure 24 shows that individuals with a space shift of the form $\mathbf{w}_i = \sigma_{s_i}\mathbf{A}_1$ have memory and concentration impaired almost simultaneously. For these individuals, the ordering between language and praxis is not changed. Individuals with a space-shift of the form $\mathbf{w}_i = -\sigma_{s_i}\mathbf{A}_1$ have language and praxis impaired nearly at the same time and the ordering of these two cognitive functions is slightly changed. We also note that the relative delay between memory and concentration varies greatly for individuals with space shifts in the direction of the first independent component. The effect of the second independent component $\mathbf{A}_2$ is illustrated in the last row of Figure 24. We note that in this direction, the relative timing between memory and concentration is not greatly changed but the ordering between language and praxis changes. These results show that the cognitive functions tend to evolve by pairs: memory & concentration, language & praxis. The individual space-shift impacts the relative delay and the ordering of these cognitive functions.

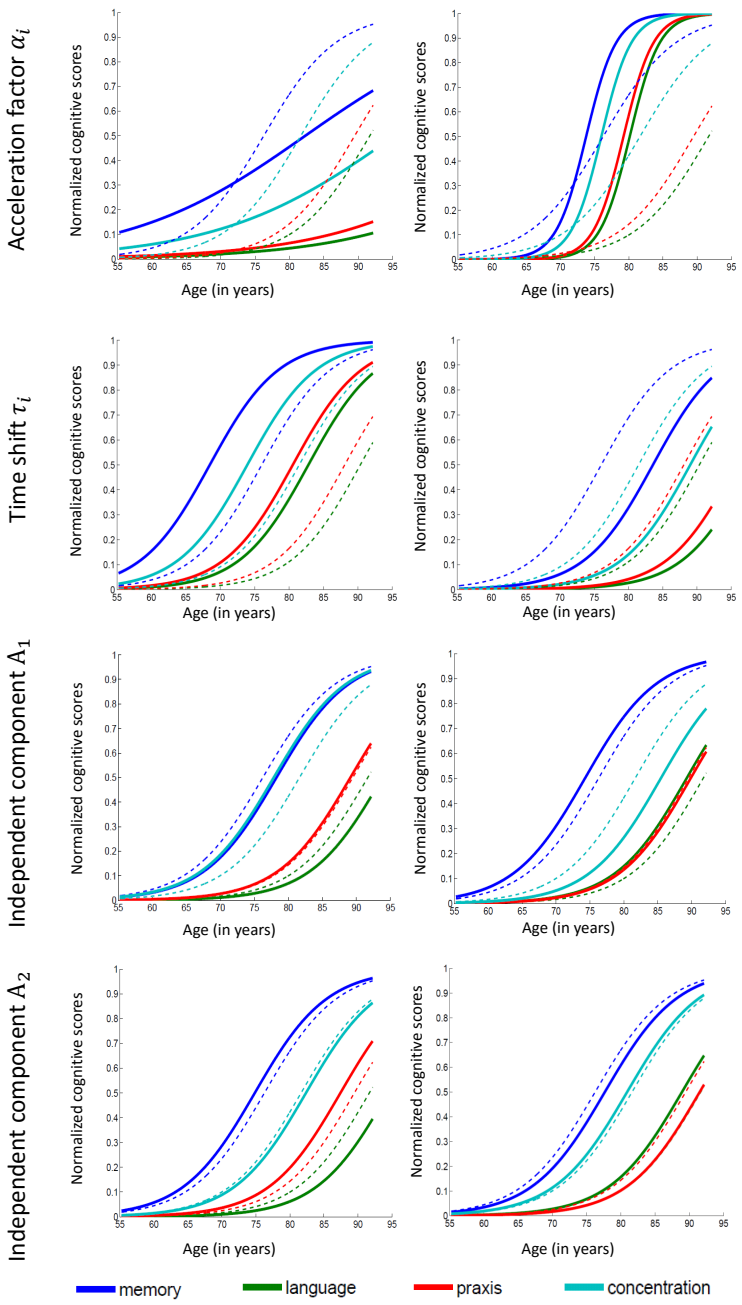


Figure 24 – Variability of the average trajectory in terms of space-shift and acceleration factor. First (respectively second row): plot of $t \mapsto \boldsymbol{\gamma}_{0,\boldsymbol{\delta}}(e^{\mp\sigma_\xi}(t-t_0)+t_0)$ (respectively $t \mapsto \boldsymbol{\gamma}_{0,\boldsymbol{\delta}}(t \pm \sigma_\tau)$) with $t_0 = 72$ years and $\sigma_\xi = 0.9$, $\sigma_\tau = 7.5$ years. Third (respectively fourth) row: plots of parallels $\eta^{\mp\sigma_{s_i}\mathbf{A}_i}(\boldsymbol{\gamma}_{0,\boldsymbol{\delta}},\cdot)$ $(i=1,2)$ in the direction given by the two estimated independent components, with $\sigma_{s_1} = 2.66$ and $\sigma_{s_2} = 2.68$ are the standard deviations of the estimates $(s_{1,i})_{1\leq i\leq p}$ and $(s_{2,i})_{1\leq i\leq p}$.

VII.3.4 Results with observations in $]0,1[^{13}$

For the third longitudinal dataset, the observations $(\mathbf{y}_{i,j})_{i,j}$ are considered perturbations of a point in $]0,1[^{13}$, where each component of the vector $\mathbf{y}_{i,j}$ corresponds to the normalized score of a specific task or question of the ADAS-Cog test. The results obtained without averaging the scores are presented below. As one might expect, these results are similar with the ones presented above, where the items of the ADAS-Cog test were averaged by cognitive function.

The MCMC-SAEM was used with the logistic propagation model to analyze this longitudinal dataset. The algorithm was run with $N_s = 1, 2, 3$ and, as before, we note that increasing the number of sources allowed to decrease the residual noise among the experiments: $\sigma^2 = 0.02$ for $N_s = 1$, $\sigma^2 = 0.0162$ for $N_s = 2$ and $\sigma^2 = 0.0159$ for $N_s = 3$. Because the residual noise was almost similar for $N_s = 2$ and $N_s = 3$ sources, we choose to report here the results obtained with the less complex model. As a consequence, we report the results obtained with 2 independent sources. The average trajectory is given in Figure 25, where each curve rep-

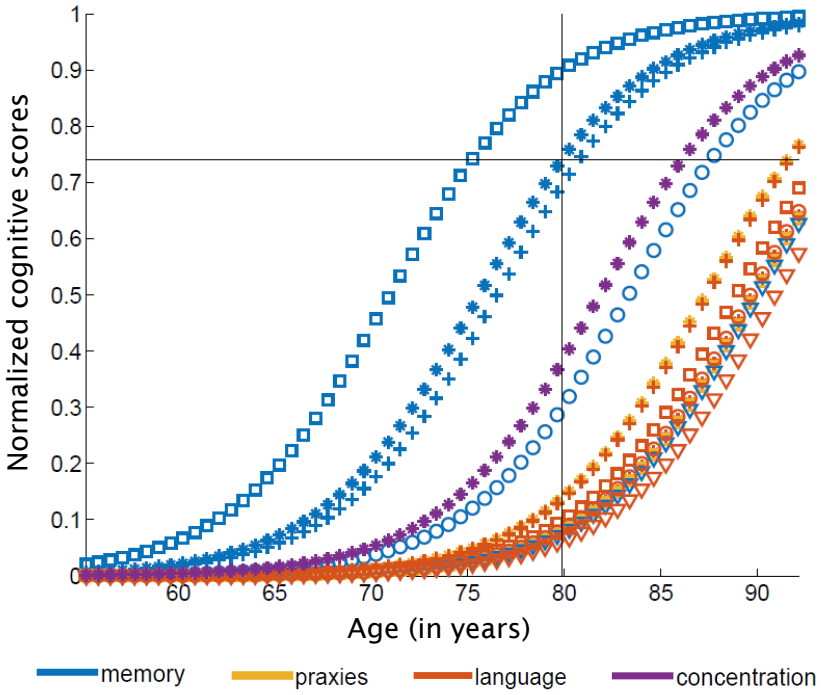


Figure 25 – The estimated average trajectory. In blue: the average trajectory of progression for the 5 memory-related items (item 1:∗, item 4:□, item 7:○, item 8:+ and item 9:△). In orange: average trajectory for the 5 language-related items (item 2:∗, item 5:□, item 10:○, item 11:+ and item 12:△). In yellow: average progression trajectory for the 2 praxis-related items (item 3:∗ and item 6:□). In purple: average progression trajectory for the concentration-related item (item 13:∗).

resents the temporal progression of one specific item of the ADAS-Cog test. The estimated fixed effects are $p_0 = 0.74$, $t_0 = 79.88$ years, $v_0 = 0.047$ unit per year, and $\boldsymbol{\delta} = [0; -14; -11; 4.6; -13; -14; -7.7; -0.9; -14.4; -14.05; -11.80; -15.3292]$ years. This means that, on average, the memory-related items (items $1, 4, 7, 8, 9$) reach the value $p_0 = 0.74$ at respectively $t_0, t_0 - \delta_4, t_0 - \delta_7, t_0 - \delta_8$ and $t_0 - \delta_9$ years, which corresponds to respectively 79.88, 75.2, 87.6, 80.7 and 94.3 years. The concentration item reaches the same value at $t_0 - \delta_{13} = 86.1$ years. The progression of the concentration item is followed by praxis and language items.

The estimated standard deviation of the time shifts is $\sigma_\tau = 8.3$ years. Recall that with the previous dataset, the estimated standard deviation was $\sigma_\tau = 7.5$ years. Therefore, the average disease propagation model is shifted by ± 8.3 years for 95% of the population, which is quite similar to the shift obtained in the previous experiment. The standard deviation of the log-acceleration factor is $\sigma_\xi = 0.8$, whereas it was 0.9 for the previous experiment. Again, the variability of the log-acceleration factor among the population is similar to the one obtained above. The effect of the acceleration factor is illustrated in the first row of Figure 26. In the second and third rows of Figure 26, the first and second independent components illustrates the spatial variability in the measurements and the variability in the relative timing of the cognitive impairments. The first independent direction shows that some memory items and language items are shifted in time with respect to the other ones, especially for memory item 4 (□) and item 7 (○). The ordering of the memory item 7 (○) and the concentration item is inverted for individuals with a space shift $\mathbf{w}_i = -\sigma_{s_{i,1}} A_1$. For those individuals, praxis items are impaired later, after the language items 2 (∗), items 12 (△) and item 5 (□). The second independent component shows a greater variability for the memory-related items than for the first independent components, in particular for memory item 9 (△) and item 4 (□). For individuals with a space shift $\mathbf{w}_i = \sigma_{s_{i,2}} A_2$, language-related items might be impaired later than the average individual, especially for the language item 12 (△). The estimated log-acceleration factor ξ_i and time shift τ_i are plotted for each individual in Figure 27 (a). As for Figure 22 (a), we can observe that the individuals who have a positive (respectively negative) time shift (they are evolving ahead, respectively behind, the average trajectory) are the individuals who converted late (respectively early) to AD. This means that the individual time-shifts correlate well with the age at which a given individual was diagnosed with AD. However, we can note that, in this experiment, there is a negative correlation, equal to -0.4, between the estimated log-acceleration factors and time shifts. This means that there is a tendency for early onset patients to be fast progressers. As with the previous dataset, the evaluation criteria described in Section VII.2 is used to evaluate how well the estimated time reparametrizations put individuals into correspondence. The histogram given in Figure 27 (b) shows that the error on the estimation of the age of conversion is similar to the one obtained in the previous section.

