Computational and statistical methods for trajectory analysis in a Riemannian geometry setting

Maxime Louis

Thèse présentée pour obtenir le grade de docteur

Université Pierre et Marie Curie

❧

École doctorale Informatique, Télécommunications et Électronique (Paris)

Computational and statistical methods for trajectory analysis in a Riemannian geometry setting

PAR : **Maxime Louis**

Sous la direction de STANLEY DURRLEMAN, Directeur de recherche INRIA Paris

MEMBRES DU JURY:

Co-directeur de thèse: Benjamin CHARLIER, Enseignant-chercheur, Université de Montpellier

Rapporteur: Xavier PENNEC, Directeur de Recherche INRIA Sophia-antipolis

Rapporteur: Marc NIETHAMMER, Professeur, UNC Chapel Hill

Examinateur : Alain TROUVÉ, Professeur, CMLA Cachan

Examinateur : Julien TIERNY, Chargé de recherche CNRS, Sorbonnes universités

Examinateur : Marleen DE BRUIJNE, Associate professor, Erasmus MC Rotterdam

Date de soutenance : 7 octobre 2019

Remerciements

Je souhaite tout d'abord adresser ma profonde gratitude à Stanley pour son encadrement sans faille tout au long de cette thèse. Je le remercie d'avoir été aussi bienveillant et de n'avoir jamais hésité à investir de l'énergie et du temps, même quand je refusais de m'orienter exactement vers la direction qu'il proposait. Ces instants passés à réfléchir tous les deux constituent un excellent souvenir pour moi, associé à un humour - pas toujours si drôle, avouons le- qe nous partageons tous les deux.

Je souhaite ensuite adresser de chaleureux remerciements à Benjamin, pour ses relectures nanométriques, son compilateur intégré de LaTeX, mais surtout pour sa capacité à ne rien lâcher même devant mes réticences les plus féroces. Je suis également heureux pour ces heures de conversations sur le canapé du labo, passées à parler maths bien entendu, mais aussi éducation, politique et escalade etc.

Je suis très reconnaissant à Xavier Pennec et à Marc Niethammer d'avoir accepté de rapporté cette thèse, pour la qualité de leurs remarques et pour l'intérêt qu'ils ont témoigné envers mon travail. Je remercie également Alain Trouvé, Julien Tierny et Marleen De Bruijne pour avoir accepté d'assister à ma soutenance et d'y amener leur expérience et leur recul.

Je voudrais continuer par remercier Olivier pour sa rigueur intellectuelle modèle et son appétit pour les conversations politiques et écologiques -qui m'aident à mettre en perspective certaines de mes dérives anarcho-écologistes! Je voudrais également le remercier pour l'énergie et le temps qu'il emploie pour faire de l'équipe aramis un lieu cordial et chaleureux. Cette aura bienveillante est palpable dans l'ambiance du labo, et c'est cette ambiance qui précisement a largement contribué à faire de ma thèse une période dont je me souviendrai avec nostalgie.

Ensuite, je tiens bien évidemment à remercier l'ensemble des membres de l'équipe, en commençant par mes compagnons de thèse. Alexandre bien évidemment, éternellement motivé, buveur de bières éponymes, aussi perfectionniste que je ne le suis pas ! Igor, cobureau comme on en voudrait toujours, dégageant plus d'énergie que le reste de la foule quand il danse. Manon, pour nos partages d'astuce en tout genre et de paroles écolos. Je tiens ensuite à remercier Simona, gardienne de l'ambiance et de la bonne cuisine, Tiziana toujours extrêmement gentille, Raphael toujours dans la lune, Jorge toujours à l'heure, Arnaud infatiguable compagnon de soirée, Pascal toujours près pour une bonne conversation, Jeremy toujours papa, Benoit toujours sympa, Catalina toujours dansante, Alexandre R toujours de très bon conseil, et tous ceux que je ne cite pas, avec qui j'ai partagé un moment.

Je tiens ensuite à remercier l'ensemble des grimpeurs avec qui j'ai pu pratiqué mon activité favorite, condition essentielle de mon bien-être. Esther, la plus flexible et tolérante des personnes que je connaisse, Alice, qui garde la pêche en tout situation, Baptiste, maitre dans l'art de chiller, Max, qui se livre à une bataille de blagues toujours plus acharnée, Faustine, que vous pouvez laisser sans problème dans des conditions apocalyptiques et retrouver de bonne humeur, Charlotte, complice de blagues en tout genre, Thomas, l'insaisissable compagnon de grimpe, sans oublier Christophe, Marion, Leni, Antoine, Quentin, Goki, Joachim, Louis, Alissa, Agathe et tant d'autres sans qui les vacances de grimpe ne seraient pas aussi drôles.

Je me dois de remercier tout particulièrement Hedi, mon nord, mon sud, mon est, mon ouest. Aussi bon coloc que grimpeur, aussi bon mathématicien que rieur, un modèle par ses valeurs, et un bonheur pour mon humeur. Je remercie ensuite bien évidemment Elise pour son sourire, Clément toujours présent pour les autres, Solène pour nos tennis et nos rires, Flora pour ses petites croyances parfois si amusantes, Malo maitre du deep learning, Catherine, Antoine, Alexandre et tous ceux avec qui j'ai pu partager un instant.

Enfin, je voudrais remercier ma famille. Mes parents Olivier et Frédérique pour leur téléguidage pendant toutes mes études et pour leur soutien indéfectible -qui s'étend bien au-delà de mon travail de thèse- ma soeur Marion, parce que sans elle rien ne serait pareil, mon frère Antoine pour sa gentillesse hors-norme et son expertise technique constante, mes nièces Ella et Diane parce qu'elles sont trop choues, ma tante Christel, mon oncle Gérard et mes cousins Arthur, Lionel, Samuel, Timothée et Maeva.

Introduction

The availability of large data sets and the increase in computational power have allowed to successfully apply machine learning methods to a wide variety of problems. This effect is most prominent in computer vision, where what used to be challenging problems are now routinely solved using convolutional networks. The advent of these automated solutions raises much interest for healthcare applications, and there is hope to provide:

- Automated diagnostic. Machine learning methods are trained to diagnose of a subject at the current time. The promise here is two-fold. First, a decrease in the number of false positives for a given pathology would allow to spare stressful, potentially invasive and expensive additional exams. Second, these methods, because virtually unlimited in their complexity and in the information that they can capture from observations, could allow to formulate diagnostics which usual approaches would not have detected.

- Prognosis i.e. the prediction of some aspects of the future state of a subject. Tasks of interest typically involve predicting a future diagnostic and the progression of symptoms such as cognitive decline. Variations around this topic include drug effect prediction and optimized treatment.

- Identification of underlying biological processes. The analysis of data sets of patients having a given disease can lead to the discovery of patterns which are informative about the nature of the disease e.g. the identification of genes involved in a disease, a particular pattern in Magnetic Resonance Imaging (MRI) etc. An interesting subclass of such approaches tackles the identification of sub-types within known diseases. Alzheimer's disease for instance is known to be heterogeneous and could in fact be a generic name for processes which vary in causes and effects.

In this PhD thesis, we tackle some of these problems in the case of neuro-degenerative diseases. We are particularly interested in modeling disease progression and in predicting the cognitive decline. Neuro-degenerative diseases are particularly challenging since their progression typically span several years, and single-time snapshot observations of subjects are much less informative than repeated observations [21]. We will therefore build methods which are able to gather information from longitudinal data sets i.e. data sets containing subjects which are repeatedly observed through time. The analysis of such data sets comes with challenges. Previous research work (e.g. [86]) was dedicated to handling the varying number of observations of each subject, managing their different time spacing and coping with the absence of known notion of time alignment: some subjects declare the disease earlier than others and some exhibit a faster progression than others. We will include solutions to face these challenges in all of our work. But we focus on dealing with new challenges which were often limiting for the approaches designed to create longitudinal disease progression models:

- High-dimensionality of the data;

- Absence of a priori knowledge to model the progression of the observations. It is possible to design models in a way that is consistent with the observed progressions when a lot of knowledge is available about an observed feature – as it is the case for cognitive scores or amyloid deposition in [88, 31]. However, in the more general case of imaging data and/or when the goal is precisely to unveil mechanisms of the disease process, such knowledge is not available. Therefore, we aim at learning the patterns of progression that are relevant to a specific disease and modality.

Riemannian geometry is often used to model disease progression and medical imaging data sets, for several reasons. First, it allows to handle data which obeys constraints e.g. symmetric positive definite tensors [80]. The constraint is directly embodied in the manifold which is considered. Second, it is in agreement with the so-called manifold hypothesis, which assumes that the data, although high-dimensional in its raw description, is in fact governed by a small number of intrinsic and hidden parameters with respect to which it varies smoothly. Third, we inherit from a vast literature in shape analysis [106], which often proposes to compare shapes by computing transformations between them and heavily relies on Riemannian geometry tools. A particular instance of this is the Large Deformation Diffeomorphic Metric Mapping (LDDMM) [74, 106], which provides a way to parametrize diffeomorphisms which act on shapes. The analysis of the shapes boils down to the comparison of diffeomorphisms which are ultimately described in a tangent vector space.

The first two parts of this thesis propose numerical methods and models for the analysis of manifold-valued data.

First, we propose a numerical scheme, the fanning scheme, for the computation of parallel transport, which remains computationally efficient as long as the inverse of the metric can be computed efficiently. The direct integration of the parallel transport equation by computation of the Christoffel symbols scales exponentially with the dimension. The Schild's and Pole's ladder [45, 61] –which constitute the sole other alternatives– are limited by the need to repeatedly compute Riemannian logarithms, an operation which is often expensive and approximate. We prove the convergence of the fanning scheme under simple conditions on the manifold. This computational block can be used in generative models on Riemannian manifolds. In spirit, the parallel transport allows to put into correspondence progressions —materialized by a tangent vector— observed at different points. It therefore constitutes a tool of choice for gathering information from multiple observed progressions on the manifold. We will show how to use this in a simple setting to predict the progression of sub-cortical structures with time in the course of Alzheimer's disease. This comes to complement existing methods for comparing trajectories on Riemannian manifolds [108, 102, 32].

Second, putting momentarily aside longitudinal data sets, we propose a generalization of Linear Discriminant Analysis [26] for manifold-valued data. In the spirit of principal geodesic analysis [27], it enables to find a geodesic sub-manifold which best summarizes the between-class differences with respect to the within-class differences. It is both a dimensionality reduction method as well as a classification method. We illustrate it on various shape data sets. Our claim is that this method is more intrinsic, in particular it is in principle independent from the coordinate system used on the manifold, when most other classification methods on manifolds heavily rely on coordinates and transformations of the data.

Until that point, we developed methods which enabled statistical analysis of manifold-valued data when the manifold is known a priori. This puts a strong limitation on the applicability of these methods. First and foremost because such Riemannian manifolds are not always available. Indeed, it is already hard in general to identify a differential manifold which is close to the data – it is a blooming field of research in itself. It is even harder to equip this manifold with a relevant Riemannian metric which is adapted to the task at hand. Note that the performances of a generative model formulated on a Riemannian manifold crucially depends on the manifold and on its metric.

Remark. One particular domain in which this is true is for the LDDMM framework. This framework postulates that shapes are obtained by diffeomorphic transformation of other shapes, where the considered diffeomorphisms belong to a fixed family of diffeomorphisms. There are of course some motivations for the construction of such families of diffeomorphisms, such as the notion of energy which should be minimized to keep the deformations as simple as possible. But we strongly believe that learning the kinds of diffeomorphisms that allow to generate an observed set of data could free us from hand-crafting these deformations and would specialize them to the task at hand. This led us to contribute to the work of Alexandre Bône [8], which we do not detail in this thesis.

Therefore, we would like to learn an adapted Riemannian manifold for the data at the same time as we learn a longitudinal model for disease progression. This involves the task of learning a Riemannian metric, which has rarely been studied so far. In [4, 2] the authors estimate a Riemannian metric so that an observed set of data maximizes the likelihood of a generative Bayesian model. But their approach makes strong assumptions on the candidate Riemannian metrics, limiting its applicability. In [53], the authors show how to use a transformation of the observation to pull-back a metric from a given space back to the observation space. They then optimize the transformation so that the data lies in regions of high volumes of the manifold equipped with this pull-back metric. To better understand how the problem of Riemannian manifold estimation can be tackled, we start by studying simple cases, both from a theoretical and experimental point of view. Namely, we looked at Riemannian manifolds so that:

- a given distribution is –close to– a normal distribution on this manifold,
- a set of observed curves are –close to– geodesics on this manifold.

We start by studying the simple case of a Riemannian metric on $\mathbb{R}$. In this setting, we provide explicit formulae and conditions for the Riemannian metric solving these two tasks. Then, we show how the first task is the one achieved by Generative Adversarial Networks [33]. For the second task, we prove an existence theorem when the set of curves obey mild assumptions. We then discuss uniqueness of the metrics which make these curves geodesics. For the experimental part, we propose a parametric family of Riemannian metrics on $\mathbb{R}^d$ and provide experimental results when optimizing cost functions associated with the two tasks, where the optimization is achieved using the automatic differentiation abilities of modern deep learning libraries.

From there we turn back to longitudinal disease progression and propose to construct a Riemannian manifold where all observed trajectories are geodesics 'parallel' to a common geodesic. The notion of parallel relies on parallel transport and is inspired from [88]. We resort there to the use of the fanning scheme for the experiments. This is a stronger version of the second task formulated above. This method, suffers from a curse of dimensionality and is impractical for imaging data. We therefore improve the approach by using a non-linear mapping from a low-dimensional latent space to the observation space. We illustrate how the optimization of this mapping amounts to a Riemannian geometry learning procedure. Experimental results in low-dimensional cases allow to recover previously seen results with the parametric metrics. Finally, a final extension of this model, presented in the last chapter of the thesis, consists in a longitudinal auto-encoder, which is able to handle any combination of modalities at any given visit in the data sets.

Organisation of the manuscript. In Part I, we present the numerical scheme for parallel transport along geodesics. We prove its linear convergence and show examples on synthetic and real data with the LDDMM framework. In Part 3, we present an extension of Linear Discriminant Analysis to manifold-valued data. This part deals with non-longitudinal data and we propose the eager reader to skip it on their first read. Finally, in part III, we provide some results when learning a Riemannian manifold optimizing the criteria specified above, as well as the longitudinal model for disease progression which does perform Riemannian manifold learning. In the Appendix, we provide short introductions to Riemannian geometry and to the LDDMM framework.

Table of contents

Introduction **i**

I Parallel transport: an efficient numerical scheme and its applications **1**

1 A numerical scheme and a proof of convergence **5**

2 Application to shape analysis [62, 10] **31**

II Geodesic Discriminant Analysis for manifold-valued data **49**

3 Geodesic Discriminant Analysis for manifold-valued data **51**

III Riemannnian geometry learning **67**

4 Riemannian metrics so as to be normally distributed **75**

5 Riemannian metrics for geodesicity **87**

6 Disease modelling using deep neural networks [66] **103**

7 Longitudinal auto-encoder for multimodal disease progression modeling [14] **115**

IV Conclusion and perspectives **127**

V Appendix **131**

8 Riemannian geometry **133**

9 Large Deformation Diffeomorphic Metric Mapping (LDDMM) **143**

List of publications **147**

Bibliography **148**

Part I

Parallel transport: an efficient numerical scheme and its applications

This part consists of two chapters. First, we give an algorithm for computing the parallel transport of a vector along a geodesic on a Riemannian manifold. We prove that the algorithm converges as the inverse of the number of discretisation steps. This part is a reproduction of the journal publication [65]. In the second part, we detail how to implement this algorithm when working with diffeomorphisms parametrized by control points and initial momenta –as introduced in details in Appendix 9– and propose several applications for the prediction of the progression of sub-cortical structures during the course of Alzheimer's disease. This second part contains a combination of the conference papers [62] and [11]. This work has been extended and included in a more comprehensive framework in [7].

Chapter 1

A numerical scheme and a proof of convergence

1.1 Introduction

Riemannian geometry has been long contained within the field of pure mathematics and theoretical physics. Nevertheless, there is an emerging trend to use the tools of Riemannian geometry in statistical learning to define models for structured data. Such data may be defined by invariance properties and therefore seen as points in quotient spaces as for shapes, orthogonal frames, or linear subspaces. They may be defined also by smooth inequalities, and therefore as points in open subsets of linear spaces, as for symmetric positive definite matrices, diffeomorphisms or bounded measurements. Such data may be considered therefore as points in a Riemannian manifolds and analysed by specific statistical approaches [109, 55, 87, 58]. At the core of these approaches lies parallel transport, an isometry between tangent spaces which allows the comparison of probability density functions, coordinates or vectors that are defined in the tangent space at different points on the manifold. The inference of such statistical models in practical situations requires efficient numerical schemes to compute parallel transport on manifolds.

The parallel transport of a given tangent vector is defined as the solution of an ordinary differential equation ([70] page 52), written in terms of the Christoffel symbols. The computation of the Christoffel symbols requires access to the metric coefficients and their derivatives, making the equation integration using standard numerical schemes very costly in situations where no closed-form formulas are available for the metric coefficients or their derivatives.

An alternative is to use Schild's ladder [45], or its faster version in the case of geodesics, the pole ladder [61]. These schemes essentially require the computation of Riemannian exponentials and logarithms at each step. Usually, the computation of the exponential may be done by integrating Hamiltonian equations and does not raise specific difficulties. By contrast, the computation of the logarithm must often be done by solving an inverse problem with the use of an optimization scheme such as a gradient descent. Such optimization schemes are approximate and sensitive to the initial conditions and to hyper-parameters, which leads to additional numerical errors –most of the time uncontrolled– as well as an increased computional cost. When closed formulas exist for the Riemannian logarithm, or in the case of Lie groups, where the logarithm can be approximated efficiently using the Baker-Campbell-Hausdorff formula (see [59]), Schild's ladder is an efficient alternative. When this is not the case, it becomes hardly tractable. A more detailed analysis of the convergence of Schild's ladder method can be found in [85].