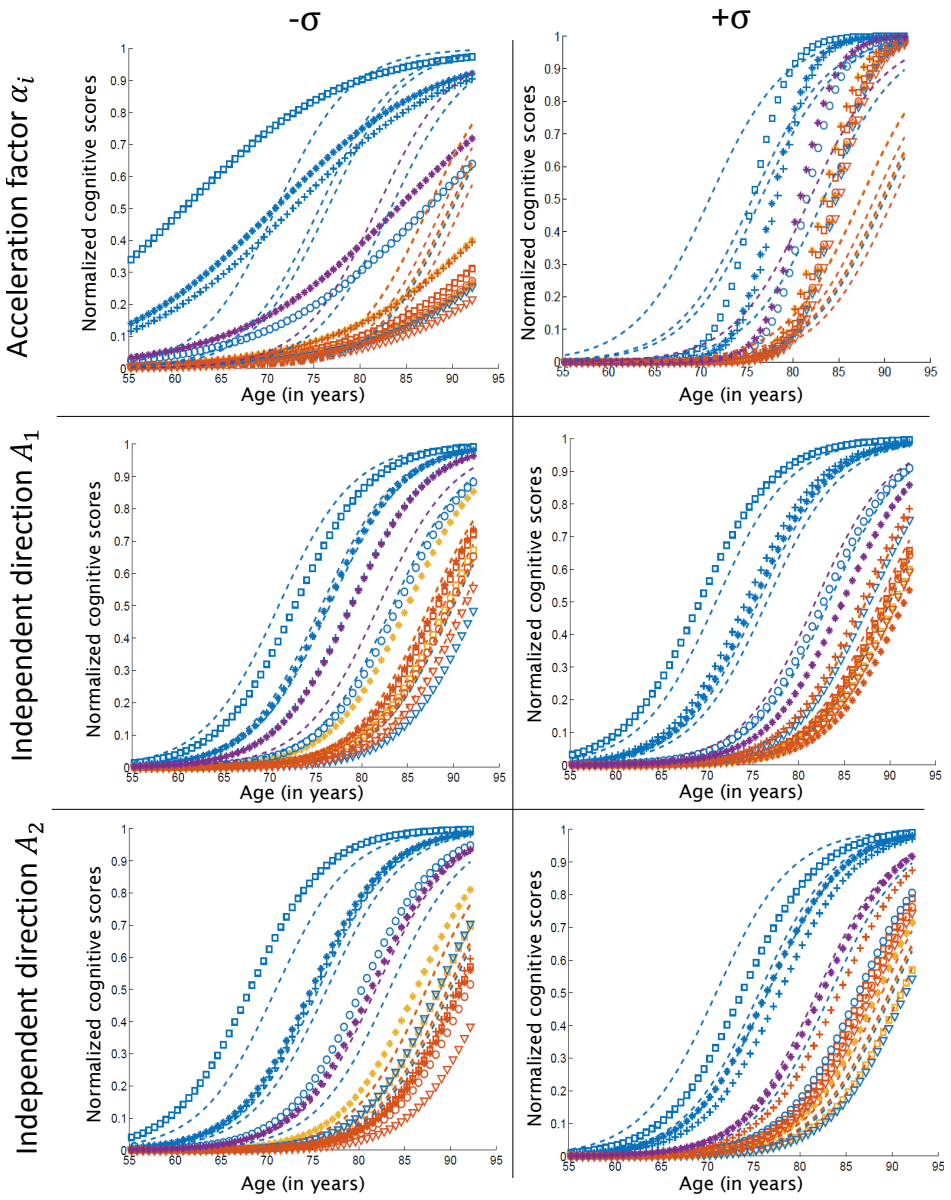
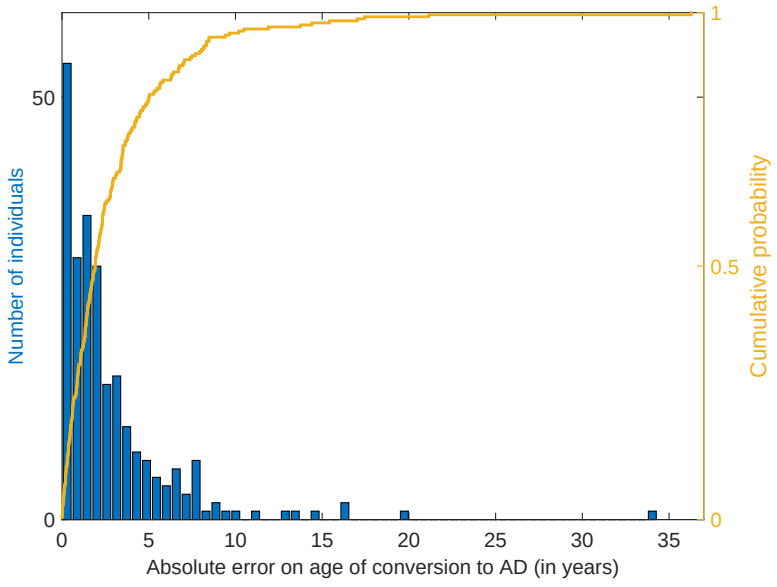


Figure 26 – First row: plot of $t \mapsto \boldsymbol{\gamma}_{0,\boldsymbol{\delta}}(\alpha(t-t_0)+t_0)$ with $\alpha = \exp(\mp\sigma_\xi)$ and $\sigma_\xi = 0.8$. Second (respectively third) rows: plots of parallels $\eta^{\mp\sigma_{s_i}\mathbf{A}_i}(\boldsymbol{\gamma}_{0,\boldsymbol{\delta}}, \cdot)$ in the direction given by the two estimated independent components, with $\sigma_{s_1} = 2.88$ and $\sigma_{s_2} = 2.54$ are the standard deviations of the estimates $(s_{1,i})_{1\leq i\leq p}$ and $(s_{2,i})_{1\leq i\leq p}$.

(a)



(b)

Figure 27 – **Figure (a)**: plot of the (MAP estimates of the) subject-specific random effects: the log-acceleration factor ξ_i is plotted against the time-shifts $t_0 + \tau_i$. Each point is colored according to the age of conversion to AD. **Figure (b)**: histogram of the errors $\left(|t_i^{\text{conv}} - \psi_i^{-1}(t^{\text{opt}})|\right)_{1 \leq i \leq 248}$ superimposed with the cumulative distribution of these errors.

VII.4 Cortical thickness measurements

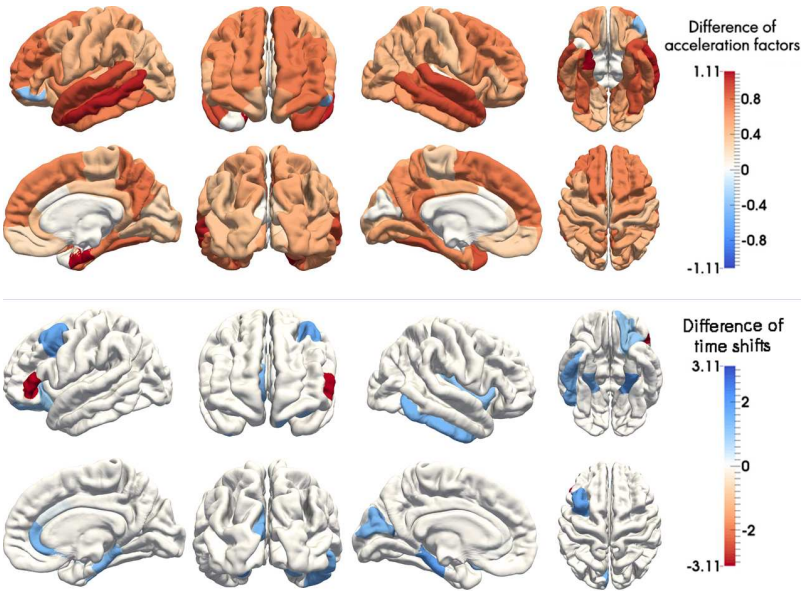
VII.4.1 The dataset

The longitudinal data used to obtain the experimental results presented below were obtained from the ADNI database. The cortical thickness measurements were collected for 725 individuals. On average, individuals have 5 time points (min: 3 ; max: 7). The longitudinal follow-up period of these individuals ranges from 6 months to 4.5 years, with an average of 2.5 years. For 95% of the population, the longitudinal follow-up period ranges from 1.5 year to 3.5 years. Diagnoses were recorded for every individual and at each visit. As for the neuropsychological test scores (see Section VII.3), these subject-specific sequences of diagnoses allowed to classify the individuals into 4 groups of interest: *stable controls* (194 individuals), *stable mild cognitive impairment*, also denoted as stable MCI (182 individuals), *stable Alzheimer patients* (162 individuals) and *converters MCI* (170 subjects). The individuals who reverted to control or MCI were not included in these groups. The cortical thickness measurements were computed using the FREESURFER software and averaged within the 34 regions of interest given by the Desikan-Killiany cortical parcellation [Desikan et al., 2006]. These measurements were analyzed using the univariate straight lines model (see V.1.2), for data in each parcel independently.

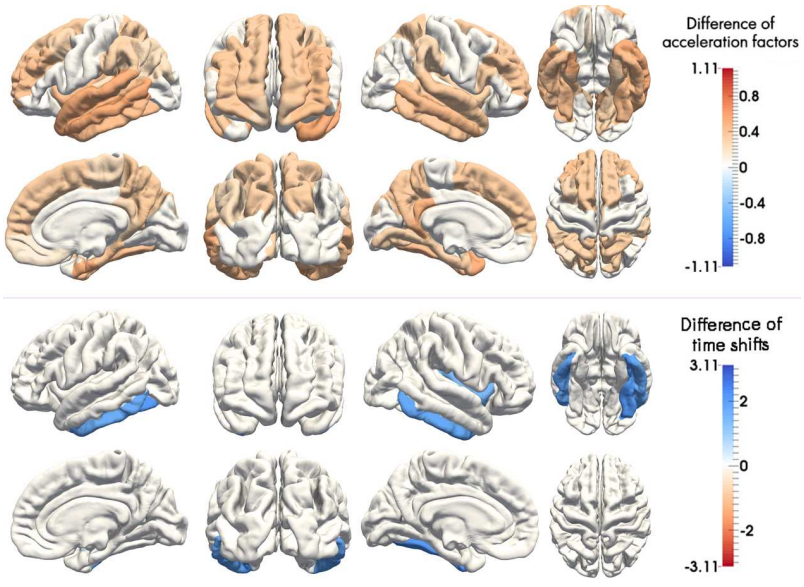
VII.4.2 Results

For each subject, MAP estimates of the time shifts and acceleration factors were obtained and the corresponding values were displayed on the cortical surface. The differences in the estimated time shifts and acceleration factors were compared between Alzheimer patients and controls in Figure 28 (a), and between converters MCI and stable MCI in Figure 28 (b). Significance level was set at 0.05, corrected for multiple comparisons using Bonferroni correction. Alzheimer patients present accelerated gray matter loss compared to stable controls in a large number of regions, with highest speed in temporal (including entorhinal cortex, parahippocampal gyrus, superior and middle temporal gyri), parietal associative (including precuneus) and frontal regions.