Another alternative is to use an equation showing that parallel transport along geodesics may be locally approximated by a well-chosen Jacobi field, up to a second order error. This idea has been suggested in [105] with further credits to [98], but without either a formal definition nor a proof of its convergence. It relies solely on the computations of Riemannian exponentials.

In this paper, we propose a numerical scheme built on this idea, which tries to limit as much as possible the number of operations required to reach a given accuracy. We will show how to use only the inverse of the metric and its derivatives when performing the different steps of the scheme. This different set of requirements makes the scheme attractive in a different set of situations than the integration of the ODE or the Schild's ladder. We will prove that this scheme converges at linear speed with the time-step, and that this speed may not be improved without further assumptions on the manifold. Furthermore, we propose an implementation which allows the simultaneous computation of the geodesic and of the transport along this geodesic. Numerical experiments on the 2-sphere and on the manifold of 3-by-3 symmetric positive definite matrices will confirm that the convergence of the scheme is of the same order as Schild's ladder in practice. Thus, they will show that this scheme offers a compelling alternative to compute parallel transport with a control over the numerical errors and the computational cost.

1.2 Rationale

1.2.1 Notations and assumptions

In this paper, we assume that γ is a geodesic defined for all time $t > 0$ on a smooth manifold $\mathcal{M}$ of finite dimension $n \in \mathbb{N}$ provided with a smooth Riemannian metric g. We denote the Riemannian exponential Exp and ∇ the covariant derivative. For $p \in \mathcal{M}$, $T_p\mathcal{M}$ denotes the tangent space of $\mathcal{M}$ at p. For all $s, t \geq 0$ and for all $w \in T_{\gamma(s)}\mathcal{M}$, we denote $\mathrm{P}_{s,t}(w) \in T_{\gamma(t)}\mathcal{M}$ the parallel transport of w from $\gamma(s)$ to $\gamma(t)$. It is the unique solution at time t of the differential equation $\nabla_{\dot\gamma(u)}\mathrm{P}_{s,u}(w) = 0$ for $\mathrm{P}_{s,s}(w) = w$. We also denote $\mathrm{J}^w_{\gamma(t)}(h)$ the Jacobi field emerging from $\gamma(t)$ in the direction $w \in T_{\gamma(t)}\mathcal{M}$, that is

$$\mathrm{J}^w_{\gamma(t)}(h) = \left.\frac{\partial}{\partial\varepsilon}\right|_{\varepsilon=0} \mathrm{Exp}_{\gamma(t)}(h(\dot\gamma(t) + \varepsilon w)) \in T_{\gamma(t+h)}\mathcal{M}$$

for $h \in \mathbb{R}$ small enough. It verifies the Jacobi equation (see for instance [70] page 111-119)

$$\nabla^2_{\dot{\gamma}} J^w_{\gamma(t)}(h) + R(J^w_{\gamma(t)}(h), \dot{\gamma}(h))\dot{\gamma}(h) = 0 \tag{1.1}$$

where R is the curvature tensor. We denote $\|\cdot\|_g$ the Riemannian norm on the tangent spaces defined from the metric g, and $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ the metric at any $p \in \mathcal{M}$. We use Einstein notations.

We fix Ω a compact subset of $\mathcal{M}$ such that Ω contains a neighborhood of $\gamma([0,1])$. We also set $w \in T_{\gamma(0)}\gamma$ and $w(t) = P_{0,t}(w)$. We suppose that there exists a coordinate system on Ω and we denote $\Phi : \Omega \longrightarrow U$ the corresponding diffeomorphism, where U is a subset of $\mathbb{R}^n$. This system of coordinates allows us to define a basis of the tangent space of $\mathcal{M}$ at any point of Ω, we denote $\left.\frac{\partial}{\partial x^i}\right|_p$ the i-th element of the corresponding basis of $T_p\mathcal{M}$ for any $p \in \mathcal{M}$. Note finally that, since the injectivity radius is a smooth function of the position on the manifold (see [70]) and since it is everywhere positive on Ω, there exists $\eta > 0$ such that for all p in Ω, the injectivity radius at p is larger than η.

The problem in this paper is to provide a way to compute an approximation of $P_{0,1}(w)$.

We suppose throughout the paper the existence of a single coordinate chart defined on Ω. In this setting, we propose a numerical scheme which gives an error varying linearly with the size of the integration step. Once this result is established, since in any case $\gamma([0,1])$ can be covered by finitely many charts, it is possible to apply the proposed method to parallel transport on each chart successively. The errors during this computation of the parallel transport would add, but the convergence result remains valid.

1.2.2 The key identity

The numerical scheme that we propose arises from the following identity, which is mentioned in [105]. Figure 1.1 illustrates the principle.

Proposition 1. *For all $t > 0$, and $w \in T_{\gamma(0)}\mathcal{M}$ we have*

$$\mathrm{P}_{0,t}(w) = \frac{\mathrm{J}^w_{\gamma(0)}(t)}{t} + \mathrm{O}\Big(t^2\Big). \tag{1.2}$$

Proof. Let $X(t) = P_{0,t}(w)$ be the vector field following the parallel transport equation: $\dot{X}^i + \Gamma^i_{kl}X^l\dot{\gamma}^k = 0$ with $X(0) = w$, where $(\Gamma^i_{kl})_{i,j,k\in\{1,\dots,n\}}$ are the Christoffel symbols associated with the Levi-Civita connection for the metric g. In normal coordinates centered at $\gamma(0)$, the Christoffel symbols vanish at $\gamma(0)$ and the equation gives: $\dot{X}^i(0) = 0$. A Taylor expansion of $X(t)$ near $t = 0$ in this local chart then reads

$$X^i(t) = w^i + \mathrm{O}\Big(t^2\Big). \tag{1.3}$$

By definition, the i-th normal coordinate of $\mathrm{Exp}_{\gamma(0)}(t(v_0 + \varepsilon w))$ is $t(v_0^i + \varepsilon w^i)$. Therefore,

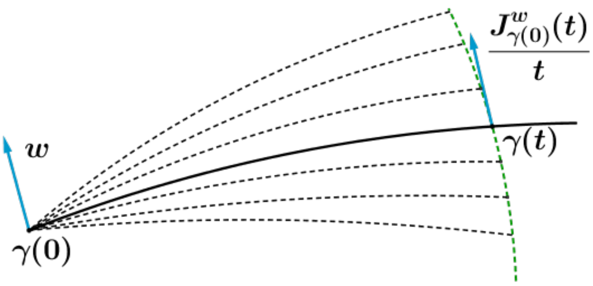


Figure 1.1: The solid line is the geodesic. The green dotted line is formed by the perturbed geodesics at time t. The blue arrows are the initial vector and its approximated parallel transport at time t.

the i-th coordinate of $\mathrm{J}^w_{\gamma(0)}(t) = \frac{\partial}{\partial \varepsilon}|_{\varepsilon=0}\mathrm{Exp}_{\gamma(0)}(t(\dot\gamma(0)+\varepsilon w))$ is tw^i. Plugging this into (1.3) yields the desired result. □

This control on the approximation of the transport by a Jacobi field suggests to divide $[0,1]$ into N intervals $[\frac{k}{N}, \frac{k+1}{N}]$ of length $h = \frac{1}{N}$ for $k = 0, \dots, N-1$ and to approximate the parallel transport of a vector $w \in T_{\gamma(0)}$ from $\gamma(0)$ to $\gamma(1)$ by a sequence of vectors $w_k \in T_{\gamma\left(\frac{k}{N}\right)}\mathcal{M}$ defined as

$$\begin{cases} w_0 = w \\ w_{k+1} = N\mathrm{J}^{w_k}_{\gamma\left(\frac{k}{N}\right)}\left(\dfrac{1}{N}\right). \end{cases} \tag{1.4}$$

With the control given in the Proposition 1, we can expect to get an error of order $\mathrm{O}\left(\frac{1}{N^2}\right)$ at each step and hence a speed of convergence in $\mathrm{O}\left(\frac{1}{N}\right)$ overall. There are manifolds for which the approximation of the parallel transport by a Jacobi field is exact e.g. Euclidean space, but in the general case, one cannot expect to get a better convergence rate. Indeed, we show in the next section that this scheme for the sphere $\mathbb{S}^2$ has a speed of convergence exactly proportional to $\frac{1}{N}$.

1.2.3 Convergence rate on $\mathbb{S}^2$

In this section, we assume that one knows the geodesic path $\gamma(t)$ and how to compute any Jacobi fields without numerical errors, and show that the approximation due to Equation (1.2) alone raises a numerical error of order $\mathrm{O}\left(\frac{1}{N}\right)$.

Let $p \in \mathbb{S}^2$ and $v \in T_p\mathbb{S}^2$ (p and v are seen as vectors in $\mathbb{R}^3$). The geodesics are the great circles, which may be written as

$$\gamma(t) = \mathrm{Exp}_p(tv) = \cos(t|v|)p + \sin(t|v|)\frac{v}{|v|},$$

where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^3$. Using spherical coordinates (θ, ϕ) on the sphere, chosen so that the whole geodesic is in the coordinate chart, we get coordinates on the tangent space at any point $\gamma(t)$. In this spherical system of coordinates, it is straightforward to see that the parallel transport of $w = p \times v$ along $\gamma(t)$ has constant coordinates, where $\times$ denote the usual cross-product on $\mathbb{R}^3$.

We assume now that $|v| = 1$. Since $w = p \times v$ is orthogonal to v, we have $\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} |v + \varepsilon w| = 0$. Therefore,

$$\begin{aligned} \mathrm{J}_p^w(t) &= \frac{\partial}{\partial \varepsilon}|_{\varepsilon=0} \left(\cos(t|v+\varepsilon w|)p + \sin(t|v+\varepsilon w|)\frac{v+\varepsilon w}{|v+\varepsilon w|} \right) \\ &= \sin(t) w \end{aligned}$$

which does not depend on p. We have $\mathrm{J}_{\gamma(t)}^w(t) = \sin(t)w$. Consequently, the sequence of vectors w_k built by the iterative process described in equation (1.4) verifies $w_{k+1} = N w_k \sin\left(\frac{1}{N}\right)$ for $k = 0, \ldots, N-1$, and $w_N = w_0 N \sin\left(\frac{1}{N}\right)^N$. Now in the spherical coordinates, $P_{0,1}(w_0) = w_0$, so that the numerical error, measured in these coordinates, is proportional to $w_0 \left(1 - \left(\frac{\sin(1/N)}{1/N}\right)^N\right)$. We have

$$\left(\frac{\sin(1/N)}{1/N}\right)^N = \exp\left(N \log\left(1 - \frac{1}{6N^2} + \mathrm{o}\big(1/N^2\big)\right)\right) = 1 - \frac{1}{6N} + \mathrm{o}\Big(\frac{1}{N}\Big)$$

yielding

$$\frac{|w_N - w_0|}{|w_0|} \propto \frac{1}{6N} + \mathrm{o}\Big(\frac{1}{N}\Big).$$

It shows a case where the bound $\frac{1}{N}$ is reached.

1.3 The numerical scheme

1.3.1 The algorithm

In general, there are no closed form expressions for the geodesics and the Jacobi fields. Hence, in most practical cases, these quantities also need to be computed using numerical methods.

Computing geodesics In order to avoid the computation of the Christoffel symbols, we propose to integrate the first-order Hamiltonian equations to compute geodesics. Let $x(t) = (x_1(t), \ldots, x_d(t))^T$ be the coordinates of $\gamma(t)$ in a given local chart, and $\alpha(t) = (\alpha_1(t), \ldots, \alpha_d(t))^T$ be the coordinates of the momentum $g_{\gamma(t)}(\dot\gamma(t), \cdot) \in T^*_{\gamma(t)}\mathcal{M}$ in the same local chart. We have then (see [106])

$$\begin{cases} \dot{x}(t) = K(x(t))\alpha(t) \\ \dot{\alpha}(t) = -\dfrac{1}{2}\nabla_x \left(\alpha(t)^T K(x(t))\alpha(t)\right) \end{cases}, \tag{1.5}$$

where $K(x(t))$, a d-by-d matrix, is the inverse of the metric g expressed in the local chart. Note that using (1.5) to integrate the geodesic equation will require us to convert initial tangent vectors into initial momenta, as seen in the algorithm description below.

Computing $\mathrm{J}^{w}_{\gamma(t)}(h)$ The Jacobi field may be approximated with a numerical differentiation from the computation of a perturbed geodesic with initial position $\gamma(t)$ and initial velocity $\dot{\gamma}(t) + \varepsilon w$ where ε is a small parameter

$$\mathrm{J}^{w}_{\gamma(t)}(h) \simeq \frac{\mathrm{Exp}_{\gamma(t)}\Big(h(\dot{\gamma}(t) + \varepsilon w)\Big) - \mathrm{Exp}_{\gamma(t)}\Big(h\dot{\gamma}(t)\Big)}{\varepsilon}, \tag{1.6}$$

where the Riemannian exponential may be computed by integration of the Hamiltonian equations (1.5) over the time interval $[t, t+h]$ starting at point $\gamma(t)$, as shown on Figure 1.2. We will also see that a choice for ε ensuring a $\mathrm{O}\Big(\frac{1}{N}\Big)$ order of convergence is $\varepsilon = \frac{1}{N}$.

The algorithm Let $N \in \mathbb{N}$. We divide $[0,1]$ into N intervals $[t_k, t_{k+1}]$ with $t_k = \frac{k}{N}$ and denote $h = \frac{1}{N}$ the size of the integration step. We initialize $\gamma_0 = \gamma(0)$, $\dot{\gamma}_0 = \dot{\gamma}(0)$, $\tilde{w}_0 = w$ and solve $\tilde{\beta}_0 = K^{-1}(\gamma_0)\tilde{w}_0$ and $\tilde{\alpha}_0 = K^{-1}(\gamma_0)\dot{\gamma}_0$. We propose to compute, at step k:

1. The new point $\tilde{\gamma}_{k+1}$ and momentum $\tilde{\alpha}_{k+1}$ of the main geodesic, by performing one step of length h of a second-order Runge-Kutta method on equation (1.5).

2. The perturbed geodesic starting at $\tilde{\gamma}_k$ with initial momentum $\tilde{\alpha}_k + \varepsilon\tilde{\beta}_k$ at time h, that we denote $\tilde{\gamma}^{\varepsilon}_{k+1}$, by performing one step of length h of a second-order Runge-Kutta method on equation (1.5).

3. The estimated parallel transport

$$\hat{w}_{k+1} = \frac{\tilde{\gamma}^{\varepsilon}_{k+1} - \tilde{\gamma}_{k+1}}{h\varepsilon}. \tag{1.7}$$

4. The corresponding momentum $\hat{\beta}_{k+1}$, by solving: $K(\tilde{\gamma}_{k+1})\hat{\beta}_{k+1} = \hat{w}_{k+1}$.

At the end of the scheme, $\tilde{w}_N$ is the proposed approximation of $P_{0,1}(w)$. Figure 1.2 illustrates the principle. A complete pseudo-code is given in appendix 1.7.1. It is remarkable that we can substitute the computation of the Jacobi field with only four calls to the Hamiltonian equations (1.5) at each step, including the calls necessary to compute the main geodesic. Note however that the step (4) of the algorithm requires to solve a linear system of size n. Solving the linear system can be done with a complexity less than cubic in the dimension (in $\mathrm{O}\Big(n^{2.374}\Big)$ using the Coppersmith–Winograd algorithm).

1.3.2 Possible variations

There are a few possible variations of the presented algorithm.

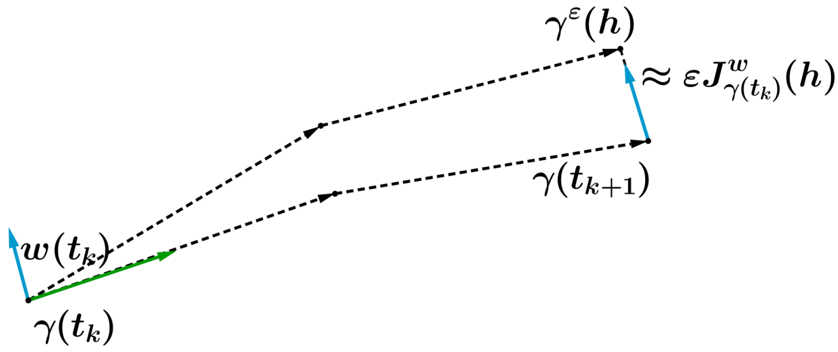


Figure 1.2: One step of the numerical scheme. The dotted arrows represent the steps of the Runge-Kutta integrations for the main geodesic γ and for the perturbed geodesic γ^{ε}. The blue arrows are the initial $w(t_k)$ and the obtained approximated transport using equation (1.6), with $h = t_{k+1} - t_k$.

1. The first variation is to use higher-order Runge-Kutta methods to integrate the geodesic equations at step (1) and (2). We prove that a second-order integration of the geodesic equation is enough to guarantee convergence and notice experimentally the absence of convergence with a first order integration of the geodesic equation. Experiments indicate a linear convergence with an improved constant using this variation. Depending on the situation, the extra computations required at each step may be counterbalanced by this increased precision.

2. The second variation uses a higher-order finite difference scheme by replacing step (2) and step (3) the following way. At the k-th iteration, compute two perturbed geodesics starting at $\tilde{\gamma}_k$ and with initial momentum $\tilde{\alpha}_k + \epsilon\tilde{\beta}_k$ (resp. $\tilde{\alpha}_k - \epsilon\tilde{\beta}_k$) at time h, that we denote $\tilde{\gamma}_{k+1}^{+\varepsilon}$ (resp. $\tilde{\gamma}_{k+1}^{-\varepsilon}$), by performing one step of length h of a second-order Runge-Kutta method on equation (1.5). Then proceed to a second-order differentiation to approximate the Jacobi field, and set:

$$\hat{w}_{k+1} = \frac{\tilde{\gamma}_{k+1}^{+\varepsilon} - \tilde{\gamma}_{k+1}^{-\varepsilon}}{2h\varepsilon}. \tag{1.8}$$

   Empirically, this variation does not seem to bring any substantial improvement to the scheme.