A similar topographical pattern was found for converters MCI compared to stable MCI but with smaller accelerations than in AD patients. On the contrary, primary motor and sensitive as well as visual cortices were spared. These results are consistent with the spatial-temporal progression patterns of neurodegeneration evidenced in histopathological studies [Braak and Braak, 1995], [Delacourte et al., 1999]. Furthermore, accelerated atrophy has also been recently shown to coincide with disease-onset [Benzinger et al., 2013]. On the other hand, the estimated time shifts were not significantly different for the vast majority of regions. In the few significant regions, the magnitude of the time-shifts was small. This is in contrast with the large time-



(a)



(b)

Figure 28 – **Figure (a)**: at the top (respectively bottom), the difference in averaged acceleration factors (respectively time shifts) between AD patients and stable controls is displayed on the cortex. **Figure (b)**: at the top (respectively bottom): the difference in averaged acceleration factors (respectively time shifts) between converters MCI and stable MCI is displayed on the cortex. In both figures, acceleration factors (and respectively time shifts) were averaged per regions of interest. Only regions where the difference was statistically significant ($p < 0.05$, corrected for multiple comparisons) were colored.

shift values found for the cognitive variables. This can be attributed to the slow but consistent age-related atrophy that is present in normal ageing subjects.

VII.5 Body fat measurements

VII.5.1 The dataset

We propose to analyze a longitudinal dataset from a prospective study on body fat accretion among a population of young girls. This dataset consists in observations of, initially, 162 young (from 8 to 18 years-old) girls from the MIT Growth and Development Study [Bandini et al., 2002,Phillips et al., 2003]. For each individual, at each visit, a measurement of body fatness is obtained through bioelectric impedance. Let $H_{i,j}$ denote the height (in cm), $W_{i,j}$ the weight (in kg) of the ith individual ($1 \leq i \leq 162$) at the jth visit. As mentioned in [Fitzmaurice et al., 2012], a measurement of total body water $\mathrm{TBW}_{i,j}$ is obtained with the following formula: $\mathrm{TBW}_{i,j} = (0.7H_{i,j}^2)/R - 0.32$, where R denotes the bioelectric impedance resistance. The percentage of body fat is obtained as follows: $y_{i,j} = \left(1 - (\mathrm{TBW}_{i,j}/0.73)W_{i,j}\right) \times 100$. The regression variable $t_{i,j}$ associated to each measurement $y_{i,j}$ is the age of the ith individual at the jth visit. In [Fitzmaurice et al., 2012], the authors mention that, at the beginning of the study, all the girls were pre-menarcheal and non-obese. They were all followed and had regular visits up to four years after the menarche. Indeed, body fat in girls starts to increase before the menarche and levels-off approximately four years after menarche. In this study, the time of menarche is known for each individual. The normalized observations of body fat were analyzed with the univariate logistic curves model (see Eq. [v.7]). Since this model assumes a monotonic progression of the measurements with time, we chose to remove from the initial dataset 11 individuals for whom the percentage of body fat was decreasing with time. For all the remaining individuals, the percentage of body fat tends to increase with time. After this step, a total of $p = 151$ individuals were remaining in the dataset.

Note that, by nature, the morphology of each individual is different. Natural differences in body shape induce a "size effect" in the measurements of body fat. In other words, the variability in the size of measurements is partly due to morphological differences between individuals. With our approach (the logistic curves model), this variability in the size of measurements would be interpreted as delay or advance in the progression of an individual. Therefore, the "size effects" in the measurements would translate as a temporal variability. In order no to over-estimate this temporal variability, it is preferable to reduce the size effects in the measurements. To address this problem, we propose to normalize the measurements. We assume that, for the ith individual, the observations $(y_{i,j})_{1 \leq j \leq k_i}$ belong to the open interval $]a_i, b_i[\subset]0, 1[$. The lower (respectively upper) bound a_i (respectively b_i) are unknown. The method proposed in this paragraph aims at estimating these bounds on the individual measurements.

Once a_i and b_i estimated, the observations of the ith individual are mapped to the Riemannian manifold $\mathbb{M} =]0,1[$ using the affine map $x \in]a_i, b_i[\mapsto (x - a_i)/(b_i - a_i) \in \mathbb{M}$. The bounds a_i and b_i are estimated by fitting a sigmoid curve, with unknown asymptotes, to the observations $(y_{i,j})_{1 \leq j \leq k_i}$. More precisely, for each individual, the following optimization problem is solved:

$$\forall i \in \{1, \dots, p\}, \ (a_i, b_i) = \underset{0<a_i<b_i<1}{\operatorname{argmin}} \sum_{j=1}^{k_i} \left[y_{i,j} - f_{a,b}(t_{i,j}) \right]^2 + \lambda (b-a)^2 \qquad \text{[vii.1]}$$

where $f_{a,b}$ is an increasing sigmoid function with asymptotes a and b. In the least squares criterion, the trade-off parameter λ is chosen equal to 10^{-1}. This optimization problem was solved using a gradient descent algorithm, implemented in MATLAB.

VII.5.2 Results

The fixed effects estimated with the MCMC-SAEM are $p_0 = 0.37$, $t_0 = 13.08$ and $v_0 = 0.13$ units per year. On average, the normalized score of an individual reaches the value $p_0 = 0.37$ at $t_0 = 13.08$ years old. The estimated standard deviation parameters $\sigma_\xi = 0.71$ and $\sigma_\tau = 2.00$ years inform on the variability of the average trajectory among the population. The average trajectory is shifted by ± 2 years with respect to the estimated time t_0 and the speed of normalized body fat increase varies, among the population, from 2.04 times slower to 2.04 times faster (Figure 29). After estimating the parameters of the model, the maximum a posteriori estimates of (η_i, τ_i) allow to consider the subject-specific time reparametrization ψ_i. This affine time reparametrization maps the individual timeline (to which $(t_{i,j})_{1 \leq j \leq k_i}$ belongs) to the reference timeline (the timeline of the average trajectory). Let t_i^{menarche} denote the age of the ith individual at menarche. Using these informations and the estimated individual trajectories, we propose to minimize the function $t \in \mathbb{R} \mapsto \sum_{1 \leq i \leq p} |t_i^{\text{menarche}} - \psi_i^{-1}(t)|$. This function has a unique minimum at $t^{\text{opt}} = 12.96$ years. The minimum t^{opt} can be understood as the age at menarche in the timeline of the normative scenario of body fat progression. We note that this age in the timeline of the average trajectory is close to the mean of $(t_i^{\text{menarche}})_{1 \leq i \leq p}$, which is 12.77 years. This average age at menarche can be mapped back to the individual timeline using the reparametrization ψ_i. For the ith individual, the age $\psi_i^{-1}(t^{\text{opt}})$ is an estimate of the actual age at menarche. An histogram of $\left(t_i^{\text{menarche}} - \psi_i^{-1}(t^{\text{opt}})\right)_{1 \leq i \leq p}$ is given in figure 30. The results given above show that the estimated age at menarche $\psi_i^{-1}(t^{\text{opt}})$ is a good approximation to the actual age at menarche t_i^{menarche} for 71% of the population. The error is greater than four years for only 7 individuals out of 151. For these individuals, the progression of body fat with time is more erratic, with periods during which measurements are decreasing. The logistic shape of the average trajectory is probably not the best choice of average trajectory for these individuals.

In [Fitzmaurice et al., 2012], the authors analyze the body fat measurements with a

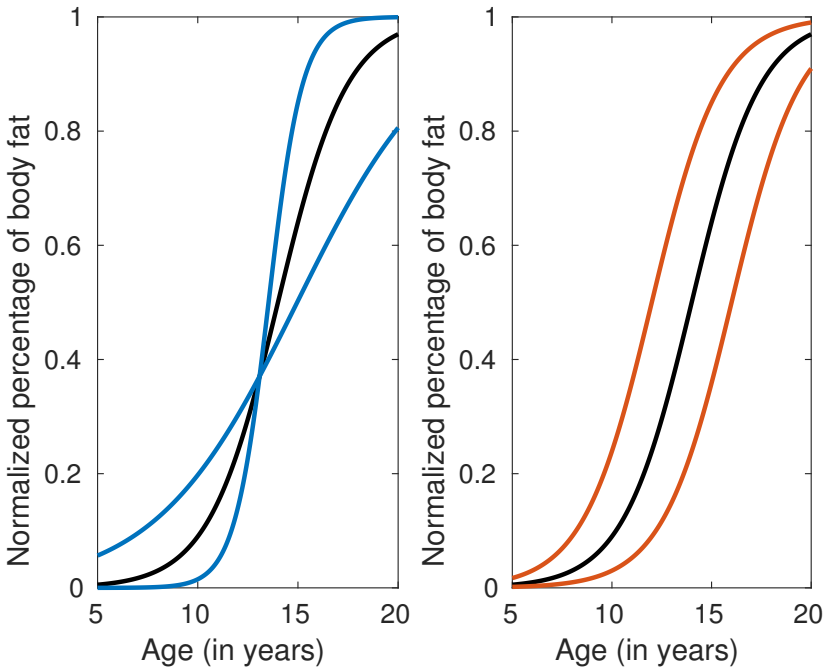


Figure 29 – Left: effect of the acceleration factor with plots of $\gamma_0(e^{\pm\sigma_\xi}(t-t_0)+t_0)$. Right: effect of the time shift with plots of $\gamma_0((t-t_0\mp\sigma_\tau)+t_0)$.

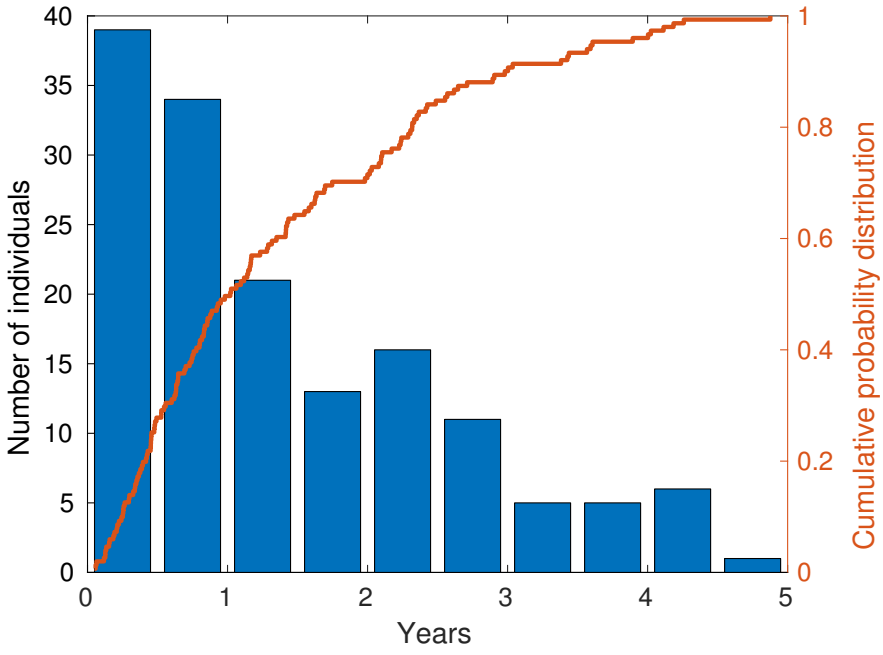


Figure 30 – Histogram of $\left(|t_i^{\text{menarche}}-\psi_i^{-1}(t^{\text{opt}})|\right)_{1\leq i\leq 151}$ superimposed with the cumulative probability distribution of this error.

piecewise linear mixed-effects model. The performance of their model and our logistic curves model can be assessed by computing the percentage of total variance explained

by both models. This percentage is defined by: $R^2 = 1 - \text{var}(\hat{\varepsilon}_{i,j})/\text{var}(\tilde{y}_{i,j})$, where $\text{var}(\hat{\varepsilon}_{i,j})$ denotes the variance of the residuals while $\text{var}(\tilde{y}_{i,j})$ denotes the variance of the (normalized) observations. The piecewise linear mixed-effects model proposed by Fitzmaurice et al. explains 84% of the variance (the coefficient R^2 is computed with the variance σ^2 of the noise $\varepsilon_{i,j}$ since the actual variance of the residuals is not given by the authors) whereas our logistic curves model explains 81% of the variance. Note that in [Fitzmaurice et al., 2012], the authors regress the percentages of body fat with respect to the time to menarche. In our approach, the age at menarche is not used during the estimation of the parameters of the model. As a matter of fact, it is an important piece of information our model tries to estimate. Even though the age to menarche was not included in our model, we note that its performance is similar to theirs (in terms of explained variance). We note also that our model requires the estimation of 6 parameters (fixed effects and variance-covariance parameters) as opposed to 10 parameters for the piecewise linear mixed-effects model. However, a drawback of our approach is the subject-specific normalization of the measurements. Indeed, as a different normalization is applied to each individual, the results cannot be used to make predictions on the evolution of body fat percentages. The piecewise linear mixed-effects model proposed in [Fitzmaurice et al., 2012] is well suited to the modeling of the progression of these measurements because it includes a change point at the age of menarche. However, it would be difficult to generalize or use this model to fit other types of measurements. Moreover, our model assumes that the time at menarche is unknown while this information is crucial in the definition of the model used by Fitzmaurice and colleagues.

VII.6 SPD matrices

VII.6.1 The dataset

We consider a synthetic dataset which consists in a repeated observations of a single diffusion tensor for one hundred individuals, simulating a progressive reduction in tensor asymmetry. The observations were not generated from the model contrary to the experiments described in Chapter VII. As a matter of fact, the observations were obtained by prescribing a hierarchical model on the eigenvalues of the diffusion tensors. At the level of the population, the eigenvalues of the diffusion tensors follow a decreasing piecewise linear evolution with a change point at 50 years old. Observations for a given individual were simulated by randomly shifting the change point (time at which the a change occurs in the speed at which eigenvalues decrease) and randomly increasing or decreasing the slopes of each eigenvalue (see Figure 31). As a result, we ensured that each individual follows a different trajectory in the space of eigenvalues and this trajectory is not given by a parallel variation of a geodesic. In particular, the initial values of the eigenvalues were different for each individual. The data was

generated for $p = 100$ individuals with, on average, 5 time points per individual.

VII.6.2 Results

The results presented below were obtained with $N_s = 1$ source. A greater number of independent sources would have been possible but many more iterations would have been necessary for the MCMC-SAEM to converge. The Bayesian tensor model with the MCMC-SAEM allowed to estimate an average trajectory of progression in the space SDP(3). This average trajectory is the geodesic which goes through the point $\overline{\mathbf{P}_0}$, at time $\overline{t_0}$, with velocity $\overline{\mathbf{V}_0}$ given by:

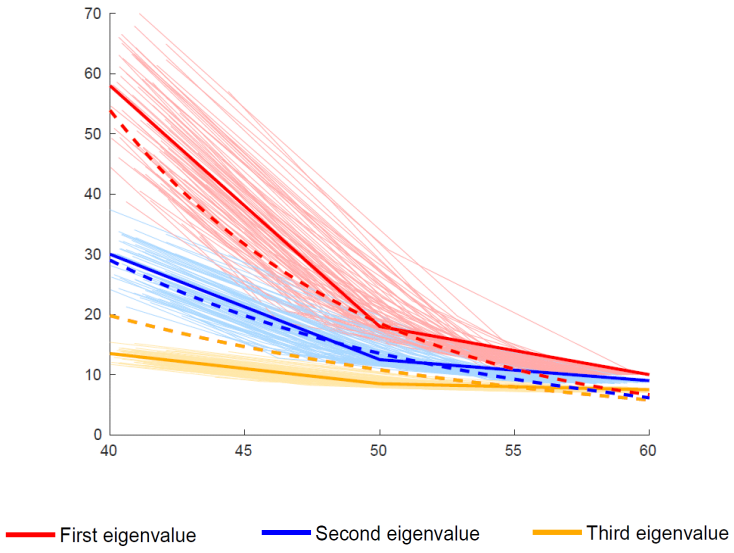
$$\overline{\mathbf{P}_0} = \begin{pmatrix} 11.30 & 0.96 & 0.68 \\ 0.96 & 9.53 & 1.21 \\ 0.68 & 1.21 & 10.19 \end{pmatrix}, \quad \overline{t_0} = 53.83,$$

and

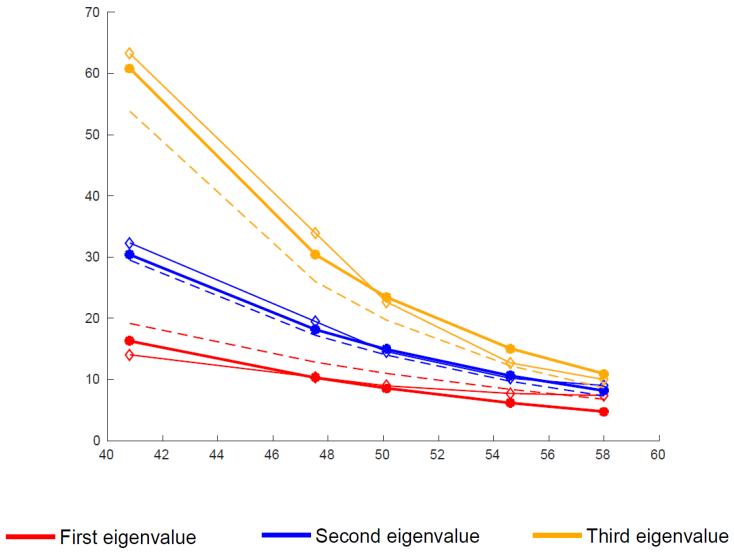
$$\overline{\mathbf{V}_0} = \begin{pmatrix} -0.99 & -0.17 & -0.20 \\ -0.17 & -0.75 & -0.27 \\ -0.20 & -0.27 & -0.85 \end{pmatrix}.$$

The evolution of the eigenvalues of the average trajectory, plotted in figure 31, is similar to the model used to generate the observations. However, the MCMC-SAEM tends to underestimate the first eigenvalue and overestimate the third eigenvalue. The variability in speed and delay of progression is captured by the estimated parameters $\sigma_\eta = 0.07$ and $\sigma_\tau = 0.5$. In figure 31 (top), we see that that, before and after the change point, eigenvalues of each individual decrease at a similar pace. This may explain why the model captured small variations in speed of progression. The standard deviation σ_τ on the parameter t_0 is much smaller. The individual acceleration factor, time shift and space shift allow to fit the average trajectory to the observations of an individual. As shown on figure 31 (bottom), the estimated individual trajectory is well adjusted to the observations of the individual.

The eigenvalues of the average estimated trajectory are smooth functions of time. Therefore, it would not have been possible to obtain piecewise-linear progression of the eigenvalues for the average trajectory. However, if t_i^* denotes the change point of each individual progression, we can validate the ability of the tensor model to put into correspondence the dynamic of each individual, using the validation method described in Section VII.2. For this dataset, the sum of errors $\sum_i |t_i^* - \psi_i^{-1}(t)|$ has a unique minimum at $t^{\text{opt}} = 49.73$ years. This minimum t^{opt} is close to 50 years, the time at which the change point occurs in the average model used to generate the data. Figure 32 shows that individual progressions are not perfectly put into correspondence by the time shift. However, for the ith individual, we make an error less than 2 years for almost 60% of the population by estimating the individual change point with $\psi_i^{-1}(t^{\text{opt}})$. Moreover, the error is less than 4 years for 90% of the population. During the simulation of this dataset, the change point of the average eigenvalues trajectory, at



(a)



(b)

Figure 31 – **Figure (a)**: In solid bold line, the average model of eigenvalues evolution for the synthetic dataset of tensors. In solid lines, the evolution of the eigenvalues for all the individuals in the dataset. In dotted line, the evolution of the eigenvalues of the average trajectory, given by the MCMC-SAEM. **Figure (b)**: The evolution of the eigenvalues of an individual. In dotted line, the eigenvalues of the average trajectory estimated by the MCMC-SAEM. With square markers, the eigenvalues of the observations for this individual. With round markers, the eigenvalues of the estimated individual trajectory.

50 years, was randomly shifted using draws from a centered Gaussian distribution with a standard deviation of 2 years. Therefore, the error $(|t_i^* - \psi_i^{-1}(t^{\text{opt}})|)_i$ is similar to the standard deviation of the distribution of the change point. However, we note that the eigenvalues of a geodesic, for the affine-invariant metric used here (see Eq. [iv.31]), or a parallel variation of a geodesic, tend to evolve as smooth functions, with an exponential shape. Therefore, it seems impossible for this model to recover precisely a piecewise linear evolution of the eigenvalues.

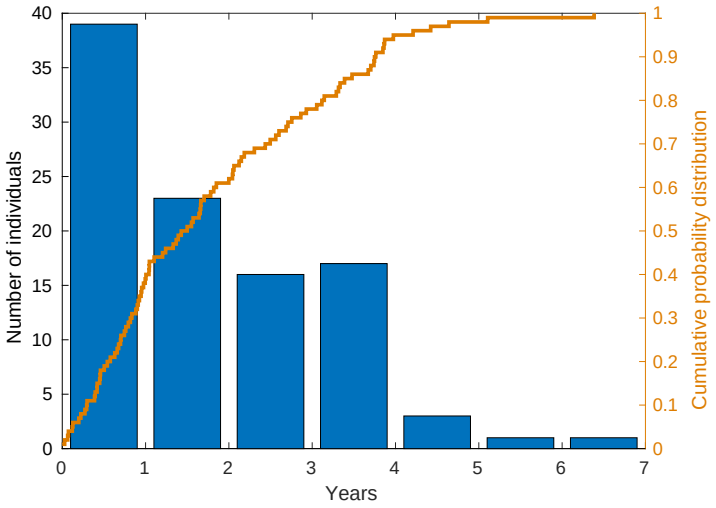


Figure 32 – Histogram of $\left(|t_i^* - \psi_i^{-1}(t^{\text{opt}})|\right)_{1 \leq i \leq 100}$ superimposed with the cumulative distribution of this error. Here, t_i^* represent the age of the change point for the ith individual and $t^{\text{opt}} = 49.73$ years.