3. The final variation of the scheme consists in adding an extra renormalization step at the end of each iteration:

   (v) Renormalize the momentum and the corresponding vector using

$$\tilde{\beta}_{k+1} = a_k\hat{\beta}_{k+1} + b_k\tilde{\alpha}_{k+1}$$

$$\tilde{w}_{k+1} = K(\tilde{\gamma}_{k+1})\tilde{\beta}_{k+1}$$

   where a_k and b_k are factors ensuring $\tilde{\beta}_{k+1}^{\top}K(\tilde{\gamma}_{k+1})\tilde{\beta}_{k+1} = \beta_0^{\top}K(\gamma_0)\beta_0$ and

$\tilde{\beta}_{k+1}^\top K(\tilde{\gamma}_{k+1})\tilde{\alpha}_{k+1} = \beta_0^\top K(\gamma_0)\alpha_0$. Indeed, the quantities $\beta(t)^\top K(\gamma(t))\beta(t)$ and $\beta(t)^\top K(\gamma(t))\alpha(t)$ are preserved along the parallel transport. This extra step is cheap even when the dimension is large. Empirically, it leads to the same rate of convergence with a smaller constant.

We will show that the proposed algorithm and the variations 1 and 2 ensure convergence of the final estimate. We do not prove convergence with the variation 3, but this additionnal step can be expected to improve the quality of the approximation at each step, at least when the discretization is sufficiently thin, by enforcing the conversation of quantities which should be conserved. Note that the best accuracy for a given computational cost is not necessarily obtained with the method in Section 1.3.1, but might be attained with one of the proposed variations, as a bit more computations at each step may be counter-balanced by a smaller constant in the convergence rate.

1.3.3 The convergence Theorem

We obtain the following convergence result, guaranteeing a linear decrease of the error with the size of the step h.

Theorem 1. *We suppose here the hypotheses stated in Section 1.2.1. Let $N \in \mathbb{N}$ be the number of integration steps. Let $w \in T_{\gamma(0)}\mathcal{M}$ be the vector to be transported. We denote the error*

$$\delta_k = \|P_{0,t_k}(w) - \tilde{w}_k\|_2$$

where $\tilde{w}_k$ is the approximate value of the parallel transport of w along γ at time t_k and where the 2-norm is taken in the coordinates of the chart Φ on Ω. We denote ε the parameter used in the step (2) and $h = \frac{1}{N}$ the size of the step used for the Runge-Kutta approximate solution of the geodesic equation.
If we take $\varepsilon = h$, then we have

$$\delta_N = O\Big(\frac{1}{N}\Big).$$

We will see in the proof and in the numerical experiments that choosing $\varepsilon = h$ is a recommended choice for the size of the step in the differentiation of the perturbed geodesics. Further decreasing ε has no visible effect on the accuracy of the estimation and choosing a larger ε lowers the quality of the approximation.

Note that our result controls the 2-norm of the error in the global system of coordinates, but not directly the metric norm in the tangent space at $\gamma(1)$. This is due to the fact that $\gamma(1)$ is not accessible, but only its approximation $\tilde{\gamma}_N$ computed by the Runge-Kutta integration of the Hamiltonian equation. However, Theorem 1 implies that the couple $(\tilde{\gamma}_N, \tilde{w}_N)$ converges towards $(\gamma(1), P_{0,1}(w))$ using the ℓ^2 distance on $\mathcal{M} \times T\mathcal{M}$ and a coordinate system in a neighborhood of $\gamma(1)$, which is equivalent to any distance on $\mathcal{M} \times T\mathcal{M}$ on this neighborhood and hence is the right notion of convergence.

We give the proof in the next Section. The proof of some technical lemmas used in the proof are given in Section 1.8.

1.4 Proof of the convergence Theorem 1

We prove the convergence of the algorithm.

Proof. We will denote, as in the description of the algorithm in Section 1.3, $\gamma_k = \gamma(t_k)$, $\tilde{\gamma}_k = \tilde{\gamma}(t_k)$ its approximation in the algorithm. Let N be a number of discretization steps and $k \in \{1, \ldots, N\}$. We build an upper bound on the error δ_{k+1} from δ_k. We have

$$\begin{aligned}
\delta_{k+1} &= \|w_{k+1} - \tilde{w}_{k+1}\|_2 \\
&\leq \underbrace{\left\| w_{k+1} - \frac{\mathrm{J}_{\gamma_k}^{w_k}(h)}{h} \right\|_2}_{(1)} + \underbrace{\left\| \frac{\mathrm{J}_{\gamma_k}^{w_k}(h)}{h} - \frac{\mathrm{J}_{\gamma_k}^{\tilde{w}_k}(h)}{h} \right\|_2}_{(2)} \\
&\quad + \underbrace{\left\| \frac{\mathrm{J}_{\gamma_k}^{\tilde{w}_k}(h)}{h} - \frac{\mathrm{J}_{\tilde{\gamma}_k}^{\tilde{w}_k}(h)}{h} \right\|_2}_{(3)} + \underbrace{\left\| \frac{\mathrm{J}_{\tilde{\gamma}_k}^{\tilde{w}_k}(h)}{h} - \frac{\tilde{\mathrm{J}}_{\tilde{\gamma}_k}^{\tilde{w}_k}(h)}{h} \right\|_2}_{(4)}
\end{aligned}$$

where

- $\tilde{\gamma}_k$ is the approximation of the geodesic coordinates at step k.
- $w_k = w(t_k)$ is the exact parallel transport.
- $\tilde{w}_k$ is its approximation at step k
- $\tilde{\mathrm{J}}$ is the approximation of the Jacobi field computed with finite difference: $\tilde{\mathrm{J}}_{\tilde{\gamma}_k}^{\tilde{w}_k} = \frac{\tilde{\gamma}_{k+1}^{\varepsilon} - \tilde{\gamma}_{k+1}}{\varepsilon}$.
- $\mathrm{J}_{\tilde{\gamma}_k}^{\tilde{w}_k}(h)$ is the exact Jacobi field computed with the approximations $\tilde{w}$, $\tilde{\gamma}$ and $\tilde{\dot{\gamma}}$ *i.e.* the Jacobi field defined from the geodesic with initial position $\tilde{\gamma}_k$, initial momentum $\tilde{\alpha}_k$, with a perturbation $\tilde{w}_k$.

We provide upper bounds for each of these terms. We start by assuming $\|w_k\|_2 \leq 2\|w_0\|_2$, before showing it is verified for any $k \leq N$ when N is large enough. We could assume more generally $\|w_k\|_2 \leq C\|w_0\|_2$ for any $C > 1$. The idea is to get a uniform control on the errors at each step by assuming that $\|w_k\|_2$ does not grow too much, and to show afterwards that the control we get is tight enough to ensure, when the number of integration steps is large, that we do have $\|w_k\|_2 \leq 2\|w_0\|_2$.

Term (1) This is the intrinsic error when using the Jacobi field. We show in Proposition **??** that for h small enough

$$\left\| P_{t_k,t_{k+1}}(w_k) - \frac{\mathrm{J}^{w_k}_{\gamma_k}(h)}{h} \right\|_{g(\gamma(t_{k+1}))} \leq \; Ah^2 \|w_k\|_g = Ah^2 \|w_k\|_g.$$

Now, since g varies smoothly and by equivalence of the norms, there exists $A' > 0$ such that

$$\left\| P_{t_k,t_{k+1}}(w_k) - \frac{\mathrm{J}^{w_k}_{\gamma(k)}(h)}{h} \right\|_2 \leq A'h^2 \|w_k\|_2 \leq 2A'h^2 \|w_0\|_2 \tag{1.9}$$

Term (2) Lemma 1.8 show that for h small enough

$$\left\| \frac{\mathrm{J}^{w_k}_{\gamma(t_k)}(h)}{h} - \frac{\mathrm{J}^{\tilde{w}_k}_{\gamma(t_k)}(h)}{h} \right\|_2 \leq (1 + Bh)\delta_k. \tag{1.10}$$

Term (3) This term measures the error linked to our approximate knowledge of the geodesic γ. It is proved in Appendix 1.8 that there exists a constant $C > 0$ which does not depend on k or h such that

$$\left\| \frac{\mathrm{J}^{\tilde{w}_k}_{\gamma_k}(h)}{h} - \frac{\tilde{\mathrm{J}}^{\tilde{w}_k}_{\gamma_k}(h)}{h} \right\|_2 \leq Ch^2. \tag{1.11}$$

Term (4) This is the difference between the analytical computation of J and its approximation. It is proved in Appendix 1.8 and 1.8 that if we use a Runge-Kutta method of order 2 to compute the geodesic points $\gamma^{\varepsilon}_{k+1}$ and γ_{k+1} and a first-order differentiation to compute the Jacobi field as described in the step (3) of the algorithm, or if we use two perturbed geodesics $\gamma^{\varepsilon}_{k+1}$ and $\gamma^{-\varepsilon}_{k+1}$ and a second-order differentiation method to compute the Jacobi field as described in equation (1.8), there exists $D \geq 0$ which does not depend on k such that:

$$\left\| \frac{\mathrm{J}^{\tilde{w}_k}_{\gamma(t_k)} - \tilde{\mathrm{J}}^{\tilde{w}_k}_{\gamma(t_k)}}{h} \right\|_2 \leq D(h^2 + \varepsilon h). \tag{1.12}$$

Note that this majoration is valid as long as $\tilde{w}_k$ is bounded by a constant which does not depend on k or N, which we have assumed so far.

Gathering equations (1.9), (1.10), (1.11) and (1.12), there exists a constant $F > 0$ such that for all k such that $\|w_i\|_2 \leq \|w_0\|_2$ for all $i \leq k$:

$$\delta_{k+1} \leq (1 + Bh)\delta_k + F(h^2 + h\varepsilon). \tag{1.13}$$

Combining those inequalities for $k = 1, \dots, s$ where $s \in \{1, \dots, N\}$ is such that $\|w_k\|_2 \leq$

$2\|w_0\|_2$ for all $k \leq s$, we obtain a geometric series whose sum yields

$$\delta_s \leq \frac{F(h^2 + h\varepsilon)}{Bh}(1 + Bh)^{s+1}. \tag{1.14}$$

We now show that for a large enough number of integration steps N, this implies that $\|w_k\|_2 \leq 2\|w_0\|_2$ for all $k \in \{1, \dots, N\}$. We proceed by contradiction, assuming that there exist arbitrary large $N \in \mathbb{N}$ for which there exists $u(N) \leq N$ – that we take minimal – such that $\|w_{u(N)}\|_2 > 2\|w_0\|_2$. For any such $N \in \mathbb{N}$, since $u(N)$ is minimal with that property, we can still use equation (1.14) with $s = u(N)$:

$$\delta_{u(N)} \leq \frac{F(h^2 + h\varepsilon)}{Bh}(1 + Bh)^{u(N)+1}. \tag{1.15}$$

Now, $h = \frac{1}{N}$ so that

$$\delta_{u(N)} \leq \frac{F(h + \varepsilon)}{B}(1 + Bh)^{u(N)+1} \leq \frac{F(h + \varepsilon)}{B}(1 + Bh)^{\frac{1}{h}+1}. \tag{1.16}$$

But we have, on the other hand:

$$\|w_0\|_2 < |\|\tilde{w}_{u(N)}\|_2 - \|w_0\|_2| \leq \|\tilde{w}_{u(N)} - w_0\|_2 \leq \frac{F(h + \varepsilon)}{B}(1 + Bh)^{\frac{1}{h}+1} \tag{1.17}$$

Taking $\varepsilon \leq h$, which we will keep as an assumption in the rest of the proof, the term on the right goes to zero as $h \to 0$ – *i.e.* as $N \to \infty$ – which is a contradiction. So for N large enough, we have $\|w_k\|_2 \leq 2\|w_0\|_2$ and equation (1.14) holds for all $k \in \{1, \dots, N\}$. With $s = N$, equation (1.14) reads:

$$\delta_N \leq \frac{F(h^2 + h\varepsilon)}{Bh}(1 + Bh)^{N+1}.$$

We see that choosing $\varepsilon = \frac{1}{N}$ yields an optimal rate of convergence: choosing a larger value deteriorates the accuracy of the scheme while choosing a lower value still yields an error in $\mathrm{O}\left(\frac{1}{N}\right)$. Setting $\varepsilon = \frac{1}{N}$:

$$\delta_N \leq \frac{2F}{BN}\left(1 + \frac{B}{N}\right)^{N+1} = \frac{2F}{BN}\left(\exp(B) + \mathrm{o}\left(\frac{1}{N}\right)\right).$$

Eventually, there exists $G > 0$ such that, for $N \in \mathbb{N}$ large enough

$$\delta_N \leq \frac{G}{N}.$$

□

1.5 Numerical experiments

1.5.1 Setup

We implemented the numerical scheme on simple manifolds where the parallel transport is known in closed form, allowing us to evaluate the numerical error [1]. We present two examples:

- $\mathbb{S}^2$: in spherical coordinates (θ, ϕ) the metric is $g = \begin{pmatrix} 1 & 0 \\ 0 & \sin(\theta)^2 \end{pmatrix}$. We gave expressions for geodesics and parallel transport in Section 1.2.3.

- The set of 3×3 symmetric positive-definite matrices SPD(3). The tangent space at any points of this manifold is the set of symmetric matrices. In [55], the authors endow this space with the affine-invariant metric: for $\Sigma \in \mathrm{SPD}(3)$, $V, W \in \mathrm{Sym}(3)$, $g_\Sigma(V, W) = \mathrm{tr}(\Sigma^{-1} V \Sigma^{-1} W)$. Through an explicit computation of the Christoffel symbols, they derive explicit expressions for any geodesic $\Sigma(t)$ starting at $\Sigma_0 \in \mathrm{SPD}(3)$ with initial tangent vector $X \in \mathrm{Sym}(3)$: $\Sigma(t) = \Sigma_0^{\frac{1}{2}} \exp(tX) \Sigma_0^{\frac{1}{2}}$ where $\exp : \mathrm{Sym}(3) \to \mathrm{SPD}(3)$ is the matrix exponentiation. Deriving an expression for the parallel transport can also be done using the explicit Christoffel symbols, see [88]. If $\Sigma_0 \in \mathrm{SPD}(3)$ and $X, W \in \mathrm{Sym}(3)$, then

$$P_{0,t}(W) = \exp\left(\frac{t}{2} X \Sigma_0^{-1}\right) W \exp\left(\frac{t}{2} \Sigma_0^{-1} X\right).$$

The code for this numerical scheme can be written in a generic way and used for any manifold by specifying the Hamiltonian equations and the inverse of the metric. For experiments in large dimensions, we refer to [63].

Remark Note that even though the computation of the gradient of the inverse of the metric with respect to the position, $\nabla_x K$, is required to integate the Hamiltonian equations (1.5), $\nabla_x K$ can be computed from the gradient of the metric using the fact that any smooth map $M : \mathbb{R} \to GL_n(\mathbb{R})$ verifies $\frac{\mathrm{d}M^{-1}}{\mathrm{d}t} = -M^{-1} \frac{\mathrm{d}M}{\mathrm{d}t} M^{-1}$. This is how we proceeded for SPD(3): it spares some potential difficulties if one does not have access to analytical expressions for the inverse of the metric. It is however a costful operation which requires the computation of the full inverse of the metric at each step.

1.5.2 Results

Errors measured in the chosen system of coordinates confirm the linear behavior in both cases, as shown on Figures 1.3 and 1.4.

[1] A modular Python version of the code is available here: `https://gitlab.icm-institute.org/maxime.louis/parallel-transport`

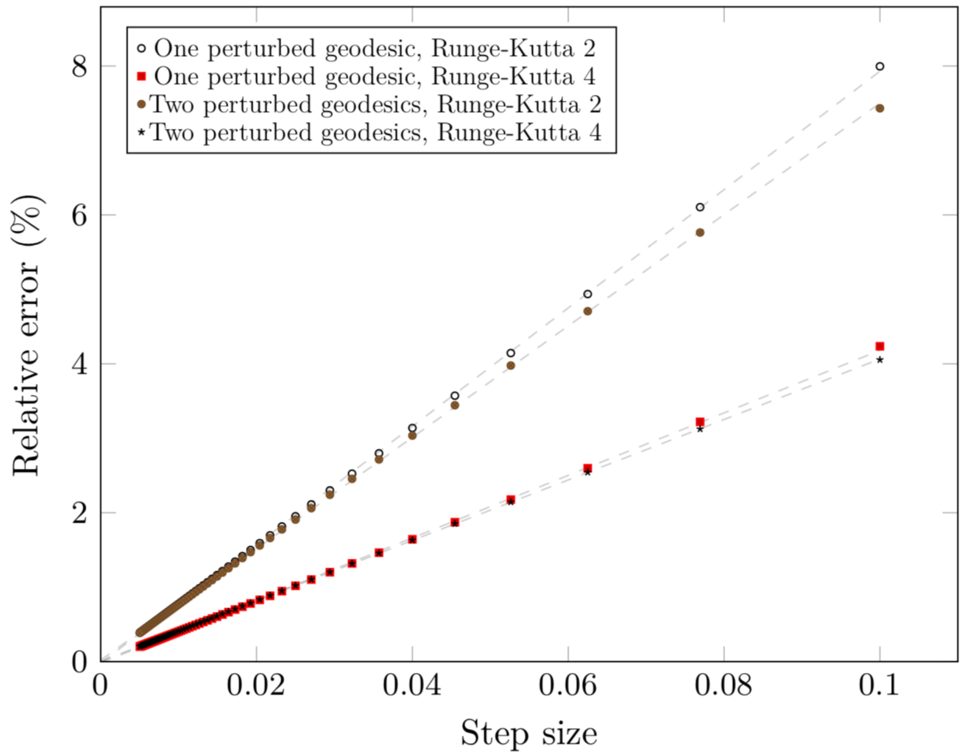


Figure 1.3: Relative errors for the 2-Sphere in different settings, as functions of the step size, with initial point, velocity and initial w kept constant. The dotted lines are linear regressions of the measurements. Runge-Kutta 2 (resp. 4) indicates that a Runge-Kutta method or order 2 (resp. 4) is used for the integration of the geodesic equation.

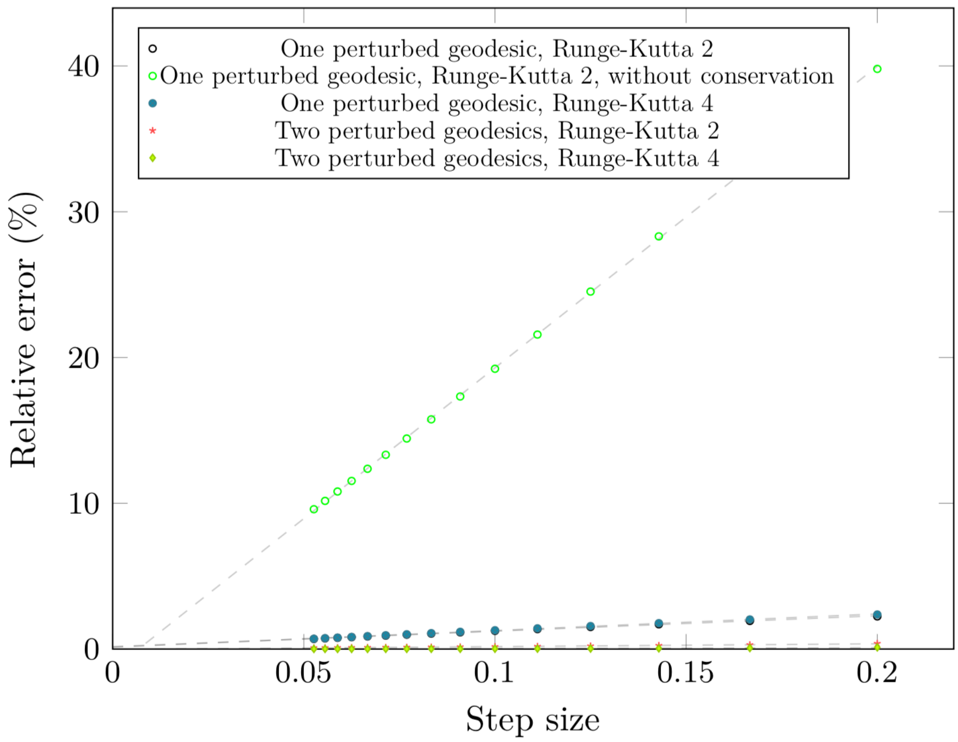


Figure 1.4: Relative errors for SPD(3) in different settings, as functions of the step size, with initial point, velocity and initial w kept constant. The dotted lines are linear regressions. Runge-Kutta 2 (resp. 4) indicate that a second-order (resp. fourth order) Runge-Kutta integration has been used to integrate the geodesic equations at steps (1) and (2). Without conservation indicates that step 3 has not been used.

We assessed the effect of a higher order for the Runge-Kutta scheme in the integration of geodesics. Using a fourth order method increases the accuracy of the transport in both cases, by a factor 2.3 in the single geodesic case. A fourth order method is twice as expensive as a second order method in terms of number of calls to the Hamiltonian equations, hence in this case it is the most efficient way to reach a given accuracy.

We also investigated the effect of using the variation 3 of the algorithm, which enforces conservation of the transported vector norm and of its scalar product with the geodesic velocity. Doing so yields an exact transport for the sphere because it is of dimension 2 and the conservation of two quantities is enough to ensure an exact transport –up to the fact that the geodesic is computed approximately– so that the actually observed error is the error in the integration of the geodesic equation. It yields a dramatically improved transport of the same order of convergence for SPD(3) (see Figure 1.4). The complexity of this operation is very low, and we recommend to always use it. It can be expected however that the effect of the enforcement of these conservations will lower as the dimension increases, since it only fixes two components of the transported vector.

We also confirmed numerically that without a second-order method to integrate the geodesic equations at steps (1) and (2) of the algorithm, the scheme does not converge. This is not in contradiction with Theorem 1 which supposes this integration is done with

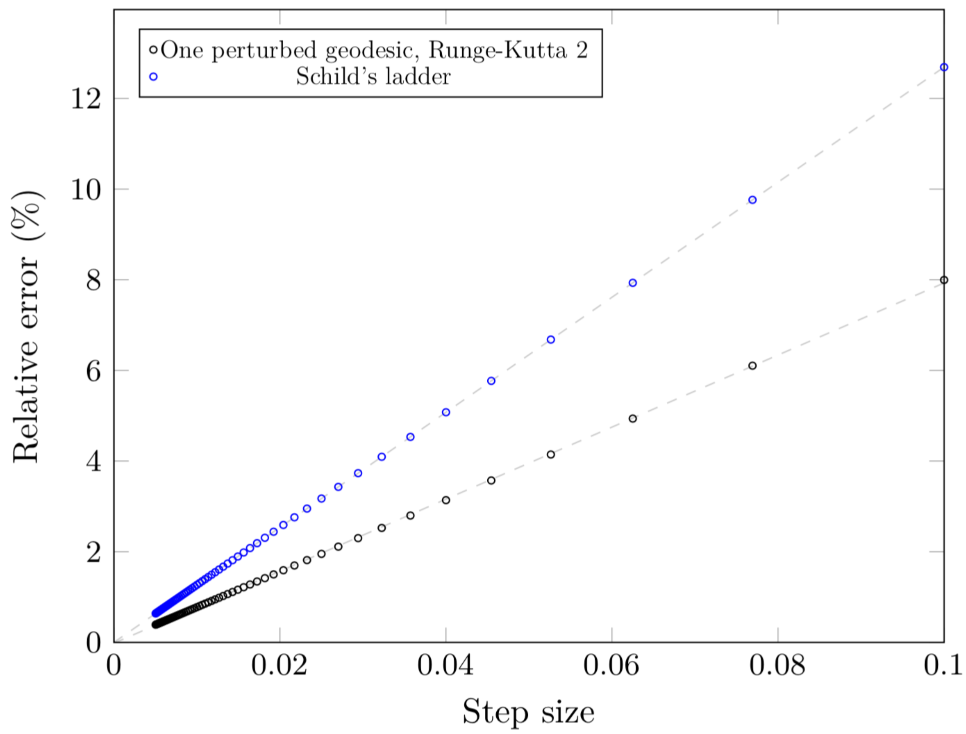


Figure 1.5: Relative error of Schild's ladder scheme compared to the fanning scheme (double geodesic, Runge-Kutta 2) proposed here, in the case of $\mathbb{S}^2$.

a second-order Runge Kutta method.

Finally, using two geodesics to compute a central-finite difference for the Jacobi field is 1.5 times more expensive than using a single geodesic, in terms of number of calls to the Hamiltonian equations, and it is therefore more efficient to compute two perturbed geodesics in the case of the symmetric positive-definite matrices.

1.5.3 Comparison with Schild's ladder

We compared the relative errors of the fanning scheme with Schild's ladder. We implemented Schild's ladder on the sphere and compared the relative errors of both schemes on a same geodesic and vector. We chose this vector to be orthogonal to the velocity, since the transport with Schild's ladder is exact if the transported vector is colinear to the velocity. We use a closed form expression for the Riemannian logarithm in Schild's ladder, and closed form expressions for the geodesic. The results are given in Figure 1.5.

1.6 Conclusion

We proposed a new method, the fanning scheme, to compute parallel transport along a geodesic on a Riemannian manifold using Jacobi fields. In contrast to Schild's ladder, this method does not require the computation of Riemannian logarithms, which may not be given in closed form and potentially hard to approximate. We proved that the error of the scheme is of order $\mathrm{O}\left(\frac{1}{N}\right)$ where N is the number of discretization steps, and that it cannot be improved in the general case, yielding the same convergence rate as Schild's ladder. We also showed that only four calls to the Hamiltonian equations are necessary at each step to provide a satisfying approximation of the transport, two of them being used to compute the main geodesic.

A limitation of this scheme is to only be applicable when parallel transporting along geodesics, and this limitation seems to be unavoidable with the identity it relies on. Note also that the Hamiltonian equations are expressed in the cotangent space whereas the approximation of the transport computed at each step lies in the tangent space to the manifold. Going back and forth from cotangent to tangent space at each iteration is costly if the metric is not available in closed-form, as it requires the inversion of a system. In very high dimensions this might limit the performances

1.7 Pseudo-code and proofs

1.7.1 Pseudo-code

We give a pseudo-code description of the numerical scheme. Here, $G(p)$ denotes the metric matrix at p for any $p \in \mathcal{M}$.

function ParallelTransport(x_0, α_0, w_0, N)
  function $\mathrm{V}(x, \alpha)$
    return $\mathrm{K}(x)\alpha$
  end function

  function $\mathrm{F}(x, \alpha)$
    return $-\frac{1}{2}\nabla_x\left(\alpha^T \mathrm{K}(x)\alpha\right)$ ▷ in closed form or by finite differences
  end function

▷ γ_0 coordinates of $\gamma(0)$
▷ α_0 coordinates of $G(\gamma(0))\dot{\gamma}(0) \in T^*_{\gamma(0)}\mathcal{M}$
▷ w_0 coordinates of $w \in T_{\gamma(0)}\mathcal{M}$
▷ β_0 coordinates of $G(\gamma(0))w_0$
▷ N number of time-steps

  $h = 1/N$, $\varepsilon = 1/N$
  for $k = 0, \dots, (N-1)$ **do**
▷ integration of the main geodesic
    $\gamma_{k+\frac{1}{2}} = \gamma_k + \frac{h}{2}v_k$
    $\alpha_{k+\frac{1}{2}} = \alpha_k + \frac{h}{2}\mathrm{F}(\gamma_k, \alpha_k)$
    $\gamma_{k+1} = \gamma_k + h\mathrm{V}(\gamma_{k+\frac{1}{2}}, \alpha_{k+\frac{1}{2}})$

$\alpha_{k+1} = \alpha_k + h\mathrm{F}(\gamma_{k+\frac{1}{2}}, \alpha_{k+\frac{1}{2}})$

▷ perturbed geodesic equation in the direction w_k

$\gamma^{\varepsilon}_{k+\frac{1}{2}} = \gamma_k + \frac{h}{2} v(\gamma_k, \alpha_k + \epsilon \beta_k)$

$\alpha^{\varepsilon}_{k+\frac{1}{2}} = \alpha_k + \varepsilon\beta_k + \frac{h}{2}\mathrm{F}(()\gamma^{\varepsilon}_k, \alpha_k + \varepsilon\beta_k$

$\gamma^{\varepsilon}_{k+1} = \gamma^{\varepsilon}_k + h\mathrm{V}(\gamma^{\varepsilon}_{k+\frac{1}{2}}, \alpha^{\varepsilon}_k + \frac{1}{2})$

▷ Jacobi field by finite differences

$\hat{w}_{k+1} = \frac{\gamma^{\varepsilon}_{k+1} - \gamma_{k+1}}{h\varepsilon}$

$\hat{\beta}_{k+1} = g(\gamma_{k+1}) w_{k+1}$ ▷ Use explicit g or solve $K(\gamma_{k+1})\hat{\beta}_{k+1} = \hat{w}_{k+1}$

▷ Conserve quantities

Solve for a, b:

$\beta_0^\top K(\gamma_0)\beta_0 = (a\hat{\beta}_{k+1} + b\alpha_{k+1})^\top K(\tilde{\gamma}_{k+1})(a\hat{\beta}_{k+1} + b\alpha_{k+1}),$

$\alpha_0^\top K(\gamma_0)\alpha_0 = (a\hat{\beta}_{k+1} + b\alpha_{k+1})^\top K(\tilde{\gamma}_{k+1})(a\hat{\beta}_{k+1} + b\alpha_{k+1}, v_{k+1})$

$\beta_{k+1} = a\hat{\beta}_{k+1} + b\alpha_{k+1}$ ▷ parallel transport

$w_{k+1} = K(\gamma_{k+1})\beta_{k+1}$

end for

return γ_N, α_N, w_N

▷ γ_N approximation of $\gamma(1)$

▷ α_N approximation of $G(\gamma(1))\dot{\gamma}(1)$

▷ w_N approximation of $\mathrm{P}_{\gamma(0),\gamma(1)}(w_0)$

end function

1.8 Proofs

A lemma to change coordinates

We recall that we suppose the geodesic contained within a compact subset Ω of the manifold $\mathcal{M}$. We start with a result controlling the norms of change-of-coordinates matrices. Let $pin\mathcal{M}$ and $q = \mathrm{Exp}_p(v)$ where $\|v\|_g \leq \frac{\eta}{2}$, where $\eta > 0$ is a lower bound on the injectivity radius on Ω. We consider two bases of $T_q\mathcal{M}$: one defined from the global system of coordinates, that we denote B^{Φ}_q, and another made of the normal coordinates centered at p, built from the coordinate on $T_p\mathcal{M}$ obtained from the coordinate chart Φ, that we denote B^N_q. We can therefore define $\Lambda(p, q)$ as the change-of-coordinates matrix between B^{Φ}_q and B^N_q. The operator norms $|||\cdot|||$ of these matrices are bounded over Ω in the following sense:

Lemma 1. *There exists $L \geq 0$ such that for all $p \in K$ and for all $q \in K$ such that $q = \mathrm{Exp}_p(v)$ for some $v \in T_p\mathcal{M}$ with $\|v\|_g \leq \frac{\eta}{2}$, we have*

$$|||\Lambda(p, q)||| \leq L$$

and

$$|||\Lambda^{-1}(p,q)||| \leq L.$$

Proof. Any two norms on $T_q\mathcal{M}$ are equivalent, and the norm bounds of the coordinate change smoothly depend on p and q by smoothness of the metric. Hence the result. $\square$

This lemma allows us to translate any bound on the components of a tensor in the global system of coordinates into a bound on the components of the same tensor in any of the normal systems of coordinates centered at a point of the geodesic, and *vice versa*.

Transport and connection

We prove a result connecting successive covariant derivatives to parallel transport:

Proposition 2. *Let V be a vector field on $\mathcal{M}$. Let $\gamma : [0,1] \to \mathcal{M}$ be a geodesic. Then*

$$\nabla^k_{\dot\gamma} V(\gamma(t)) = \left.\frac{\mathrm{d}^k}{\mathrm{d}h^k}\right|_{h=0} P^{-1}_{t,t+h}(V(\gamma(t+h))). \tag{1.18}$$

Proof. Let $E_i(0)$ be an orthonormal basis of $T_{\gamma(0)}\mathcal{M}$. Using the parallel transport along γ, we get orthonormal basis $E_i(s)$ of $T_{\gamma(t)}\mathcal{M}$ for all t. For $t \in [0,1]$, denote $(a_i(t))_{i=1,\ldots,n}$ the coordinates of $V(\gamma(t))$ in the basis $(E_i(t))_{i=1,\ldots,n}$. We have

$$\frac{\mathrm{d}^k}{\mathrm{d}h^k} P^{-1}_{t,t+h}(V(\gamma(t+h)) = \frac{\mathrm{d}^k}{\mathrm{d}h^k} P^{-1}_{t,t+h}\Big(\sum_{i=1}^n a_i(t+h)E_i(t+h)\Big) = \sum_{i=1}^n \frac{\mathrm{d}^k a_i(t+h)}{\mathrm{d}h^k} E_i(t)$$

because $P^{-1}_{t,t+h}E_i(t+h) = E_i(t)$ does not depend on h. On the other hand

$$\nabla^k_{\dot\gamma} V(\gamma(t)) = \nabla^k_{\dot\gamma} \sum_{i=1}^n a_i(t)E_i(t) = \sum_{i=1}^n \nabla^k_{\dot\gamma}(a_i(t))E_i(t) = \sum_{i=1}^n \frac{\mathrm{d}^k a_i(t+h)}{\mathrm{d}h^k} E_i(t)$$

by definition of $E_i(s)$. $\square$

A stronger version of Proposition 1

From there, we can prove a stronger version of Proposition 1. As before, η denotes a lower bound on the injectivity radius of $\mathcal{M}$ on Ω.

Proposition 3. *There exists $A \geq 0$ such that for all $t \in [0,1[$, for all $w \in T_{\gamma(t)}\mathcal{M}$ and for all $h < \frac{\eta}{\|\dot\gamma(t)\|_g}$ we have*

$$\left\| P_{t,t+h}(w) - \frac{\mathrm{J}^w_{\gamma(t)}(h)}{h} \right\|_g \leq Ah^2 \|w\|_g.$$

Proof. Let $t \in [0,1[$, $w \in T_{\gamma(t)}\mathcal{M}$ and $h < \frac{\eta}{\|\dot\gamma(t)\|_g}$ *i.e.* such that $J^w_{\gamma(t)}(h)$ is well defined. From Lemma 2, for any smooth vector field V on $\mathcal{M}$,

$$\nabla^k_{\dot\gamma(t)} V(\gamma(t)) = \left.\frac{\mathrm{d}^k}{\mathrm{d}h^k}\right|_{h=0} P^{-1}_{t,t+h}(V(\gamma(t+h)). \tag{1.19}$$

We will use this identity to obtain a development of $V(\gamma(t+h)) = J^w_{\gamma(t)}(h)$ for small h.