Part VIII

Conclusion and perspectives

Summary

VIII.1 Conclusive summary

In this dissertation, we proposed a Bayesian mixed-effects model, called *generic spatiotemporal model*, for the spatiotemporal analysis of longitudinal manifold-valued measurements. The framework of Riemannian manifolds enables to consider almost any kind of longitudinal observations lying in a space defined by smooth constraints. The generic model inherits a hierarchical structure from its fixed and random effects, which allow to describe the model both at the group and individuals level. At the population level, the fixed effects of the model define a group-average trajectory of progression, which is a geodesic on a Riemannian manifold. In order to define individual trajectories of progression, we introduced the concept of "parallel variation" of a curve on a Riemannian manifold. For each individual, a "parallel" to the average trajectory is defined using an individual-specific space shift. This parallel is then reparametrized in time using an affine individual-specific time reparametrization defined by its acceleration factor and time shift. The concept of parallel variation enforces an orthogonality constraint on the space shifts and we proposed methods and algorithms to include this orthogonality constraint into the generic model. Eventually, the random effects of the model, namely acceleration factors, time shifts and space shifts, allow to put into correspondence trajectories of progression across individuals. In other words, the space shifts allow to spatially register the trajectories whereas the acceleration factor and time shifts allow to temporally register the timeline of each individual to the reference timeline. The distribution of these random effects allows to learn a distribution of trajectories of changes on a Riemannian manifold.

Chapter V shows how a large variety of mixed-effects models for longitudinal data may be derived from the generic abstract model. We proposed particular cases of the generic model which can be used to analyze specific types of longitudinal observations. The *logistic curves model* (respectively *straight lines model*) is designed for longitudinal normalized or bounded (respectively unbounded) scalar observations. The *progression models* may be used to model the temporal progression of a family of features or biological characteristics. These models assumes the same shape of progression for each feature, but shifted in time. As a result, it models the joint temporal progression of these features, but also to estimate the relative delay between those. These models can be used to determine an ordering of the observed features. Finally, we proposed a model called *SPD matrices model*, which is designed for the analysis of longitudinal datasets of positive symmetric definite matrices (like covariance matrices).

In Chapter VI, we reviewed and discussed several methods for the inference in nonlinear mixed-effects models. We chose to use a stochastic version of the EM algorithm, namely the Monte Carlo Markov Chain Stochastic Approximation EM (MCMC-SAEM) algorithm, to estimate the parameters of the generic spatiotemporal model. This choice is motivated by the fact that the MCMC-SAEM offers theoretical guarantees of convergence and, with specific families of samplers, does not require any computation of derivative. Since the model is defined in a Riemannian manifold framework, we

discussed possible solutions to overcome difficulties encountered when deriving the MCMC-SAEM for the generic spatiotemporal model. In this chapter, we also evaluated and validated the proposed algorithm in settings of varying complexity. In particular, the MCMC-SAEM is validated on a synthetic longitudinal dataset of 3×3 covariance matrices, which provides an example of observations on a multivariate curved Riemannian manifold. We also discussed the use of numerical schemes within the MCMC-SAEM. Eventually, we presented in chapter VII data-driven models of progression estimated from longitudinal health data. Longitudinal datasets of normalized neuropsychological test scores are analyzed with the unstructured logistic progression model to derive a normative scenario of the progressive impairment of cognitive functions during the onset of Alzheimer's disease. We also analyzed longitudinal observation of cortical thickness for individuals from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort and obtained results which are consistent with previous knowledge on the pathophysiology of Alzheimer's disease. The experimental results obtained with the longitudinal dataset of body fat measurements and with the simulated longitudinal dataset of SPD matrices show that the generic spatiotemporal model successfully puts in correspondence the progression of individuals.

VIII.2 Limitations

The proposed methodology and algorithm are subject to several assumptions. Some of these assumptions are discussed in the previous chapters of this dissertation. In this section, we review and discuss some limitations of the generic approach.

VIII.2.1 The monotonicity assumption

The straight lines model (Eq. [v.5]), logistic curves model (Eq. [v.7]) or progression models (see Section V.3) assume a monotonic evolution of the measurements. The monotonicity of the progression makes sense, for example, for the analysis of neuropsychological test scores in the context of Alzheimer's disease. However, this assumption may not make sense for other neurodegenerative diseases such as multiple sclerosis (MS) for which patients may experience remission periods. This assumption may also not make sense for other types of longitudinal observations such as behavioral, mental health test questions, where the progression for healthy individuals might not be monotonic.

VIII.2.2 Assumptions on the Riemannian manifold $\mathbb{M}$

Throughout this dissertation, the assumption that the Riemannian manifold $\mathbb{M}$ is a (connected) *open* subset of the Euclidean space $\mathbb{R}^N$ played an important role. This

assumption is particularly useful for the MCMC-SAEM since it legitimates the use of symmetric random walk Metropolis-Hastings algorithms within the Gibbs sampler. However, this assumption clearly does not hold for simple Riemannian manifolds such as the 2-sphere $\mathbb{S}^2$ or the torus $\mathbb{T}^2 \subset \mathbb{R}^3$ defined by: $\mathbb{T}^2 = \{(x,y,z) \in \mathbb{R}^3,\, z^2 + \big(2 - \sqrt{x^2+y^2}\big)^2 = 1\}$. Indeed, the Gibbs sampler with symmetric random walk proposal would draw samples which are not necessarily on the manifold. Possible solutions might consist in considering the projection of a multivariate Gaussian distribution on the Riemannian manifold, as long as this leads to a tractable density function. It may also be possible to consider specific probability distributions such as the Von Misses distribution for the 2-sphere $\mathbb{S}^2$ or the Bivariate Von Misses distribution for the torus $\mathbb{T}^2$, but at the expense of the generalization to other Riemannian manifolds.

VIII.3 Perspectives

This section discusses several possible improvements to the generic spatiotemporal model.

VIII.3.1 Multi-class approach

A possible development would consist in using the model for unsupervised clustering. This could be done by replacing the probability distribution in Section IV.3.4 by mixtures of probability distributions. As a result, the generic spatiotemporal model could be used to automatically create subsets of individuals among the population sharing similar spatiotemporal patterns. Another possible approach to unsupervised clustering would be to fit the generic spatiotemporal model to a longitudinal dataset and estimate the individual random effects as it was done in Sections VII.3. Then, algorithms such as K-Means or mixture of Gaussians could be used to classify these individual random effects into subgroups. However, such classification method could be directly included into the model as mixtures of probability distributions, following the work in [Marin et al., 2005]. For instance, consider the longitudinal dataset analyzed in Section VII.3.2. Recall that this longitudinal dataset consists in univariate normalized neuropsychological test scores for 1393 individuals from the ADNI database. Let $\boldsymbol{\theta}^*$ denote the parameters estimated by the MCMC-SAEM for this dataset. The individual log-acceleration factors $(\xi_i)_{1\leq i\leq p}$ are estimated with two different methods. The first method was used to produce results presented in Chapter VII: (ξ_i, τ_i) are estimated for each individual by maximizing the joint conditional distribution $q(\xi_i, \tau_i \mid \mathbf{y}_i, \boldsymbol{\theta}^*)$. With the second method, (ξ_i, τ_i) were estimated by minimizing the nonlinear least squares $\sum_{j=1}^{k_i} |y_{i,j} - \gamma_0\big(\psi_i(t_{i,j})\big)|^2$. In the second method, no constraint is enforced on the distribution of the individual random effects. Equivalently, the second method is similar to the first, without prior distribution on these individual random effects. Figure 33 gives a normalized histogram of the log-acceleration factors obtained with both meth-

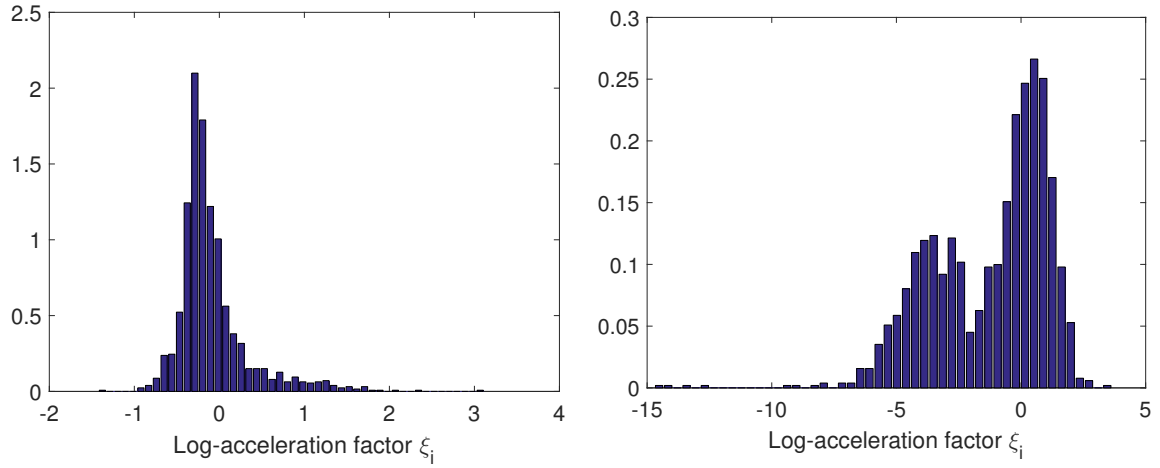


Figure 33 – **Left**: Normalized histogram of the individual log-acceleration factors ξ_i estimated by maximizing the joint posterior distribution $q(\xi, \tau_i \mid \mathbf{y}_i, \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ are the model parameters estimated by the MCMC-SAEM in Section VII.3.2. **Right**: Normalized histogram of the individual log-acceleration factors ξ_i estimated by solving a nonlinear least squares problem. No prior distribution was assumed for the log-acceleration factors.

ods. These results give the idea that a mixture of distributions could better explain the data and allow to identify a subpopulation of slow progressers (the left peak in the second histogram) and another subpopulation of fast progressers (the right peak in the second histogram).

VIII.3.2 Using the generic model with multi-modal data

Other possible developments include the extension of the generic spatiotemporal model to multi-modal longitudinal observations. Indeed, at each time point and for each individual, the observations could be the concatenation of images (like Magnetic Resonance (MR) images, for example) and observations of biological features such as the concentration of a protein in the cerebrospinal fluid or neuropsychological test scores. A possible application of these methodological developments would be to map the onset of clinical symptoms on the anatomical and functional changes of the brain seen in MR images. Still, this would require to extend the generic spatiotemporal model to longitudinal shape analysis. This could be done by specifying the generic model on the group of diffeomorphisms. However, this might raise several methodological difficulties. In particular, the parallel transport is not known explicitly for the group of diffeomorphisms. Therefore, the MCMC-SAEM would have to be used with a numerical scheme to approximate the parallel transport. Moreover, considering multi-modal data will lead to work in a high-dimensional Riemannian [product] manifold. This high-dimensional setting might lead to consider more efficient MCMC samplers for the

MCMC-SAEM.