We have $J^w_{\gamma(t)}(0) = 0$, $\nabla_{\dot\gamma} J^w_{\gamma(t)}(0) = w$, $\nabla^2_{\dot\gamma} J^w_{\gamma(t)}(0) = -R(J^w_{\gamma(t)}(0), \dot\gamma(0))\dot\gamma(0) = 0$ using equation (1.1) and finally

$$\begin{aligned} \|\nabla^3_{\dot\gamma} J^w_{\gamma(t)}(h)\|_g &= \|\nabla_{\dot\gamma}(R)(J^w_{\gamma(t)}(h), \dot\gamma(h))\dot\gamma(h) + R(\nabla_{\dot\gamma} J^w_{\gamma(t)}(h), \dot\gamma(h))\dot\gamma(h)\|_g \\ &\leq \|\nabla_{\dot\gamma} R\|_\infty \|\dot\gamma(h)\|_g^2 \|J^w_{\gamma(t)}(h)\|_g + \|R\|_\infty \|\dot\gamma(h)\|^2_g \|\nabla_{\dot\gamma} J^w_{\gamma(t)}(h)\|_g, \end{aligned} \tag{1.20}$$

where the ∞-norms, taken over the geodesic and the compact Ω, are finite because the curvature and its derivatives are bounded. Note that we used $\nabla_{\dot\gamma}\dot\gamma = 0$ which holds since γ is a geodesic. In normal coordinates centered at $\gamma(t)$, we have $J^w_{\gamma(t)}(h)^i = hw^i$. Therefore, if we denote $g_{ij}(\gamma(t+h))$ the components of the metric in normal coordinates, we get using Einstein notations

$$\|J^w_{\gamma(t)}(h)\|^2_g = h^2 g_{ij}(\gamma(t+h)) w^i w^j.$$

To obtain an upper bound for this term which does not depend on t, we note that the coefficients of the metric in the global coordinate system are bounded on Ω. Using Lemma 1, we get a bound $M \geq 0$ valid on all the systems of normal coordinates centered at a point of the geodesic, so that

$$\|J^w_{\gamma(t)}(h)\|_g \leq hM\|w\|_2.$$

By equivalence of the norms as seen in Lemma (1), and because g varies smoothly, there exists $N \geq 0$ such that

$$\|J^w_{\gamma(t)}(gh)\|_g \leq hMN\|w\|_g \tag{1.21}$$

where the dependence of the majoration on t has vanished, and the result stays valid for all $h < \max\left(\frac{\eta}{\|\dot\gamma(t)\|_g}, 1-t\right)$ and all w. Similarly, there exists $C > 0$ such that

$$\|\nabla_{\dot\gamma} J^w_{\gamma(s)}(h)\| \leq C\|w\|_g, \tag{1.22}$$

at any point and for any $h < \max\left(\frac{\eta}{\|\dot\gamma(t)\|_g}, 1-t\right)$. Gathering equations (1.20), (1.21) and (1.22), we get that there exists a constant $A \geq 0$ which does not depend on t, h or w such that

$$\left\|\nabla^3_{\dot\gamma} J^w_{\gamma(s)}(h)\right\|_g \leq A\|w\|_g. \tag{1.23}$$

Now using equation (1.19) with $V(\gamma(t+h)) = J^w_{\gamma(t)}(h)$ and a Taylor's formula, we get

$$P^{-1}_{t,t+h}(J^w_{\gamma(t)}(h)) = hw + h^3 r(h,w)$$

where r is the remainder of the expansion, controlled in equation (1.23). We thus get

$$\left\| \frac{J^w_{\gamma(t)}(h)}{h} - P_{t,t+h}(w) \right\|_g = \|P_{t,t+h}(h^3 r(w,h))\|_g.$$

Now, because the parallel transport is an isometry, we can use our control (1.23) on the remainder to get

$$\left\| \frac{J^w_{\gamma(t)}(h)}{h} - P_{t,t+h}(w) \right\|_g \le \frac{A}{6} h^2 \|w\|_g.$$

□

A Lemma to control error accumulation

At every step of the scheme, we compute a Jacobi field from an approximate value of the transported vector. We need to control the error made with this computation from an already approximate vector. We provide a control on the 2-norm of the corresponding error, in the global system of coordinates.

Lemma 2. *There exists $B \ge 0$ such that for all $t \in [0,1[$, for all $w_1, w_2 \in T_{\gamma(t)}\mathcal{M}$ and for all $h \le \frac{\eta}{\|\dot\gamma(t)\|_g}$ small enough, we have :*

$$\left\| \frac{J^{w_1}_{\gamma(t)}(h) - J^{w_2}_{\gamma(t)}(h)}{h} \right\|_2 \le (1 + Bh)\|w_1 - w_2\|_2. \tag{1.24}$$

Proof. Let $t \in [0,1[$ and $h \le \frac{\eta}{\|\dot\gamma(t)\|_g}$. We denote $p = \gamma(t)$, $q = \gamma(t+h)$. We use the exponential map to get normal coordinates on a neighborhood V of p from the basis $\left(\frac{\partial}{\partial x^i}\Big|_p\right)_{i=1,\dots,n}$ of $T_p\mathcal{M}$. Let's denote $\left(\frac{\partial}{\partial y^i}\Big|_r\right)_{i=1,\dots,n}$ the basis obtained in the tangent space at any point r of V from this system of normal coordinates centered at p. At any point r in V, there are now two different bases of $T_r\mathcal{M}$: $\left(\frac{\partial}{\partial y^i}\Big|_r\right)_{i=1,\dots,n}$ obtained from the normal coordinates and $\left(\frac{\partial}{\partial x^i}\Big|_r\right)_{i=1,\dots,n}$ obtained from the coordinate system Φ. Let $w_1, w_2 \in T_p\mathcal{M}$ and denote w^i_j for $i \in \{1,\dots,n\}$, $j \in \{1,2\}$ the coordinates in the global system Φ. By definition, the basis $\left(\frac{\partial}{\partial y^i}\Big|_p\right)_{i=1,\dots,n}$ and the basis $\left(\frac{\partial}{\partial x^i}\Big|_p\right)_{i=1,\dots,n}$ coincide, and in particular, for $j \in \{1,2\}$:

$$w_j = (w_j)^i \left.\frac{\partial}{\partial x^i}\right|_p = (w_j)^i \left.\frac{\partial}{\partial y^i}\right|_p.$$

If $i \in \{1, \dots, n\}$, $j \in \{1, 2\}$, the j-th coordinate of $J^{w_i}_{\gamma(t)}(h)$ in the basis $\left(\frac{\partial}{\partial y^i}\Big|_q \right)_{i=1,\dots,n}$ is

$$J^{w_j}_{\gamma(t)}(h)^i = \frac{\partial}{\partial \varepsilon}\bigg|_{\varepsilon=0} (\mathrm{Exp}_p(h(v + \varepsilon w_j)))^i = \frac{\partial}{\partial \varepsilon}\bigg|_{\varepsilon=0} (h(v + \varepsilon w_j))^i = h w^i_j.$$

Let $\Lambda(\gamma(t+h), \gamma(t))$ be the change-of-coordinate matrix of $T_{\gamma(t+h)}$ from the basis $\left(\frac{\partial}{\partial y^i}\Big|_q \right)_{i=1,\dots,n}$ to the basis $\left(\frac{\partial}{\partial x^i}\Big|_q \right)_{i=1,\dots,n}$. Λ varies smoothly with t and h, and is the identity when $h = 0$. Hence, we can write an expansion

$$\Lambda(\gamma(t+h), \gamma(t)) = Id + hW(t) + O(h^2).$$

The second order term depends on the second derivative of Λ with respect to h. Restricting ourselves to a compact subset of $\mathcal{M}$, as in Lemma 1, we get a uniform bound on the norm of this second derivative thus getting a control on the operator norm of $\Lambda(\gamma(t+h), \gamma(t))$, that we can write, for h small enough

$$|||\Lambda(\gamma(t+h), \gamma(t))||| \leq (1 + Bh)$$

where B is a positive constant which does not depend on h or t. Now we get

$$\left\| \frac{J^{w_1}_{\gamma(t)}(h) - J^{w_2}_{\gamma(t)}(h)}{h} \right\|_2 = \|\Lambda(\gamma(t+h), \gamma(t))(w_1 - w_2)\|_2 \leq (1 + Bh)\, \|w_1 - w_2\|_2$$

which is the desired result. □

Proof that we can compute the geodesic simultaneously with a second-order method

We give here a control on the error made in the scheme when computing the main geodesic approximately and simultaneously with the parallel transport. We assume that the main geodesic is computed with a second-order method, and we need to control the subsequent error on the Jacobi field. The computations are made in global coordinates, and the error is measured by the 2-norm in these coordinates. $\Phi : \Omega \to U$ denotes the corresponding diffeomorphism. We denote $\eta > 0$ a lower bound on the injectivity radius of $\mathcal{M}$ on Ω and $\varepsilon > 0$ the parameter used to compute the perturbed geodesics at step (2).

Proposition 4. *There exists $A > 0$ such that for all $t \in [0, 1[$, for all $h \in [0, 1 - t]$, for all $w \in T_{\gamma(t)}\mathcal{M}$:*

$$\left\| \frac{\mathrm{J}^{\tilde{w}_k}_{\gamma_k}(h)}{h} - \frac{\mathrm{J}^{\tilde{w}_k}_{\tilde{\gamma}_k}(h)}{h} \right\|_2 \leq Ah^2.$$

Proof. Let $t \in [0,1[$, $h \in [0, 1-t]$, and $w \in T_{\gamma(t)}\mathcal{M}$. The term rewrites

$$\left\| \frac{\mathrm{J}^{\tilde{w}_k}_{\gamma_k}(h)}{h} - \frac{\mathrm{J}^{\tilde{w}_k}_{\tilde{\gamma}_k}(h)}{h} \right\|_2 = \left\| \left.\frac{\partial \mathrm{Exp}_{\gamma_k}(h\dot{\gamma}_k + x\tilde{w}_k)}{\partial x}\right|_{x=0} - \left.\frac{\partial \mathrm{Exp}_{\tilde{\gamma}_k}(h\dot{\tilde{\gamma}}_k + x\tilde{w}_k)}{\partial x}\right|_{x=0} \right\|_2. \tag{1.25}$$

This is the difference between the derivatives of two solutions of the same differential equation (1.5) with two different initial conditions. More precisely, we define $\Pi : \Phi(\Omega) \times B_{\mathbb{R}^n}(0, \|\tilde{\gamma}_k\| + 2\varepsilon\|\tilde{w}_k\|) \times [0,\eta]) \to \mathbb{R}^n$ such that $\Pi(p_0, \alpha_0, h)$ are the coordinates of the solutions of the Hamiltonian equation at time h with initial coordinates p_0 and initial momentum α_0. Π is the flow, in coordinates, of the geodesic equation. We can now rewrite equation (1.25)

$$\left\| \frac{\mathrm{J}^{\tilde{w}_k}_{\gamma_k}(h)}{h} - \frac{\mathrm{J}^{\tilde{w}_k}_{\tilde{\gamma}_k}(h)}{h} \right\|_2 = \left\| \left.\frac{\partial \Pi(\gamma_k, \dot{\gamma}_k + \varepsilon\tilde{w}_k, h)}{\partial \varepsilon}\right|_{\varepsilon=0} - \left.\frac{\partial \Pi(\tilde{\gamma}_k, \dot{\tilde{\gamma}}_k + \varepsilon\tilde{w}_k, h)}{\partial \varepsilon}\right|_{\varepsilon=0} \right\|_2.$$

By Cauchy-Lipschitz theorem and results on the regularity of the flow, Π is smooth. Hence, its derivatives are bounded over its compact set of definition. Hence there exists a constant $A > 0$ such that

$$\left\| \frac{\mathrm{J}^{\tilde{w}_k}_{\gamma_k}(h)}{h} - \frac{\mathrm{J}^{\tilde{w}_k}_{\tilde{\gamma}_k}(h)}{h} \right\|_2 \leq A\left(\|\tilde{\gamma} - \gamma\|_2 + \left\|\dot{\tilde{\gamma}} - \dot{\gamma}\right\|_2 \right)$$

where we can once again assume A independent of t and h. In coordinates, we use a second-order Runge-Kutta method to integrate the geodesic equation (1.5) so that the cumulated error $\|\tilde{\gamma} - \gamma\|_2 + \left\|\dot{\tilde{\gamma}} - \dot{\gamma}\right\|_2$ is of order h^2. Hence, there exists a positive constant B which does not depend on h, t or w such that

$$\left\| \frac{\mathrm{J}^{\tilde{w}_k}_{\gamma_k}(h)}{h} - \frac{\mathrm{J}^{\tilde{w}_k}_{\tilde{\gamma}_k}(h)}{h} \right\|_2 \leq Bh^2.$$

□

Numerical approximation with a single perturbed geodesic

We prove a lemma which allows to control the error we make when we approximate numerically the Jacobi field using steps (3) and (2) of the algorithm:

Lemma 3. *For all $L > 0$, there exists $A > 0$ such that for all $t \in [0,1[$, for all $h \in [0, \frac{\eta}{\|\dot{\gamma}(t)\|_g}]$ and for all $w \in T_{\gamma(t)}\mathcal{M}$ with $\|w\|_2 < L$ – in the global system of coordinates – we have*

$$\left\| \frac{\mathrm{J}^{w}_{\gamma(t)}(h) - \tilde{\mathrm{J}}^{w}_{\gamma(t)}(h)}{h} \right\|_2 \leq A(h^2 + \varepsilon h)$$

where $\tilde{\mathrm{J}}^w_{\gamma(t)}(h)$ *is the numerical approximation of* $\mathrm{J}^w_{\gamma(t)}(h)$ *computed with a single perturbed geodesic and a first-order differentiation method.*

Proof. Let $L > 0$. Let $t \in [0, 1[$, $h \in [0, \frac{\eta}{\|\dot\gamma(t)\|_g}]$ and $w \in T_{\gamma(t)}\mathcal{M}$. We split the error term into two parts

$$\left\|\frac{\mathrm{J}^w_{\gamma(t)}(h)}{h} - \frac{\tilde{\mathrm{J}}^w_{\gamma(t)}(h)}{h}\right\|_2 \le \left\|\underbrace{\frac{\mathrm{J}^w_{\gamma(t)}(h)}{h} - \frac{\mathrm{Exp}_{\gamma(t)}(h(\dot\gamma(t)+\varepsilon w)) - \mathrm{Exp}_{\gamma(t)}(h\dot\gamma(t))}{\varepsilon h}}_{(1)}\right\|_2 +$$
$$\left\|\underbrace{\frac{\mathrm{Exp}_{\gamma(t)}(h(\dot\gamma(t)+\varepsilon w)) - \mathrm{Exp}_{\gamma(t)}(h\dot\gamma(t)) - \tilde{\mathrm{Exp}}_{\gamma(t)}(h(\dot\gamma(t)+\varepsilon w)) + \tilde{\mathrm{Exp}}_{\gamma(t)}(h\dot\gamma(t))}{\varepsilon h}}_{(2)}\right\|_2$$

where Exp is the Riemannian exponential and $\tilde{\mathrm{Exp}}$ is the numerical approximation of this Riemannian exponential computed thanks to the Hamiltonian equations. When running the scheme, these computations are done in the global system of coordinates.

Term (1) Let $i \in \{1, \dots, n\}$ and let $F^i : (x, t, w) \mapsto \mathrm{Exp}[h\dot\gamma(t) + xw]^i$. We have

$$\begin{aligned}\frac{\mathrm{J}^w_{\gamma(t)}(h)^i}{h} &- \frac{\mathrm{Exp}[h(\dot\gamma(t)+\varepsilon w)]^i - \mathrm{Exp}[h\dot\gamma(t)]^i}{\varepsilon h}\\ &= \frac{1}{h}\frac{\partial F^i(\varepsilon h, t, w)}{\partial \varepsilon}\bigg|_{\varepsilon=0} - \frac{F^i(\varepsilon h, t, w) - F^i(0, t, w)}{\varepsilon h}\\ &= \frac{\partial F^i(x, t, w)}{\partial x}\bigg|_{x=0} - \frac{F^i(\varepsilon h, t, w) - F^i(0, t, w)}{\varepsilon h}.\end{aligned}$$

This is the error when performing a first-order differentiation on $x \mapsto F^i(x, t, w)$ at 0. This error is of order ϵh and will depend smoothly on t and w. Since $t \in [0, 1]$ and imposing $\|w\|_2 < L$, there exists B which does not depend on t or w such that

$$\left|\frac{\mathrm{J}^w_{\gamma(t)}(h)^i}{h} - \frac{\mathrm{Exp}[h\dot\gamma(t) + \varepsilon h w]^i - \mathrm{Exp}[h\dot\gamma(t)]^i}{\varepsilon h}\right| \le B\varepsilon h$$

so that there exists $C > 0$ such that for all t, for all h and for all w with $\|w\|_2 \le L$

$$\left\|\frac{\mathrm{J}^w_{\gamma(t)}(h)}{h} - \frac{\mathrm{Exp}[h\dot\gamma(t) + \varepsilon h w] - \mathrm{Exp}[h\dot\gamma(t)]}{\varepsilon h}\right\|_2 \le C\varepsilon h.$$

Term (2) We rewrite the Hamiltonian equation $\dot x(t) = F_1(x(t), \alpha(t))$ and $\dot\alpha(t) = F_2(x(t), \alpha(t))$. We denote $x^\varepsilon, \alpha^\varepsilon$ the solution of this equation (in the global system of coordinates) with initial conditions $x^\varepsilon(0) = x_0 = \gamma(t)$ and $\alpha^\varepsilon(0) = \alpha_0^\varepsilon = K(x_0)^{-1}(\dot\gamma(t) + \varepsilon w)$. We denote $\tilde x^\varepsilon$ the result after one step of length h of the integration of the same equation using a second-order Runge-Kutta method with parameter $\delta \in]0,1]$. The term (2) rewrites

$$\tfrac{1}{\varepsilon h}\|(x^\varepsilon(h) - x^0(h)) - (\tilde{x}^\varepsilon - \tilde{x}^0)\|_2.$$

First, we develop x^ε in the neighborhood of 0:

$$x^\varepsilon(h) = x_0 + h\dot{x}_0 + \frac{h^2}{2}\ddot{x}_0 + \int_0^h \frac{(h-t)^2}{2}\dddot{x}^\varepsilon(t)\mathrm{d}t. \tag{1.26}$$

We have for the last term

$$\left\|\int_0^h \frac{(h-t)^2}{2}\dddot{x}^\varepsilon(t)\mathrm{d}t - \int_0^h \frac{(h-t)^2}{2}\dddot{x}^0(t)\mathrm{d}t\right\|_2 = \left\|\int_0^h\int_0^{+\varepsilon} \frac{(h-t)^2}{2}\partial_\varepsilon\dddot{x}^\varepsilon(u,t)\mathrm{d}u\mathrm{d}t\right\|_2,$$

x^ε being the solution of a smooth ordinary differential equation with smoothly varying initial conditions, it is smooth in time and with respect to ε. Hence, when the initial conditions are within a compact, $\partial_\varepsilon\dddot{x}^\varepsilon$ is bounded, hence there exists $D > 0$ such that

$$\left\|\int_0^h \frac{(h-t)^2}{2}\dddot{x}^\varepsilon(t)\mathrm{d}t - \int_0^h \frac{(h-t)^2}{2}\dddot{x}^0(t)\mathrm{d}t\right\|_2 \leq Dh^3\varepsilon.$$

After computations of the first and second order terms, we get

$$\begin{aligned} x^\varepsilon(h) =& x_0 + h(\dot{\gamma}(0) + \varepsilon w) + \\ & \frac{h^2}{2}\Big((\nabla_x K)(x_0)[K(x_0)\alpha_0^\varepsilon]\alpha_0^\varepsilon + K(x_0)F_2(x_0,\alpha_0^\varepsilon)\Big) + \mathrm{O}\Big(h^3|\varepsilon|\Big). \end{aligned} \tag{1.27}$$

Now we focus on the approximation $\tilde{x}^\varepsilon$. One step of a second-order Runge Kutta method with parameter δ gives:

$$\begin{aligned} \tilde{x}^\varepsilon &= x_0 + h\Big[\Big(1 - \frac{1}{2\delta}\Big)F_1(x_0,\alpha_0^\varepsilon) + \frac{1}{2\delta}F_1\Big(x_0 + \delta hF_1(x_0,\alpha_0^\varepsilon), \alpha_0^\varepsilon + \delta hF_2(x_0,\alpha_0^\varepsilon)\Big)\Big] \\ &= x_0 + h\Big[\Big(1 - \frac{1}{2\delta}\Big)K(x_0)\alpha_0^\varepsilon + \frac{1}{2\delta}K\Big(x_0 + \delta hK(x_0)\alpha_0^\varepsilon\Big)\Big(\alpha_0^\varepsilon + \delta hF_2(x_0,\alpha_0^\varepsilon)\Big)\Big] \end{aligned}$$

We use a Taylor expansion for K:

$$\begin{aligned} K\Big(x_0 + \delta hK(x_0)\alpha_0^\varepsilon\Big) =&\ K(x_0) + \delta h(\nabla_x K)(x_0)[K(x_0)\alpha_0^\varepsilon] + \\ & \frac{(\delta h)^2}{2}(\nabla_x K)^2[K(x_0)\alpha_0^\varepsilon, K(x_0)\alpha_0^\varepsilon] + \mathrm{O}\Big(h^3\Big) \end{aligned}$$

Injecting this into the previous expression for x^ε, we get after development

$$\begin{aligned}\tilde{x}^\varepsilon &= x_0 + hK(x_0)(\alpha_0^\varepsilon)\\ &\quad + \frac{h^2}{2}\Big[K(x_0)F_2(x_0,\alpha_0^\varepsilon) + (\nabla_x K)(x_0)[K(x_0)\alpha_0^\varepsilon]\alpha_0^\varepsilon\Big]\\ &\quad + \frac{h^3\delta}{4}\Big[(\nabla_x K)(x_0)[\alpha_0^\varepsilon]F_2(x_0,\alpha_0^\varepsilon) + (\nabla_x K)^2[K(x_0)\alpha_0^\varepsilon, K(x_0)\alpha_0^\varepsilon]\alpha_0^\varepsilon\Big] + \mathrm{O}\Big(h^4\Big).\end{aligned}$$

The third order terms of $\tilde{x}^\varepsilon - x^0$ is then proportionnal to

$$\begin{aligned}&(\nabla_x K)(x_0)[\alpha_0^\varepsilon]F_2(x_0,\alpha_0^\varepsilon) - (\nabla_x K)(x_0)\alpha_0^0 F_2(x_0,\alpha_0^0)\\ &+ (\nabla_x K)^2[K(x_0)\alpha_0^\varepsilon, K(x_0)\alpha_0^\varepsilon]\alpha_0^\varepsilon - (\nabla_x K)^2[K(x_0)\alpha_0^0, K(x_0)\alpha_0^0]\alpha_0^0.\end{aligned}$$

Both these terms are the differences of smooth functions at points whose distance is of order $\varepsilon\|w\|_2$. Because those functions are smooth, and we are only interested in these majorations for points in Ω and tangent vectors in a compact ball in the tangent space, this third order term is bounded by $Eh^3\varepsilon\|w\|_2$ where E is a positive constant which does not depend on the position on the geodesic. Finally, the zeroth, first and second-order terms of x^ε and $\tilde{x}^\varepsilon$ cancel each other, so that there exists $D \geq 0$ such that

$$\|(x^\varepsilon(h) - x^0(h)) - (\tilde{x}^\varepsilon(h) - \tilde{x}^0(h))\|_2 \leq (h^3\varepsilon + Eh^3\varepsilon)$$

which concludes. □

Numerical approximation with two perturbed geodesics

We suppose here that the computation to get the Jacobi field is done using two perturbed geodesics, and a second-order differentiation as described in equation (1.8).

Lemma 4. *For all $L > 0$, there exists $A > 0$ such that for all $t \in [0,1[$, for all $h \in [0,1-t]$ and for all $w \in T_{\gamma(t)}\mathcal{M}$ with $\|w\|_2 < L$ –in the global system of coordinates – we have*

$$\left\|\frac{\mathrm{J}^w_{\gamma(t)}(h) - \tilde{\mathrm{J}}^w_{\gamma(t)}(h)}{h}\right\|_2 \leq A(h^2 + \varepsilon h),$$

where $\tilde{\mathrm{J}}^w_{\gamma(t)}(h)$ is the numerical approximation of $\mathrm{J}^w_{\gamma(t)}(h)$ computed with two perturbed geodesics and a central finite differentiation method. We consider that this approximation is computed in the global system of coordinates.

The proof is similar to the one above.

Chapter 2

Application to shape analysis [62, 10]

2.1 Introduction

The primary pathological developments of a neuro-degenerative disease such as Alzheimer's are believed to spring long before the first symptoms of cognitive decline. Subtle gradual structural alterations of the brain arise and develop along the disease course, in particular in the hippocampi regions, whose volumes are classical bio-markers in clinical trials. Among other factors, those transformations ultimately result in the decline of cognitive functions, which can be assessed through standardized tests. Being able to track and predict future structural changes in the brain is therefore key to estimate the individual stage of disease progression, to select patients and provide endpoints in clinical trials.

To this end, we propose here to predict the future shape of brain structures segmented from MRIs. We propose a methodology based on three building blocks : extrapolate from the past progression of a subject ; transfer the progression of a reference subject observed over a longer time period to new subjects ; and refine this transfer with information about the relative disease dynamics extracted from cognitive evaluations. Instead of limiting ourselves to specific features such as volumes, we propose to see each observation of a patient at a given time-point as a segmented surface mesh in a shape space.

In computational anatomy, shape spaces can be defined via the action of a group of diffeomorphisms [5, 97, 99]. In this framework, one may estimate a flow of diffeomorphisms such that a shape continuously deformed by this flow best fits repeated observations of the same subject over time, thus leading to a subject-specific spatio-temporal trajectory of shape changes [57, 81]. If the flow is geodesic in the sense of a shortest path in the group of diffeomorphisms, this problem is called geodesic regression [25, 57, 81, 28] and may be thought of as the extension to Riemannian manifolds of the linear regression concept. It is tempting then to use such regression to infer the future evolution of the shape given several past observations. To the best of our knowledge, the predictive power of such a method has not yet been extensively assessed. We will demonstrate that satisfying results can only be obtained when large numbers of data points over extensive periods of time are available, and that poor ones should be expected in the more interesting use-case scenario of a couple of observations.

In such situations, an appealing workaround would be to transfer previously acquired knowledge from another patient observed over a longer period of time. This idea requires the definition of a spatio-temporal matching method to transport the trajectory of shape changes into a different subject space. Several techniques have been proposed to register image time series of different subjects [73, 100]. They often require time series to have the same number of images, or to have correspondences between images across time series, and are therefore unfit for prognosis purposes. Parallel transport in groups of diffeomorphisms has been recently introduced to infer deformation of follow-up images from baseline matching [90, 60]. Such paradigms have been used mostly to transport spatio-temporal trajectories to the same anatomical space for hypothesis testing [83, 35]. Two main methodologies have emerged: either by parallel-transporting the time series along the baseline matching as in [28], or by parallel-transporting the baseline matching along the time series as in [87]. We evaluate both in this paper, according to their predictive power.

To compute this parallel transport operation, we use the fanning scheme described in the previous chapter, which we re-detail for the LDDMM diffeomorphisms. On the way, we use the proposed applications to control the behaviour of the numerical scheme in a very high dimensional setting.

Section 2.2 gives the theoretical background and the detailed steps of the algorithm, in the LDDMM context. Section 2.3 describes how parallel transport can be used to perform future shape prediction. Section 2.4 Section 2.5 concludes.

2.2 Parallel transport in the context of shape analysis

In this chapter, we work on a manifold of diffeomorphisms which is a particular instance of the LDDMM framework. Appendix 9 gives more details both about the theoretical and the computational aspects of this construction.

2.2.1 The chosen family of diffeomorphisms

The LDDMM-derived construction proposed in [22] provides an effective way to build a family of diffeomorphims acting on the d-dimensional ambient space $\mathbb{R}^d$. Time-varying vector fields $v_t(.)$ are generated by convolution of a Gaussian kernel $k(x,y) = \exp\left[-\frac{\|x-y\|^2}{2\sigma^2}\right]$ over n_{cp} time-varying control points $c(t) = [c_i(t)]_i$, weighted by n_{cp} associated momenta $\alpha(t) = [\alpha_i(t)]_i$, i.e. $v_t(.) = \sum_{i=1}^{n_{cp}} k\left(.\,, c_i(t)\right) \alpha_i(t)$. The set of such vector fields forms a Reproducible Kernel Hilbert Space (RKHS).

Those vector fields are then integrated along $\partial_t \phi_t(.) = v_t[\phi(.)]$ from $\phi_0 = \mathrm{Id}$ into a flow of diffeomorphisms. In [74], the authors showed that the kernel-induced distance between ϕ_0 and ϕ_1 –which can be seen as the deformation kinetic energy– is minimal i.e. the obtained flow is geodesic when the control points and momenta satisfy the Hamiltonian equations :

$$\dot{c}(t) = \mathrm{K}_{c(t)}\alpha(t), \quad \dot{\alpha}(t) = -\frac{1}{2}\ \nabla_{c(t)} \left\{\alpha(t)^T\, \mathrm{K}_{c(t)}\, \alpha(t)\right\}, \tag{2.1}$$

where $K_{c(t)}$ is the kernel matrix. A diffeomorphism is therefore fully parametrized by its initial control points c and momenta α. We denote it $\Phi_{c,\alpha}$.

Those Hamiltonian equations can be integrated with a Runge-Kutta scheme without computing the Christoffel symbols, thus avoiding the associated curse of dimensionality. The obtained diffeomorphisms then act on shapes embedded in $\mathbb{R}^d$, such as images or meshes, and we denote $\star$ this action: it is direct application point by point for meshes, and composition by ϕ_1^{-1} for images.

We now define $\mathcal{G}_c = \left\{\Phi_{c,\alpha} | c \in \mathbb{R}^{nd}, c_i \neq c_j\ \forall i \neq j, \alpha \in \mathbb{R}^{nd}\right\}$. $\mathcal{G}_c$ is the whole family of diffeomorphisms that we are considering in our experiments. Each diffeomorphism $\Phi_{c,\alpha}$ of $\mathcal{G}_c$ corresponds to a point and a co-tangent vector on the landmark manifold defined in the Appendix 9.3. This opens the way to parallel transporting $\Phi_{c,\alpha}$ along $\Phi_{c,\alpha'}$ by parallel transporting the co-tangent vector α along the geodesic γ with initial position c and initial momentum α'. At the end of the transport, we obtain a new diffeomorphism $\Phi_{\tilde{c},\tilde{\alpha}}$ that can act on the newly observed shape. This is the heuristic we use to transport observed progressions onto new subjects.

2.2.2 Parallel transport on $\mathcal{G}_c$

We are now ready to describe the fanning scheme in this particular context. This transport formally occurs on the landmark manifold.

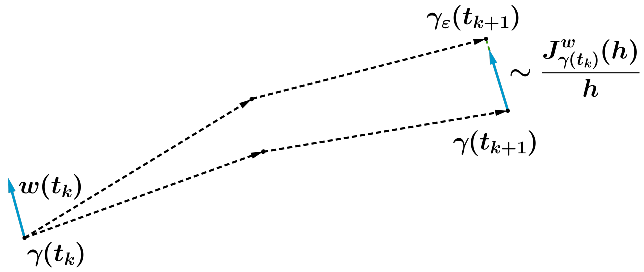


Figure 2.1: Step of the parallel transport of the vector w (blue arrow) along the geodesic γ (solid black curve). J^w_γ is computed by central finite difference with the perturbed geodesics γ_ε and $\gamma_{-\varepsilon}$, integrated with a second-order Runge-Kutta scheme (dotted black arrows). A fan of geodesics is formed.

Algorithm in this diffeomorphism case.

Divide $[0,1]$ into N intervals of length $h = \frac{1}{N}$ where $N \in \mathbb{N}$. We note ω_k the momenta of the transported diffeomorphism, c_k the control points and α_k the momenta of the geodesic γ at time $\frac{k}{N}$. Iteratively :

(i) Compute the main geodesic control points c_{k+1} and momenta α_{k+1}, using a Runge-Kutta 2 method.

(ii) Compute the control points $c_{k+1}^{\pm h}$ of the perturbed geodesics $\gamma_{\pm h}$ with initial momenta and control points $(\alpha_k \pm h\omega_k,\, c_k)$, using a Runge-Kutta 2 method.

(iii) Approximate the Jacobi field J_{k+1} by central finite difference :

$$\mathrm{J}_{k+1} = \frac{c_{k+1}^{+h} - c_{k+1}^{-h}}{2h}. \tag{2.2}$$

(iv) Compute the transported momenta $\tilde{\omega}_{k+1}$ according to equation (1.2) :

$$\mathrm{K}_{c_{k+1}} \tilde{\omega}_{k+1} = \frac{\mathrm{J}_{k+1}}{h}. \tag{2.3}$$

(v) Correct this value with $\omega_{k+1} = \beta_{k+1}\tilde{\omega}_{k+1} + \delta_{k+1}\alpha_{k+1}$, where β_{k+1} and δ_{k+1} are normalization factors ensuring the conservation of $\|\omega\|_{V_c} = \omega_k^T K_{c_k} \omega_k$ and of $\langle \alpha_k, \omega_k \rangle_{c_k} = \alpha_k^T K_{c_k} \omega_k$.

A step of the scheme is illustrated in Figure 2.1. The Jacobi field is computed with only four calls to the Hamiltonian equations. This operation scales quadratically with the dimension of the manifold, which makes this algorithm practical in high dimension, unlike Christoffel-symbol-based solutions. Step ((iv)) –solving a linear system of size n_{cp}– is the most expensive one, but remained within reasonable computational time in the investigated examples which features up to $n = 3000$ control points in dimension $d = 3$.

In the previous sections, we proved the convergence of this scheme, and showed that the error increases linearly with the size of the step used. The convergence is guaranteed as long as the step ((ii)) is performed with a method of order at least two. A first order method in step ((iii)) is also theoretically sufficient to guarantee convergence. Those variations will be studied in Subsection 2.4.2.

2.3 Method for future shape prediction

In this section, we explain the different predictive models that are used to tackle the task of shape progression prediction.

Let $(y_j)_{j=1,...,n_i}$ be a time series of segmented surface meshes for a given subject $i \in \{1, ..., N\}$, obtained at the ages $(t_j)_{j=1,...,n_i}$. Under the action of the flow of diffeomorphisms, an initial template shape T is continuously deformed and describes a trajectory in the shape space, which we denote $t \mapsto \gamma_{(c^0,\alpha^0)}(T,t) = \Phi_{c(t),\alpha(t)} \star T$. We endow the surface meshes with a varifold norm $\|\cdot\|$ which allows to measure a data attachment term between meshes without point correspondence [22].

Geodesic regression

In the spirit of linear regression, one can perform geodesic regression in the shape space by estimating the "intercept" T and the "slope" (c, α) such that $\gamma_{(c,\alpha)}(T, \cdot)$ minimizes the loss:

$$\inf_{c,\alpha,T} \sum_{j=1}^{n_i} \|\gamma_{(c,\alpha)}(T, t_j) - y_j\|^2 + R(c, \alpha) \tag{2.4}$$

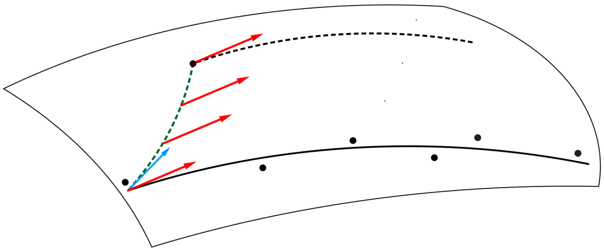
where R is a regularization term which penalizes the kinetic energy of the deformation. We estimate a solution of equation (2.4) with a Nesterov gradient descent as implemented in the software Deformetrica (`www.deformetrica.org`), where the gradient with respect to the control points, the momenta and the template is computed with a backward integration of the data attachment term along the geodesic [23]. In practice, we will fix the initial points to a regularly spaced set within a box containing the observations. This allows a to avoid one computation of inverse convolution for the parallel transport, needed when transporting a deformation along another one with different initial control points.