VIII.3.3 Towards MCMC methods for high-dimensional settings

If the generic spatiotemporal model is used to analyze multi-modal data, the dimension N of the space of observations $\mathbb{M}$ might increase greatly. As a result, using the generic model in a high-dimensional setting might raise several methodological and computational challenges. The block Metropolis-Hastings-within-Gibbs sampler proposed in Section VI.3.2 might have difficulties in sampling efficiently the target distribution if the dimension of the space of latent variables becomes very large. To address this problem, we could consider other MCMC samplers such as the Metropolis Adjusted Langevin Algorithm [Atchadé, 2006, Beskos, 2014]. Other possible improvements would consist in using or adapting MCMC methods, such as [Maclaurin and Adams, 2014, Shang et al., 2015], which only use subsets of the dataset.

VIII.3.4 Personalization and prediction

Finally, the generic spatiotemporal model could be used for prediction purposes. Given some individual observations, we could determine how well the generic spatiotemporal model allows to predict the observation at the next time point. In addition to this, the model could allow to predict the evolution of biological features and estimate the time to the onset of symptoms and provide insightful informations to help diagnose specific diseases.

Appendices

A A review of mixed-effects models and their limitations in the context of manifold-valued data

Remarkable developments have been made in the past 30 years regarding the methodology for analyzing longitudinal data. Advancements in computer softwares and technologies have facilitated these methodological developments, their implementation and use in a large spectrum of disciplines. This section presents a powerful and flexible family of statistical models for longitudinal data analysis called *mixed-effects models* or *mixed models*. Initially, these models were introduced to handle *clustered data* (data in which the measurements can be grouped into several "classes"). Longitudinal data can be considered as a particular case of clustered data since multiple observations are made for the same individual. Mixed-effects models have become popular for several reasons. Among these, they offer the advantage of handling missing data and unbalanced repeated measures with uneven spacing of the measurements in time. Moreover, they model the observations as a function of *fixed effects* and *random effects*. The terms "fixed effects" and "random effects" were formally introduced in 1947 by Eisenhart in [Eisenhart, 1947] and lead to the terminology *mixed models*. Fixed and random effect provide the model with a hierarchical structure, allowing to describe the model at different *levels*. In numerous applications, two-levels models are considered, where the fixed (respectively random) effects describe the model at the *population* (respectively *individuals*) level. However, models with more than two levels appear in the literature. In particular, a third level is sometimes used to model clusters of individuals.

Two types of mixed-effects models have been extensively studied in the literature: *linear mixed-effects models* and *nonlinear mixed-effects models*. A comprehensive overview of mixed-effects models for longitudinal data can be found in [Diggle et al., 2002, Fitzmaurice et al., 2008, Verbeke and Molenberghs, 2009, Fitzmaurice et al., 2012].

A.1 Linear mixed-effects (LME) models

An LME model is the *random slope and intercept model*, which assumes that a longitudinal dataset $(\mathbf{y}_{i,j}, t_{i,j})_{1\leq i\leq p,\, 1\leq j\leq k_i}$ arise from:

$$\mathbf{y}_{i,j} = t_{i,j}(\overline{\mathbf{A}} + \mathbf{A}_i) + (\overline{\mathbf{B}} + \mathbf{B}_i) + \boldsymbol{\varepsilon}_{i,j} \qquad \text{[viii.1]}$$

where $(t_{i,j})_{1\leq j\leq k_i}$ denotes the time points at which the observations of the ith individual were obtained. This model assumes that, for all $i \in \{1, \ldots, p\}$, $\mathbf{A}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}_A)$, $\mathbf{B}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}_B)$ and $\boldsymbol{\varepsilon}_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \mathrm{I}_{k_i})$ and are independent of each other. Note that the random slope and intercept model is a particular case of the generic LME model

(see Eq. [ii.1]) since it may be written with:

$$\mathbf{X}_i = \mathbf{Z}_i = \begin{bmatrix} t_{i,1} & 1 \\ \vdots & \vdots \\ t_{i,k_i} & 1 \end{bmatrix}, \boldsymbol{\alpha} = \begin{bmatrix} \overline{\mathbf{A}} \\ \overline{\mathbf{B}} \end{bmatrix}, \text{ and } \boldsymbol{\beta}_i = \begin{bmatrix} \mathbf{A}_i \\ \mathbf{B}_i \end{bmatrix}. \quad \text{[viii.2]}$$

For $i \in \{1, \ldots, p\}$, let $\mathbf{D}_i(t) = (\overline{\mathbf{A}} + \mathbf{A}_i)t + (\overline{\mathbf{B}} + \mathbf{B}_i)$. Then, Eq. [viii.1] may be written: $\mathbf{y}_{i,j} = \mathbf{D}_i(t_{i,j}) + \boldsymbol{\varepsilon}_{i,j}$. This last equation shows that, under this LME model, the observations of the ith individual can be seen as random samples along an *individual trajectory* $\mathbf{D}_i$, which is obtained as follows: starting from the *average trajectory* $\overline{\mathbf{D}}(t) = t\overline{\mathbf{A}} + \overline{\mathbf{B}}$, the random effects $\mathbf{A}_i$ and $\mathbf{B}_i$ are used to transform, *via* a change of slope and intercept, the straight line $\overline{\mathbf{D}}$ into $\mathbf{D}_i$. The random effect $\mathbf{A}_i$ informs on whether the ith individual is progressing faster or slower than the average trajectory $\overline{\mathbf{D}}$, whereas the random effect $\mathbf{B}_i$ informs on the distribution of the measurements at time $t = 0$.

B Methods for the inference in linear mixed-effects models

In this section, we consider the LME model (see Eq. [ii.1]) where, for all $i \in \{1, \ldots, p\}$, $\mathbf{R}_i = \sigma^2 \mathrm{I}_{k_i}$. Moreover, we assume that the variance-covariance matrix $\mathbf{D} \in \mathrm{Spd}(q)$ is a function of some unknown parameters $\tilde{\boldsymbol{\theta}}$. Let $\boldsymbol{\theta} = (\tilde{\boldsymbol{\theta}}, \sigma^2)$ denote the variance-covariance parameters of the model. Hence, the parameters of the model are: $(\boldsymbol{\alpha}, \boldsymbol{\theta})$.

As mentioned above, estimating the parameters of this model is done by using maximum likelihood methods or restricted maximum likelihood methods. *Restricted likelihood methods* were introduced in [Thompson Jr, 1962] to reduce the small-sample biais of maximum likelihood methods. Both methods proceed as follows: first, the maximum likelihood estimator of $\boldsymbol{\alpha}$, denoted by $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$, is obtained. This estimator, which is a function of the unknown variance-covariance parameters $\boldsymbol{\theta}$ is plugged into the likelihood to obtain a *profiled likelihood*. This profiled likelihood is then maximized a Newton-Raphson algorithm or an EM algorithm ([Dempster et al., 1977] ; see [Lindstrom and Bates, 1988, Laird et al., 1987] for the use of the EM algorithm in LME models).

Conditionally on $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$, the vector $\mathbf{y}$ is normally distributed with mean $\mathbf{X}\boldsymbol{\alpha}$ and variance-covariance matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \mathbf{Z}\mathbf{D}\mathbf{Z}^\top + \sigma^2 \mathrm{I}_n$. The log-likelihood $\log q(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta})$ of $\mathbf{y}$ is given by:

$$\log q(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) = -\frac{1}{2} \log \det \boldsymbol{\Gamma}(\boldsymbol{\theta}) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta})(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}). \quad \text{[viii.3]}$$

If the variance-covariance parameters $\boldsymbol{\theta}$ are known, the log-likelihood $q(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta})$ can be maximized with respect to $\boldsymbol{\alpha}$, yielding an estimate $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ given by:

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \left(\mathbf{X}^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta})\mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta})\mathbf{y}. \quad \text{[viii.4]}$$

The estimator $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ is the *best linear unbiased estimator* (BLUE) of $\boldsymbol{\alpha}$ since it depends linearly on $\mathbf{y}$ and it has minimum mean square error among the class of linear unbiased estimator of $\boldsymbol{\alpha}$. Since the conditional distribution of $\mathbf{y}$ is normal, given $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$, the estimator $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ can be derived from the Gauss-Markov theorem (see [Harville, 1976]). Plugging the estimator $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ into the log-likelihood $\log q(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta})$ leads to the *profiled log-likelihood* $\log q(\mathbf{y} \mid \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \boldsymbol{\theta})$. In [Harville, 1977] and [Lindstrom and Bates, 1988], the authors use the profiled log-likelihood to estimate the variance-covariance parameters of a LME model. Indeed, by definition of the BLUE of $\boldsymbol{\alpha}$:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\theta}} \log q(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \log q(\mathbf{y} \mid \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \boldsymbol{\theta}). \qquad \text{[viii.5]}$$

However, maximizing the profiled log-likelihood with respect to $\boldsymbol{\theta}$ is a difficult problem which usually does not yield closed-form solutions. In order to ensure the positive definiteness of the matrix $\mathbf{D}$, Pinheiro and Bates in [Pinheiro and Bates, 1996] propose five different possible parametrizations of variance-covariance matrices to ensure positive definiteness and compare them in terms of computational efficiency and statistical interpretability.

B.1 Restricted likelihood

The drawback of maximizing the profiled log-likelihood to produce estimates of the variance-covariance parameters is that the maximum likelihood estimates (MLE) are biased as they do not take into account the loss of degree of freedom from the estimation of the fixed effects with $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$. To address this problem, the likelihood is replaced with the *restricted likelihood*. By definition, the restricted likelihood method consists not in the likelihood of the full data $\mathbf{y}$ but the likelihood of $n - s$ *error contrasts* $\mathbf{Ay}$, where s is the rank of the matrix $\mathbf{X}$. The matrix $\mathbf{A}$ is given by: $\mathbf{A} = \mathrm{I}_n - \mathbf{X}\big(\mathbf{X}^\top\mathbf{X}\big)^{-1}\mathbf{X}^\top$, the orthogonal projection onto $\big(\mathrm{Im}\mathbf{X}\big)^\perp$. In [Harville, 1974], the authors show that the likelihood of $\mathbf{Ay}$ is given by:

$$q(\mathbf{Ay} \mid \boldsymbol{\theta}) = q(\mathbf{y} \mid \hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \boldsymbol{\theta}) - \frac{1}{2}\log\det\big(\mathbf{X}^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta})\mathbf{X}\big). \qquad \text{[viii.6]}$$

Let $\tilde{\mathbf{D}}$ be the *scaled covariance matrix* such that $\mathbf{D} = \sigma^2\tilde{\mathbf{D}}$. The matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ can be written $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \sigma^2\big(\mathbf{Z}\tilde{\mathbf{D}}\mathbf{Z}^\top + \mathrm{I}_n\big)$. It follows that the restricted likelihood Eq. [viii.6] can be used to provide an *unbiased* estimate of σ^2:

$$\hat{\sigma^2}(\tilde{\boldsymbol{\theta}}) = \frac{1}{n-s}\big(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}(\tilde{\boldsymbol{\theta}})\big)^\top\boldsymbol{\Gamma}^{-1}(\tilde{\boldsymbol{\theta}})\big(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}(\tilde{\boldsymbol{\theta}})\big). \qquad \text{[viii.7]}$$

Substituting this estimator in Eq. [viii.6] yields a profiled restricted likelihood which only depends on the variance-covariance parameters $\tilde{\boldsymbol{\theta}}$. Maximizing this profiled restricted likelihood with respect to $\tilde{\boldsymbol{\theta}}$ is a nonlinear optimization problem. In [Lindstrom and Bates, 1988], the authors suggest to use a Newton-Raphson algorithm to maximize

this function. In addition to this, they note that removing σ^2 from the parameters to estimate reduces the number of iterations required and improves the overall convergence behavior. They also note that the Newton-Raphson algorithm fails to converge when applied directly to the log-likelihood of $\mathbf{y}$, but is able to maximize the profiled (restricted) log-likelihood. However, the authors do not mention whether the objective function is convex. Indeed, if the objective function is not convex (and not unimodal), the Newton-Raphson algorithm could fail to converge.