Once an optimum is found, we obtain a description of the progression of the brain structures which lies in the tangent space at the identity of the group of diffeomorphisms. It is natural to attempt to extrapolate from the obtained geodesic to obtain a prediction of the progression of the structures.

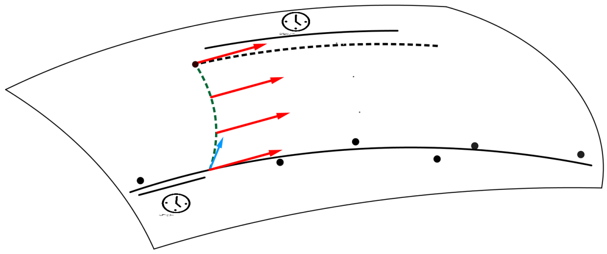
Two methods to transport spatio-temporal trajectories of shapes

As it will be demonstrated in Subsection 2.4.3, geodesic regression extrapolation produces an accurate prediction only if data over a long time span is available for the subject. This is not compatible with the goal of early prognosis.

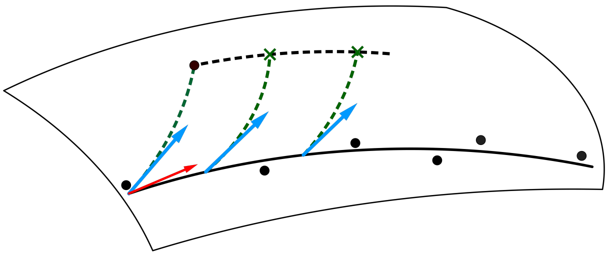
As proposed in [60, 105], given a reference geodesic, we use the Riemannian parallel transport to generate a new trajectory. We first perform a baseline matching between the reference subject and the new subject, which can be described as a vector in the tangent space of the group of diffeomorphisms. Two paradigms are available to obtain a parallel trajectory. [90] advises to transport the reference regression along the geodesic



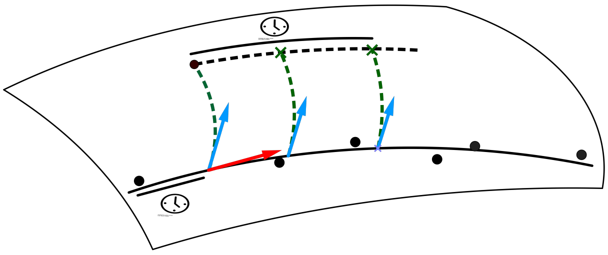
(A1) Geodesic parallelization. Blue arrow: baseline matching. Red arrows: transported regression. Black dotted line : exponentiation of the transported regression.



(A2) Reparametrized geodesic parallelization. Matching time and exp-parallel trajectory are reparametrized.



(B1) Exp-parallelization. Red arrow: geodesic regression. Blue arrows: transported baseline matching. Black dotted line : exp-parallelization of the reference geodesic for the given subject.



(B2) Reparametrized exp-parallelization. Matching time and exp-parallel trajectory are reparametrized.

which defines the matching between base and new observation and then shoot. In the shape space, this generates a geodesic starting at the baseline shape ; for this reason, we call this solution *geodesic parallelization*, and is illustrated in Figure (A1). On the other hand, [87] advocates to transport the matching vector along the reference geodesic and then build a trajectory with this transported vector from every point of the reference geodesic, as described on Figure (B1). We will call this procedure *exp-parallelization*.

To implement these parallel shifting methods, we use the algorithm described in Subsection 2.2.2.

Cognitive scores dynamics

The protocol described in the previous section has two main drawbacks. First, the choice of the matching time in the reference trajectory is arbitrary : the baseline is purely a convenience choice and ideally the matching should be performed at similar stages of the disease. Second, it does not take into account the pace of progression of the subject. In [87], the authors propose a statistical model allowing to learn, in an unsupervised manner, dynamical parameters of the subjects from ADAS-cog test results, a standardized cognitive test designed for disease progression tracking. More specifically, they suppose that each patient follows a parallel to a mean trajectory, with a time reparametrization :

$$\psi(t) = e^{\eta}(t - t_0 - \tau) + t_0 \tag{2.5}$$

which maps the subject time to a normalized time frame, where $\eta, \tau \in \mathbb{R}$. A high (resp. low) τ hence corresponds to a fast (resp. slow) progression of the scores, when a negative (resp. positive) τ corresponds to an early decay (resp. late decay) of those scores. In the data set introduced below, the acceleration factors $(e^{\eta_i})_i$ range from 0.15 to 6.01 and the time-shifts $(\tau_i)_i$ from -20.6 to 22.8 years, thus showing a tremendous variability in the individual dynamics of the disease, which must be taken into account.

With these dynamic parameters, the shape evolution can be adjusted by reparametrizing the parallel trajectory with the same formula (2.5), as illustrated on Figures (A2) and (B2).

2.4 Results

In this section, we describe the data used, we provide an analysis of the behaviour of the scheme in this high-dimensional setting and we discuss the results obtained for each of the three proposed prediction methods.

2.4.1 Data, pre-processing, parameters and performance metric

MRIs are extracted from the ADNI database, where only MCI converters (subjects who ultimately convert to Alzheimer's disease) with 7 visits or more are kept, for a total of ($N = 74$) subjects and 634 visits. Subjects are observed for a period of time ranging from 4 to 9 years (5.9 on average), with 12 visits at most. The 634 MRIs are segmented using the FreeSurfer software. The extracted brain masks are then affinely registered towards the Colin 27 Average Brain using the FSL software. The estimated transformations are finally applied to the pairs of caudates, hippocampi and putamina subcortical structures.

All diffeomorphic operations i.e. matching, geodesic regression estimation, shooting, exp-parallelization and geodesic parallelization are performed thanks to the Deformetrica software previously mentioned. A varifold distance with Gaussian kernel width of 3 mm for each structure and a deformation kernel width of 5 mm are chosen. The time discretization resolution is set to 2 months.

The chosen performance metric between two sets of meshes is the Dice coefficient, that is the sum of the volumes of the intersections of the corresponding meshes, divided by the total sum of the volumes. We only measure the volume of the intersection between corresponding structures. The Dice coefficient is comprised between 0 and 1 : it equals 1 for a perfect match, and 0 for disjoint structures.

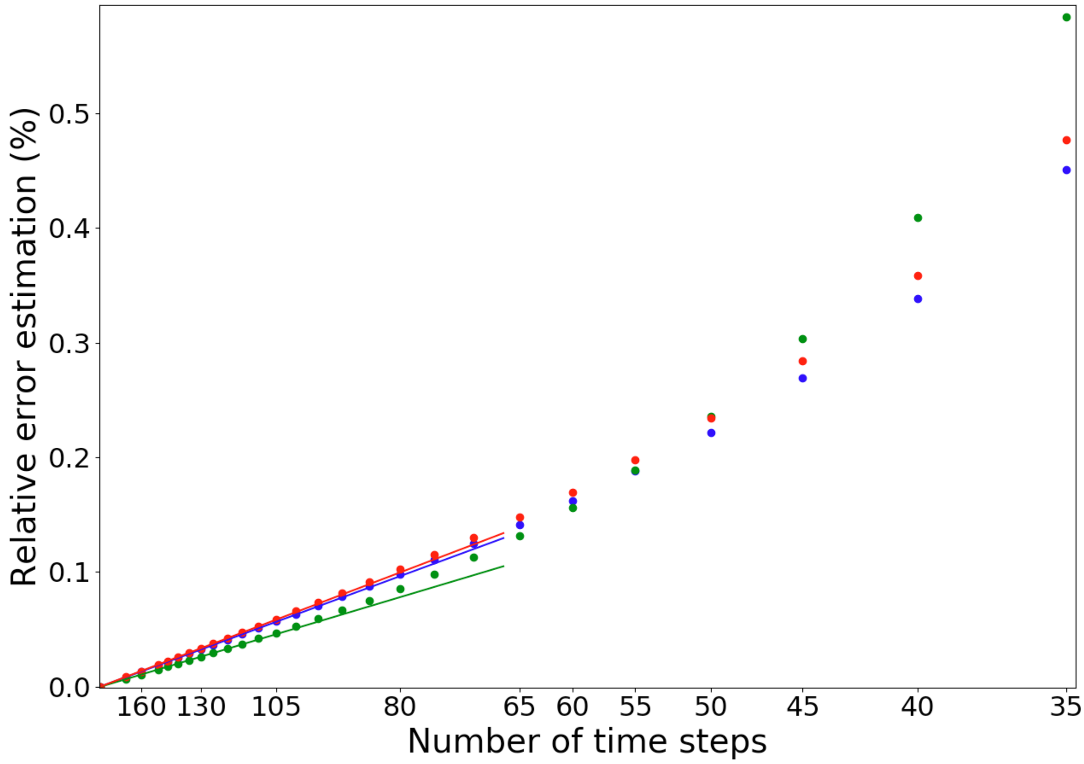


Figure 2.1: Empirical relative error of the parallel transport in a high-dimensional setting. In blue the proposed algorithm, in green the WEC variant, in red the RK4 variant.

2.4.2 Estimating the error associated to a single parallel transport

To study the error in this high-dimensional setting, we compute the parallel transport for a varying number of discretization steps N, thus obtaining increasingly accurate estimations. We then compute the empirical relative errors, taking the most accurate computation as reference.

Arbitrary reference and target subjects being chosen, Figure 2.1 gives the results for the proposed algorithm and three variations : without enforcing the conservations at step ((v)) [WEC], using a Runge-Kutta of order 4 at step ((ii)) [RK4], and using a single perturbed geodesic to compute J at step ((iii)) [SPG]. We recover a linear behavior with the length of the step $\frac{1}{N}$ in all cases. The SPG variant converges much slower, and is excluded from the following considerations.

For the other algorithms, the empirical relative error remains below 5% with 15 steps or more, and below 1% with 25 steps or more. The slopes of the asymptotic linear behaviors, estimated with the last 10 experimental measurements, range from 0.10 for the RK4 method to 0.13 for the WEC one. Finally, an iteration takes respectively -on a single CPU- 4.26, 4.24 and 8.64 seconds for the proposed algorithm, the WEC variant and the RK4 one. Therefore the initially detailed algorithm in Section 2.2 seems to achieve the best trade-off between accuracy and speed in the considered experimental setting.

2.4.3 Geodesic regression extrapolation

The acceleration factor $\exp^{\eta}$ in equation (2.5) encodes the rate of progression of each patient. Multiplying this coefficient with the actual observation window gives a notion of the absolute observation window length, in the disease time referential. Only the 22 first subjects according to this measure have been considered for this section : they are indeed expected to feature large structural alterations, making the geodesic regression procedure more accurate. The geodesic regression predictive performance is compared to a naive one consisting in leaving the last observed brain structures in the learning data set unchanged.

Table 2.1 presents the results obtained for varying learning data set and extrapolation extents. We perform a Mann-Whitney test with the null hypothesis that the observed Dice coefficients distributions obtained with the [reg] and [naive] procedure are the same. It allows us to obtain statistical significance levels to compare the 2 methods. The extrapolated meshes are satisfying only in the case where all but one data points are used to perform the geodesic regression, achieving a high Dice index and outperforming the naive one, by a small margin though and failing to reach the significance level ($p=0.25$). When the window of observation becomes narrower, the prediction accuracy decreases and becomes worse than the naive one. Indeed, the lack of robustness of the – although standard – segmentation pipeline imposes a high noise level, which seems to translate into a too low signal-to-noise ratio after extrapolation from only a few observations.

Figure 2.2 displays an extrapolated geodesic regression for a specific subject, with a large learning period of 72 months, and a prediction at 108 months from the baseline (Dice performance of 0.74 versus 0.65 with the naive approach).

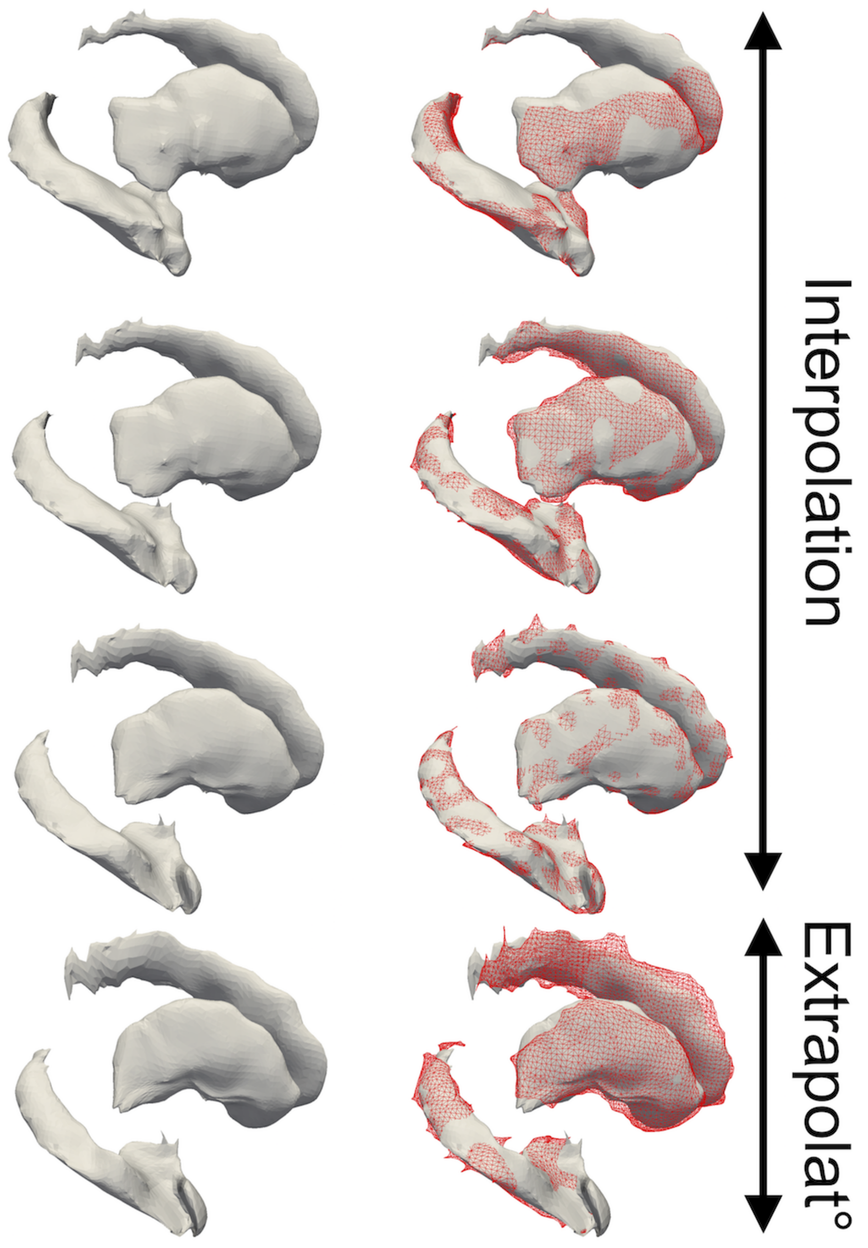


Figure 2.2: Extrapolated geodesic regression for the subject s0671. The right hippocampus, the caudate and the putamen brain structures are representated in each sub-figure. The three first rows present the interpolated brain structures, corresponding to ages 61.2, 64.2 and 67.2 (years). The last row presents the extrapolation result at age 70.2. On the right column are added the target brain structures (red wireframes), segmented from the original images.

Learning period (months)	**Method**	**Predicted follow-up visit** M12 N=22	M24 N=21		M36 N=19		M48 N=18		M72 N=16		M96 N=5	
6	[reg]	.878	.800	***	.737	***	.624	**	.509	**	.483	
	[naive]	**.888**	**.850**		**.803**		**.708**		**.626**		**.602**	
12	[reg]	-	.839	**	.769	****	.658	**	.523	**	.465	*
	[naive]	-	**.875**		**.832**		**.735**		**.644**		**.608**	
18	[reg]	-	.885		.823	*	.738		.611		.579	
	[naive]	-	**.890**		**.851**		**.764**		**.661**		**.627**	
24	[reg]	-	-		.864		.778		.681		**.657**	
	[naive]	-	-		**.869**		**.779**		**.689**		.653	
max - 1	[reg]	**.807**	*Prediction at the most remote possible time*									
~60 months	[naive]	.797	*point (~76 months) for all subjects (N=22).*									

Table 2.1: Averaged Dice performance measures between predictions and observations for varying extents of learning data sets and extrapolation. The [reg] tag indicates the regression-based prediction, and [naive] the naive one. Each row corresponds to an increasingly large learning data set, patients being observed for widening periods of time. Each column corresponds to an increasingly remote predicted visit from baseline. Significance levels [.05, .01, .001, .0001] for the Mann-Whitney test.

2.4.4 Non reparametrized transport

Among the 22 subjects whose regression-based predictive power has been evaluated in the previous section, the two which performed best are chosen as references for the rest of this paper. Their progressions are transported onto the 73 other subjects with the two different parallel shifting methods.

In more details, for each pair of reference and target subjects, the baseline target shape is first registered to the reference baseline. The reference geodesic regression is then either geodesically or exp-parallelized. Prediction performance is finally assessed : the Dice index between the prediction and the actual observation, for the two modes of transport, are computed and compared to the Dice index between the baseline meshes and the actual observation – the only available information in the absence of a predictive paradigm.

The upper part of Table 2.2 presents the results. In most cases, the obtained meshes by the proposed protocol are of lesser quality than the reference ones, according to the Dice performance metric. The two methods of transport are essentially similarly predictive, although geodesic parallelization slightly outperforms the exp-parallelization for the M12 prediction.