B.2 Estimation of the random effects

When the variance-covariance parameters $\boldsymbol{\theta}$ are known, [Laird and Ware, 1982] provide an estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ of $\boldsymbol{\beta}$, which is given by:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \mathbf{D}\mathbf{Z}^\top \boldsymbol{\Gamma}^{-1}(\boldsymbol{\theta})\big(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})\big). \qquad \text{[viii.8]}$$

This estimator is called *best linear unbiased predictor* (BLUP) of the random effect $\boldsymbol{\beta}$. The term *predictor* is used to distinguish this estimator from estimator of the fixed effects. In [Robinson, 1991], different methods to obtain $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ are reviewed. Note that Eq. [viii.8] and Eq. [viii.4] both require to inverse the $K \times K$ matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta})$, where $K = \sum_i k_i$ denotes the number of observations for all the individuals. In [Henderson, 1950], a method to jointly obtain $\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is proposed. This methods consist in solving a linear system of equation called *mixed-model equations* (MME). This linear system is defined by:

$$\begin{bmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{X} & \mathbf{Z}^\top\mathbf{Z} + \sigma^2\mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top\mathbf{y} \\ \mathbf{Z}^\top\mathbf{y} \end{bmatrix}. \qquad \text{[viii.9]}$$

The matrix of this linear system is a $(2p+1)\times(2p+1)$ matrix, where p corresponds to the number of individuals. Usually, p is much smaller than K. As a consequence, solving the MME is computationally interesting. The MME can be obtained by maximizing the joint density function of $(\mathbf{y}, \boldsymbol{\beta})$ with respect to $(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Bibliography

References

[Allassonniere and Kuhn, 2015] Allassonniere, S. and Kuhn, E. (2015). Convergent stochastic expectation maximization algorithm with efficient sampling in high dimension. application to deformable template model estimation. Computational Statistics & Data Analysis, 91:4–19.

[Allassonnière et al., 2010] Allassonnière, S., Kuhn, E., and Trouvé, A. (2010). Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. Bernoulli, 16(3):641–678.

[Andrieu et al., 2005] Andrieu, C., Moulines, E., and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. SIAM Journal on control and optimization, 44(1):283–312.

[Arsigny et al., 2006] Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2006). Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magnetic resonance in medicine, 56(2):411–421.

[Atchadé, 2006] Atchadé, Y. F. (2006). An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. Methodology and Computing in applied Probability, 8(2):235–254.

[Bandini et al., 2002] Bandini, L., Must, A., Spadano, J., and Dietz, W. (2002). Relation of body composition, parental overweight, pubertal stage, and race-ethnicity to energy expenditure among premenarcheal girls. The American Journal Of Clinical Nutrition, 76(5):1040–1047.

[Benzinger et al., 2013] Benzinger, T. L., Blazey, T., Jack, C. R., Koeppe, R. A., Su, Y., Xiong, C., Raichle, M. E., Snyder, A. Z., Ances, B. M., Bateman, R. J., et al. (2013). Regional variability of imaging biomarkers in autosomal dominant alzheimer's disease. Proceedings of the National Academy of Sciences, 110(47):E4502–E4509.

[Beskos, 2014] Beskos, A. (2014). A stable manifold mcmc method for high dimensions. Statistics & Probability Letters, 90:46–52.

[Betancourt, 2013] Betancourt, M. (2013). A general metric for riemannian manifold hamiltonian monte carlo. In Geometric science of information, pages 327–334. Springer.

[Billingsley, 2013] Billingsley, P. (2013). Convergence of probability measures. John Wiley & Sons.

[Booth and Hobert, 1999] Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(1):265–285.

[Braak and Braak, 1995] Braak, H. and Braak, E. (1995). Staging of alzheimer's disease-related neurofibrillary changes. Neurobiology of aging, 16(3):271–278.

[Cappé et al., 2012] Cappé, O., Guillin, A., Marin, J.-M., and Robert, C. P. (2012). Population monte carlo. Journal of Computational and Graphical Statistics.

[Chi and Reinsel, 1989] Chi, E. M. and Reinsel, G. C. (1989). Models for longitudinal data with random effects and ar (1) errors. Journal of the American Statistical Association, 84(406):452–459.

[Consonni and Marin, 2007] Consonni, G. and Marin, J.-M. (2007). Mean-field variational approximate bayesian inference for latent variable models. Computational Statistics & Data Analysis, 52(2):790–798.

[Corder et al., 1993] Corder, E., Saunders, A., Strittmatter, W., Schmechel, D., Gaskell, P., Small, G., Roses, A., Haines, J., and Pericak-Vance, M. (1993). Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. Science, 261(5123):921–923.

[Cowles and Carlin, 1996] Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association, 91(434):883–904.

[Davidian and Gallant, 1992] Davidian, M. and Gallant, A. R. (1992). Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. Journal of Pharmacokinetics and Biopharmaceutics, 20(5):529–556.

[Delacourte et al., 1999] Delacourte, A., David, J., Sergeant, N., Buee, L., Wattez, A., Vermersch, P., Ghozali, F., Fallet-Bianco, C., Pasquier, F., Lebert, F., et al. (1999). The biochemical pathway of neurofibrillary degeneration in aging and alzheimer's disease. Neurology, 52(6):1158–1158.

[Delor et al., 2013] Delor, I., Charoin, J., Gieschke, R., Retout, S., and Jacqmin, P. (2013). 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):e78.

[Delyon et al., 1999] Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. Annals of statistics, pages 94–128.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological), pages 1–38.

[Desikan et al., 2006] Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage, 31(3):968–980.

[Diggle et al., 2002] Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). Analysis of longitudinal data. Oxford University Press.

[Do Carmo Valero, 1992] Do Carmo Valero, M. P. (1992). Riemannian geometry. Birkhäuser.

[Dryden et al., 2009] Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. The Annals of Applied Statistics, pages 1102–1123.

[Durrleman et al., 2011] Durrleman, S., Fillard, P., Pennec, X., Trouvé, A., and Ayache, N. (2011). Registration, atlas estimation and variability analysis of white matter fiber bundles modeled as currents. NeuroImage, 55(3):1073–1090.

[Durrleman et al., 2013] Durrleman, S., Pennec, X., Trouvé, A., Braga, J., Gerig, G., and Ayache, N. (2013). Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. International Journal of Computer Vision, 103(1):22–59.

[Eisenhart, 1947] Eisenhart, C. (1947). The assumptions underlying the analysis of variance. Biometrics, 3(1):1–21.

[Fitzmaurice et al., 2008] Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). Longitudinal data analysis. CRC Press.

[Fitzmaurice et al., 2012] Fitzmaurice, G., Laird, N., and Ware, J. (2012). Applied longitudinal analysis, volume 998. John Wiley & Sons.

[Fletcher et al., 2009] Fletcher, P. T., Venkatasubramanian, S., and Joshi, S. (2009). The geometric median on riemannian manifolds with application to robust atlas estimation. NeuroImage, 45(1):S143–S152.

[Fletcher, 2011] Fletcher, T. (2011). 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.

[Förstner and Moonen, 2003] Förstner, W. and Moonen, B. (2003). A metric for covariance matrices. In Geodesy-The Challenge of the 3rd Millennium, pages 299–309. Springer.

[Gallot et al., 1990] Gallot, S., Hulin, D., and Lafontaine, J. (1990). Riemannian geometry, volume 3. Springer.

[Gelman and Rubin, 1992] Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical science, pages 457–472.

[Geweke, 2004] Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. Journal of the American Statistical Association, 99(467):799–804.

[Girolami and Calderhead, 2011] Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(2):123–214.

[Harville, 1976] Harville, D. (1976). Extension of the gauss-markov theorem to include the estimation of random effects. The Annals of Statistics, pages 384–395.

[Harville, 1974] Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. Biometrika, 61(2):383–385.

[Harville, 1977] Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, 72(358):320–338.

[Henderson, 1950] Henderson, C. R. (1950). Estimation of genetic parameters. In Biometrics, volume 6, pages 186–187.

[Hoffman and Gelman, 2014] Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. Journal of Machine Learning Research, 15(1):1593–1623.

[Hong et al., 2014] Hong, Y., Singh, N., Kwitt, R., and Niethammer, M. (2014). Time-warped geodesic regression. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 105–112. Springer.

[Hyvärinen et al., 2004] Hyvärinen, A., Karhunen, J., and Oja, E. (2004). Independent component analysis, volume 46. John Wiley & Sons.

[Jordan et al., 1999] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. Machine learning, 37(2):183–233.

[Kheyfets et al., 2000] Kheyfets, A., Miller, W. A., and Newton, G. A. (2000). Schild's ladder parallel transport procedure for an arbitrary connection. International Journal of Theoretical Physics, 39(12):2891–2898.

[Kuhn and Lavielle, 2004] Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure. ESAIM: Probability and Statistics, 8:115–131.

[Laird et al., 1987] Laird, N., Lange, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the em algorithm. Journal of the American Statistical Association, 82(397):97–105.

[Laird and Ware, 1982] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics, pages 963–974.

[Lang, 1972] Lang, S. (1972). Differential manifolds, volume 212. Springer.

[Lavielle and Mentré, 2007] Lavielle, M. and Mentré, F. (2007). Estimation of population pharmacokinetic parameters of saquinavir in hiv patients with the monolix software. Journal of pharmacokinetics and pharmacodynamics, 34(2):229–249.

[Lee, 2003] Lee, J. M. (2003). Smooth manifolds. In Introduction to Smooth Manifolds, pages 1–29. Springer.

[Lee, 2006] Lee, J. M. (2006). Riemannian manifolds: an introduction to curvature, volume 176. Springer Science & Business Media.

[Lenglet et al., 2006] Lenglet, C., Rousson, M., Deriche, R., and Faugeras, O. (2006). Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor mri processing. Journal of Mathematical Imaging and Vision, 25(3):423–444.

[Liang et al., 2011] Liang, F., Liu, C., and Carroll, R. (2011). Advanced Markov chain Monte Carlo methods: learning from past samples, volume 714. John Wiley & Sons.

[Lindstrom and Bates, 1988] Lindstrom, M. J. and Bates, D. M. (1988). Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association, 83(404):1014–1022.

[Lindstrom and Bates, 1990] Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. Biometrics, pages 673–687.

[Lorenzi et al., 2011] Lorenzi, M., Ayache, N., and Pennec, X. (2011). Schild's ladder for the parallel transport of deformations in time series of images. In Information Processing in Medical Imaging, pages 463–474. Springer.

[Lorenzi et al., 2015] Lorenzi, M., Pennec, X., Frisoni, G. B., and Ayache, N. (2015). Disentangling normal aging from alzheimer's disease in structural magnetic resonance images. Neurobiology of aging, 36:S42–S52.

[Maclaurin and Adams, 2014] Maclaurin, D. and Adams, R. P. (2014). Firefly monte carlo: Exact mcmc with subsets of data. arXiv preprint arXiv:1403.5693.

[Marin et al., 2005] Marin, J.-M., Mengersen, K., and Robert, C. P. (2005). Bayesian modeling and inference on mixtures of distributions. Handbook of statistics, 25:459–507.

[Meyn and Tweedie, 2012] Meyn, S. P. and Tweedie, R. L. (2012). Markov chains and stochastic stability. Springer Science & Business Media.

[Michor, 2008] Michor, P. W. (2008). Topics in differential geometry, volume 93. American Mathematical Soc.

[Moakher and Zéraï, 2011] Moakher, M. and Zéraï, M. (2011). The riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. Journal of Mathematical Imaging and Vision, 40(2):171–187.

[Mohs et al., 1997] Mohs, R. C., Knopman, D., Petersen, R. C., Ferris, S. H., Ernesto, C., Grundman, M., Sano, M., Bieliauskas, L., Geldmacher, D., Clark, C., et al. (1997). Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the alzheimer's disease assessment scale that broaden its scope. Alzheimer Disease & Associated Disorders, 11:13–21.

[Muralidharan and Fletcher, 2012] Muralidharan, P. and Fletcher, P. T. (2012). Sasaki metrics for analysis of longitudinal data on manifolds. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 1027–1034. IEEE.

[Musso and Tricerri, 1988] Musso, E. and Tricerri, F. (1988). Riemannian metrics on tangent bundles. Annali di matematica pura ed applicata, 150(1):1–19.

[Ng et al., 2014] Ng, B., Dressler, M., Varoquaux, G., Poline, J.-B., Greicius, M., and Thirion, B. (2014). Transport on riemannian manifold for functional connectivity-based classification. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 405–412. Springer.

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

[Pennec et al., 2006] Pennec, X., Fillard, P., and Ayache, N. (2006). A riemannian framework for tensor computing. International Journal of Computer Vision, 66(1):41–66.

[Petersen, 2006] Petersen, P. (2006). Riemannian geometry, volume 171. Springer.

[Phillips et al., 2003] Phillips, S., Bandini, L., Compton, D., Naumova, E., and Must, A. (2003). A longitudinal comparison of body composition by total body water and bioelectrical impedance in adolescent girls. The Journal of nutrition, 133(5):1419–1425.

[Pinheiro and Bates, 2006] Pinheiro, J. and Bates, D. (2006). Mixed-effects models in S and S-PLUS. Springer Science & Business Media.

[Pinheiro, 1994] Pinheiro, J. C. (1994). Topics in mixed effects models. PhD thesis, University Of Wisconsin–Madison.

[Pinheiro and Bates, 1996] Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. Statistics and Computing, 6(3):289–296.

[Ring and Wirth, 2012] Ring, W. and Wirth, B. (2012). Optimization methods on riemannian manifolds and their application to shape space. SIAM Journal on Optimization, 22(2):596–627.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. The annals of mathematical statistics, pages 400–407.

[Robert and Casella, 2009] Robert, C. and Casella, G. (2009). Introducing Monte Carlo Methods with R. Springer Science & Business Media.

[Robert and Casella, 2013] Robert, C. and Casella, G. (2013). Monte Carlo statistical methods. Springer Science & Business Media.

[Roberts et al., 1997] Roberts, G. O., Gelman, A., Gilks, W. R., et al. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. The annals of applied probability, 7(1):110–120.

[Robinson, 1991] Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. Statistical science, pages 15–32.

[Rumpf and Wirth, 2012] Rumpf, M. and Wirth, B. (2012). Discrete geodesic calculus in the space of viscous fluidic objects. arXiv preprint arXiv:1210.0822.

[Savic et al., 2011] Savic, R. M., Mentré, F., and Lavielle, M. (2011). Implementation and evaluation of the saem algorithm for longitudinal ordered categorical data with an illustration in pharmacokinetics–pharmacodynamics. The AAPS journal, 13(1):44–53.

[Scheffé, 1956] Scheffé, H. (1956). A "mixed model" for the analysis of variance. The Annals of Mathematical Statistics, pages 23–36.

[Schiratti et al., 2015a] Schiratti, J.-B., Allassonnière, S., Colliot, O., and Durrleman, S. (2015a). Learning spatiotemporal trajectories from manifold-valued longitudinal data. In Advances in Neural Information Processing Systems, pages 2404–2412.

[Schiratti et al., 2015b] Schiratti, J.-B., Allassonnière, S., Colliot, O., and Durrleman, S. (2015b). Mixed-effects model for the spatiotemporal analysis of longitudinal manifold-valued data. In 5th MICCAI Workshop on Mathematical Foundations of Computational Anatomy.

[Schiratti et al., 2015c] Schiratti, J.-B., Allassonnière, S., Routier, A., Colliot, O., and Durrleman, S. (2015c). Estimating profiles of disease progression using mixed-effects models with time reparametrization. In Organisation for Human Brain Mapping.

[Schiratti et al., 2015d] Schiratti, J.-B., Allassonnière, S., Routier, A., Colliot, O., Durrleman, S., and the ADNI (2015d). A mixed-effects model with time reparametrization for longitudinal univariate manifold-valued data. In International Conference on Information Processing in Medical Imaging, pages 564–575. Springer.

[Shang et al., 2015] Shang, X., Zhu, Z., Leimkuhler, B., and Storkey, A. J. (2015). Covariance-controlled adaptive langevin thermostat for large-scale bayesian sampling. In Advances in Neural Information Processing Systems, pages 37–45.

[Sheiner and Beal, 1980] Sheiner, L. B. and Beal, S. L. (1980). 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.

[Siegel, 1964] Siegel, C. L. (1964). Symplectic geometry. Am. J. Math., 65:1–86.

[Singh et al., 2013] Singh, N., Hinkle, J., Joshi, S., and Fletcher, P. T. (2013). A hierarchical geodesic model for diffeomorphic longitudinal shape analysis. In Information Processing in Medical Imaging, pages 560–571. Springer.

[Singh et al., 2014] Singh, N., Hinkle, J., Joshi, S., and Fletcher, P. T. (2014). An efficient parallel algorithm for hierarchical geodesic models in diffeomorphisms. In 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pages 341–344. IEEE.

[Skovgaard, 1984] Skovgaard, L. T. (1984). A riemannian geometry of the multivariate normal model. Scandinavian Journal of Statistics, pages 211–223.

[Smith, 1994] Smith, S. T. (1994). Optimization techniques on riemannian manifolds. Fields institute communications, 3(3):113–135.

[Su et al., 2011] Su, J., Dryden, I. L., Klassen, E., Le, H., and Srivastava, A. (2011). Fitting optimal curves to time-indexed, noisy observations of stochastic processes on nonlinear manifolds. Journal of Image and Vision Computing.

[Thompson Jr, 1962] Thompson Jr, W. A. (1962). The problem of negative estimates of variance components. The Annals of Mathematical Statistics, pages 273–289.

[Tierney and Kadane, 1986] Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. Journal of the american statistical association, 81(393):82–86.

[Verbeke and Molenberghs, 2009] Verbeke, G. and Molenberghs, G. (2009). Linear mixed models for longitudinal data. Springer Science & Business Media.

[Wu, 1983] Wu, C. F. J. (1983). On the convergence properties of the em algorithm. The Annals of statistics, pages 95–103.

[Yang et al., 2011] Yang, E., Farnum, M., Lobanov, V., Schultz, T., Raghavan, N., Samtani, M. N., Novak, G., Narayan, V., and DiBernardo, A. (2011). Quantifying the pathophysiological timeline of alzheimer's disease. Journal of Alzheimer's Disease, 26(4):745–753.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Méthodes et algorithmes pour l'apprentissage de modèles d'évolution spatio-temporels à partir de données longitudinales sur une variété Riemannienne

Mots clefs : Géométrie Riemannienne, Données longitudinales, Modélisation statistique, Algorithme EM stochastique

Résumé : Dans ce manuscrit, nous présentons un modèle à effets mixtes, présenté dans un cadre Bayésien, permettant d'estimer la progression temporelle d'un phénomène biologique à partir d'observations répétées, à valeurs dans une variété Riemannienne, et obtenues pour un individu ou groupe d'individus. La progression est modélisée par des trajectoires continues dans l'espace des observations, que l'on suppose être une variété Riemannienne. La trajectoire moyenne est définie par les effets fixes du modèle. Pour définir les trajectoires de progression individuelles, nous avons introduit la notion de "variation parallèle" d'une courbe sur une variété Riemannienne. Pour chaque individu, une trajectoire individuelle est construite en considérant une variation parallèle de la trajectoire moyenne et en reparamétrisant en temps cette parallèle. Les transformations spatio-temporelles sujet-spécifiques, que sont la variation parallèle et la reparamétrisation temporelle, sont définies par les effets aléatoires du modèle et permettent de quantifier les changements de direction et vitesse à laquelle les trajectoires sont parcourues. Le cadre de la géométrie Riemannienne permet d'utiliser ce modèle générique avec n'importe quel type de données définies par des contraintes lisses. Une version stochastique de l'algorithme EM, le Monte Carlo Markov Chains Stochastic Approximation EM (MCMC-SAEM), est utilisé pour estimer les paramètres du modèle au sens du maximum *a posteriori*. L'utilisation du MCMC-SAEM avec un schéma numérique permettant de calculer le transport parallèle est discuté dans ce manuscrit. De plus, le modèle et le MCMC-SAEM sont validés sur des données synthétiques, ainsi qu'en grande dimension. Enfin, nous des résultats obtenus sur différents jeux de données liés à la santé.

Title : Models and algorithms to learn spatiotemporal changes from longitudinal manifold-valued observations

Keywords : Riemannian geometry, Longitudinal data, Statistical modeling, Stochastic EM algorithm

Abstract : We propose a generic Bayesian mixed-effects model to estimate the temporal progression of a biological phenomenon from manifold-valued observations obtained at multiple time points for an individual or group of individuals. The progression is modeled by continuous trajectories in the space of measurements, which is assumed to be a Riemannian manifold. The group-average trajectory is defined by the fixed effects of the model. To define the individual trajectories, we introduced the notion of "parallel variations" of a curve on a Riemannian manifold. For each individual, the individual trajectory is constructed by considering a parallel variation of the average trajectory and reparametrizing this parallel in time. The subject-specific spatiotemporal transformations, namely parallel variation and time reparametrization, are defined by the individual random effects and allow to quantify the changes in direction and pace at which the trajectories are followed. The framework of Riemannian geometry allows the model to be used with any kind of measurements with smooth constraints. A stochastic version of the Expectation-Maximization algorithm, the Monte Carlo Markov Chains Stochastic Approximation EM algorithm (MCMC-SAEM), is used to produce produce *maximum a posteriori* estimates of the parameters. The use of the MCMC-SAEM together with a numerical scheme for the approximation of parallel transport is discussed. In addition to this, the method is validated on synthetic data and in high-dimensional settings. We also provide experimental results obtained on health data.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France