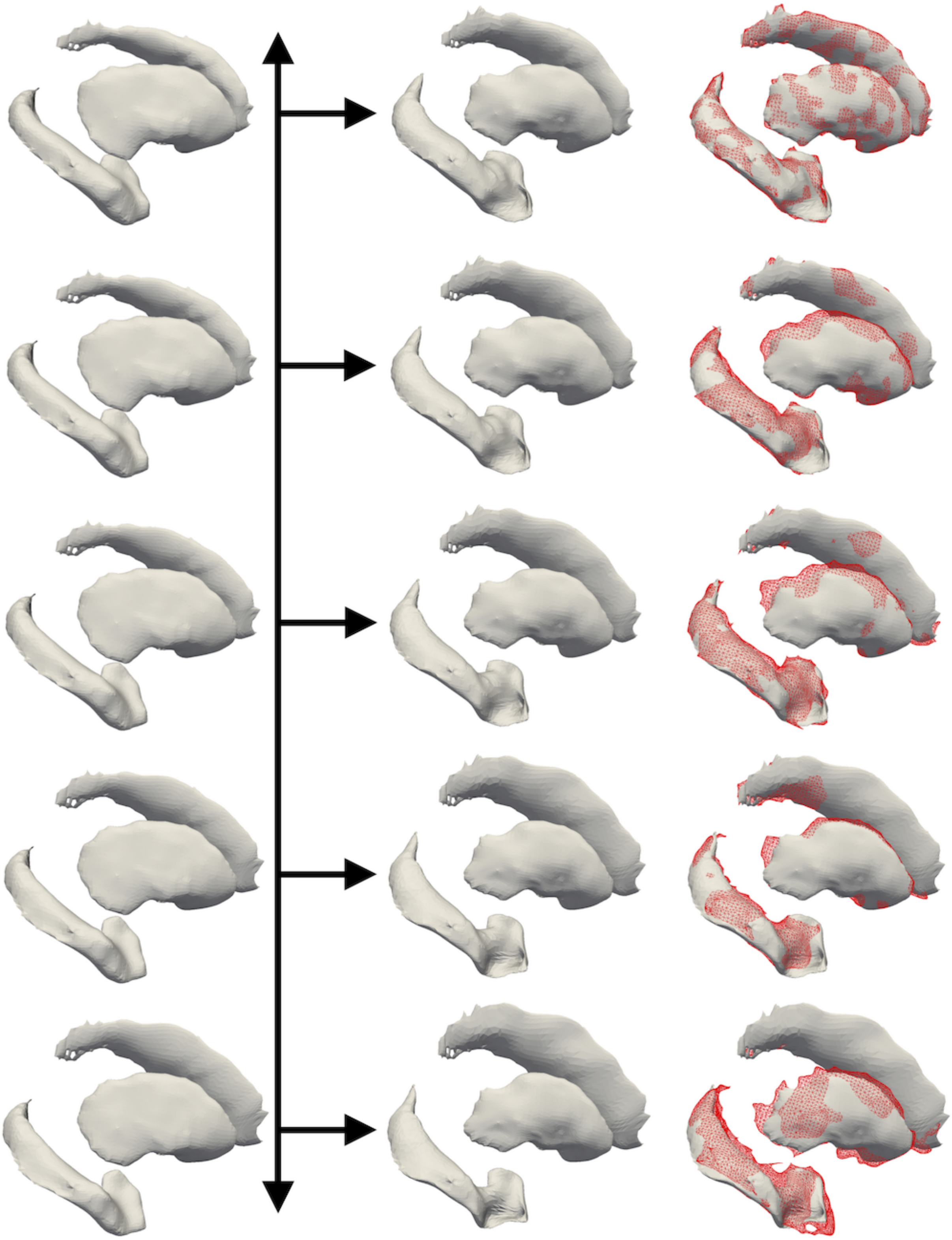


Figure 2.3: Exp-parallelization of the reference subject s0906 (first column) towards the subject s1080 (second column), giving predictions for ages 81.6, 82.6, 83.6, 84.6 and 85.6 (years). On the third column are added the target brain structures (red wireframes), segmented from the original images.

Time reparam.	**Method**	**Predicted follow-up visit** M12	M24	M36	M48	M72	M96
		N=144	N=138	N=130	N=129	N=76	N=11
Without reparam.	[exp]	* .878	.841	.799	.744	.650	.647
	[geod]	**.883**	.847 *	**.806**	.753	.664	**.661**
	[naive]	.882	**.850**	.806	**.754**	**.682**	.611
		N=140	N=134	N=123	N=113	N=62	N=17
With reparam.	[exp]	* .882	.852	.825	.796	.756	.730
	[geod]	**.888**	* **.858**	** **.831**	** ** **.802**	* ** ** **.762**	* ** * **.732** **
	[naive]	.884	.852	.809	** ** .764	** * .706	* .636

Table 2.2: Averaged Dice performance measures between predictions and obervations for two modes of transport, with or without refinement by the cognitive scores. In each cell, the first line corresponds to the exp-parallelization-based prediction [exp], the middle line to the geodesic parallelization-based one [geod], and the last line to the naive approach [naive]. Each column corresponds to an increasingly remote predicted visit from baseline. Significance levels for the Mann-Whitney test [.05, .01, .001, .0001].

2.4.5 Refining with cognitive dynamical parameters

The two reference progressions are transported through geodesic and exp-parallelization onto all remaining subjects. After time-reparametrization, the obtained parallel trajectories then deliver predictions for the brain structures. 280 parallel trajectories are obtained, delivering predictions for the brain structures.

Figure 2.4 displays a reference geodesic and an exp-parallelized curve. The predicted progression graphically matches the datapoints, and it can be noticed that the final prediction at age 85.6 (Dice 0.73) outperforms the corresponding one on Figure 2.3, obtained without time-reparametrization (Dice 0.69).

Quantitative results are presented in the lower part of Table 2.2. At the exception of the M12 prediction, both protocols outperform the naive one. The M36, M48, M72 and M96 predictions are the most impressive ones, with p-values always lesser than 1%. This shows that the pace of cognitive score evolution is well correlated with the pace of structural brain changes, and therefore allows an enhanced prediction of follow-up shapes.

No conclusion can be drawn concerning the two parallel shifting methodologies, a single weak significance result being obtained only for the M12 prediction where the geodesic parallelization method slightly outperforms the exp-parallelization one with a Dice score of 0.888 versus 0.882.

2.5 Conclusion

We detailed how to use the fanning scheme in the shape analysis context, using the LDDMM framework. Our analysis unveiled the operational qualities and computational efficiency of the scheme in high dimensions, with a empirical relative error below 1% for 25 steps only.

We then conducted a quantitative study of geodesic regression extrapolation, exhibiting its limited predictive abilities and subsequently proposed a method to transport a spatio-temporal trajectory onto a different subject space with cognitive decline-derived time reparametrization, and demonstrated its potential for prognosis. The results show how crucial the dynamics are in disease modeling, and how cross-modality data can be exploited to improve a learning algorithm. The two main paradigms that have emerged for the transport of parallel trajectories were shown to perform equally well in this prediction task. Nonetheless, the exp-parallelization offers a methodological advantage in that the generated trajectories do not depend on a particular choice of point on the reference geodesic, in contrast with the trajectories obtained by geodesic parallelization. It takes full advantage of the isometric property of the parallel transport, and eases the combination with time-warp functions based on the individual disease dynamics. In [7], the authors leverage this invariance to learn a distribution of trajectories which are all parallel to

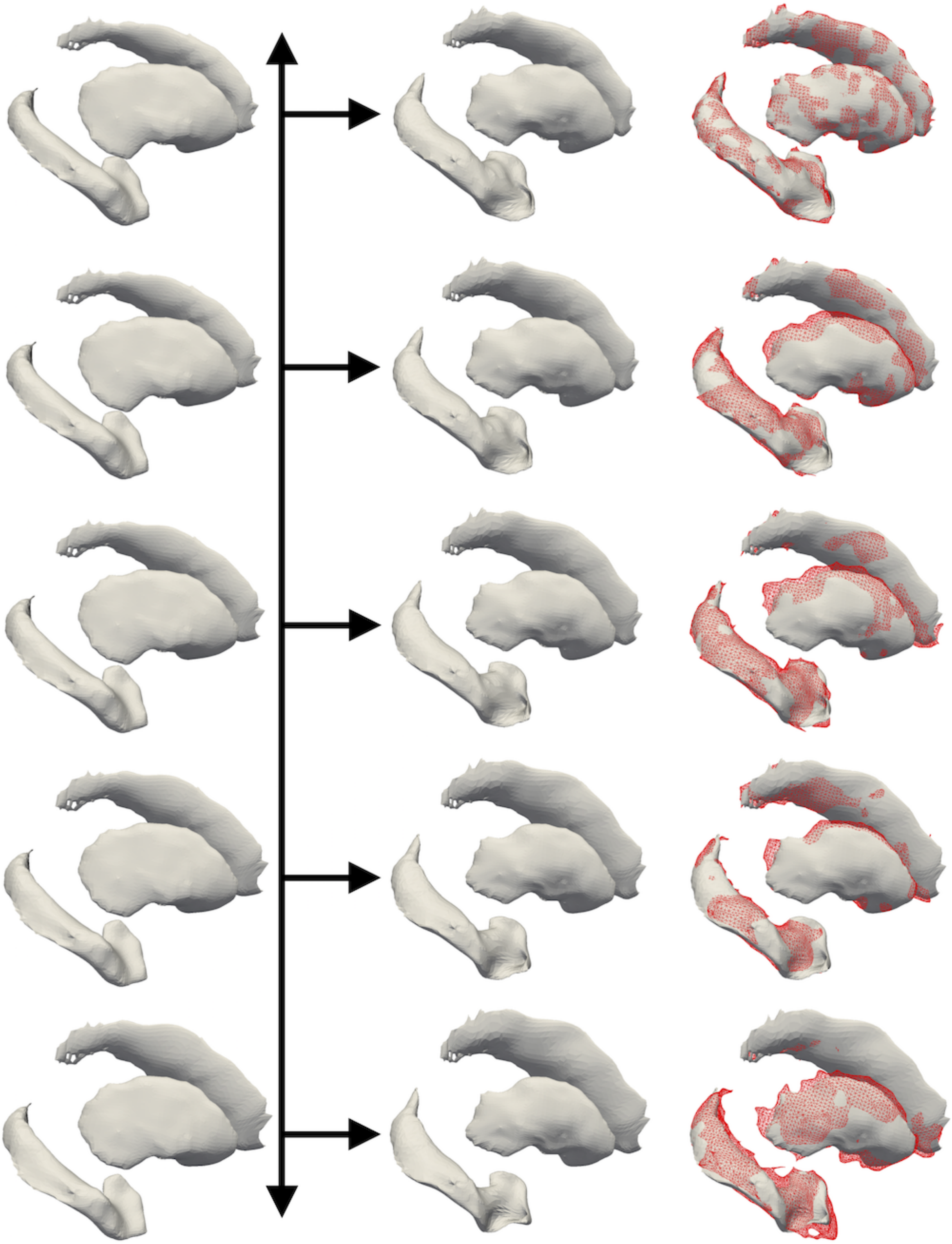


Figure 2.4: Time-reparametrized exp-parallelization of the reference subject s0906 (first column) towards the subject s1080 (second column), giving predictions for ages 81.6, 82.6, 83.6, 84.6 and 85.6 (years). On the third column are added the target brain structures (red wireframes), segmented from the original images.

a common reference geodesic, thus providing a comprehensive framework extending the preliminary approach developed here.

Part II

Geodesic Discriminant Analysis for manifold-valued data

Chapter 3

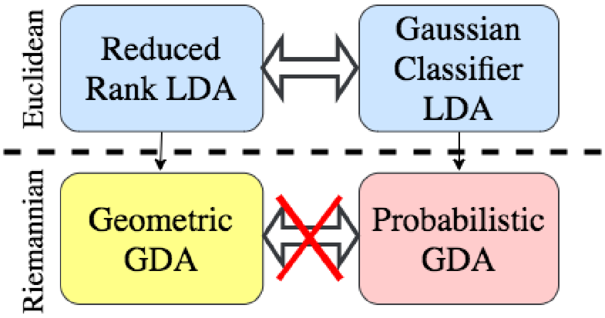
Geodesic Discriminant Analysis for manifold-valued data

This part is a reproduction of [64].

3.1 Introduction

Large quantities of high-dimensional structured data are now routinely acquired such as various types of images, videos or 2D and 3D shape data. The raw description of this kind of data does not in general reflect its intrinsic structure: it hides the generally low number of degrees of freedom that produced the observations, it is often very high-dimensional and the use of usual distances on raw data is not appropriate. To obtain a better description of the data, a standard approach is to construct or learn a low-dimensional manifold which best approximates a set of observations under a predefined criterion. If an invariance property is expected in the data, such as an invariance by rotation and scaling, it is possible to project the set of observations in the corresponding quotient space [44] or at least to build a quotiented description of the data to more classic machine learning methods as it is commonly done in scattering [69] and convolutional networks. Otherwise, the manifold structure can be learned from the data itself as it is proposed in manifold learning approaches, which project the data onto $\mathbb{R}^n$ for some small $n \in \mathbb{N}$ while trying to preserve some local or global structure observed in the high-dimensional data (see [67, 96]).

All of these approaches produce a large amount of manifold-valued data for which it is necessary to adapt usual linear machine learning approaches. The Linear Discriminant Analysis (LDA) method is a popular classification algorithm assuming a linear structure in the data. It can be formulated in two different ways. First, as a dimension reduction problem which seeks to maximize the between-class variance with respect to the within-class variance. Second, LDA can be formulated as a classification problem supposing each class is distributed as a Gaussian random variable with common covariance matrix. In this paper, we propose generalizations of those two formulations of LDA to manifold-valued data. So far, most classifications of manifold-valued data was done after having projected the data onto a common tangent space (see [23, 54]), or using a coordinate chart on the



manifold. Both of these approaches only see a simplified version of the geometry of the manifold. The proposed generalizations of LDA address this by taking into account the intrinsic geometry of the data to perform dimension reduction and classification.

The first generalization, derived in Section 3.2, that we call geometric Geodesic Discriminant Analysis (geometric GDA), is obtained by rewriting the Fisher Discriminant Ratio (FDR) –which measures the ratio between the between-class variance and the within-class variance– using geodesic distances on the manifold. We propose to build a geodesic subspace on the manifold on which this criterion is maximized.

Because the optimization of the criterion formulated for the geometric GDA is not always tractable, we proceed with a second generalization of LDA. Derived in Section 3.3, it extends the restricted Gaussian Classifier formulation of LDA. To extend this formulation to manifolds, we model the classes distributions as Riemannian exponentials of Gaussian distributions defined on a tangent space at a specific common point. We then propose to optimize the point on which this construction is centered at and to use convenient descriptions of the between-class covariance and the within-class covariance, basing our work on [82]. We call this method probabilistic GDA. We will show how to make this approach computationally efficient for a wide variety of manifolds.

It has been shown in [36] that, in the linear case, this approach is equivalent to the reduced rank LDA: it produces the same dimension reduction and classification rule. In the nonlinear case, probabilistic GDA and geometric GDA will not be equivalent.

A particular case of manifold structure can be obtained under the action of a group of diffeomorphisms on a set of shapes. Our formulation of the Large Deformation Diffeomorphic Metric Mapping (LDDMM) provides a way to parametrize a finite-dimensional manifold of diffeomorphisms. This family of diffeomorphisms then allows the comparison of shapes on which they act. We introduce this framework in Section 3.4. In Section 3.5 we provide results of the algorithm on 2D shapes extracted from the kimia-216 dataset as well as on 3D Brain structures segmented from the ADNI dataset. The probabilistic GDA is however generic and efficient enough to be applicable to a much broader family of manifolds.

Among related work, Exact Principal Geodesic Analysis (Exact PGA), initially formulated in [27], proposes to minimize the unexplained variance -measured using geodesic distances- after projection of the data onto a geodesic subspace. Several variations have been proposed, among which Bayesian PGA [110] or Horizontal Component Analysis [91]. Our work differs from these methods since we propose a supervised learning algorithm which optimizes class separation, not explained variance, to increase classification performances.

We summarize our contributions:

1. We propose geometric GDA, a generalization of the reduced rank definition of LDA to manifold-valued data.

2. We propose probabilistic GDA, a generalization of the restricted Gaussian classifier definition of LDA to manifold-valued data.

3. We illustrate the geometric GDA method on $\mathbb{S}^2$ with synthetic data and the probabilistic GDA model on the kimia-216 dataset and on a dataset of hippocampi extracted from magnetic resonance images (MRI).

3.2 Geometric Geodesic Discriminant Analysis

In this section, we introduce geometric GDA, a generalization of LDA to manifold-valued data using Fisher's approach to LDA [26]. In this paper, we consider a set of labelled observations $(y_i)_{i=1,...,N} \in \mathcal{M}$ from $C > 0$ different classes, where $\mathcal{M}$ is a smooth Riemannian manifold that we assume geodesically complete. For $p, q \in \mathcal{M}$, we note $d(p, q)$ the geodesic distance between p and q, and $s \in \mathbb{N}$ the dimension of the manifold.

If the manifold is a vector space, reduced rank LDA [26] consists in projecting the observations onto a linear subspace on which the between-class variance is maximized with respect to the within-class variance. Fisher proposed to find unit vectors a via maximization of the Fisher Discriminant Ratio (FDR):

$$0 < \frac{a^\top B a}{a^\top W a} \tag{3.1}$$

where $^\top$ denotes transposition, B is the between-class covariance matrix –the covariance matrix of the class centroids– and W is the within-class covariance matrix. To provide an expression generalizable to manifolds, we rewrite this FDR:

$$\frac{\dfrac{1}{C-1}\displaystyle\sum_{c=1}^{C}(a^\top \mu - a^\top \mu_c)^2}{\dfrac{1}{N-C}\displaystyle\sum_{c=1}^{C}\sum_{i\in I_c}(a^\top \mu_c - a^\top x_i)^2} \tag{3.2}$$

where μ is the mean of the observations, for each $c \in \{1, ..., C\}$, μ_c is the empirical mean of the class c and I_c is the set of indices of the observations of the class c. For any observation x, $a^\top x$ may be interpreted as the projection of the observation onto the space spanned by a.

If the manifold is non flat, instead of constructing a linear subspace, we will build a geodesic subspace on the manifold, as proposed for PGA in [27]. For any $m \in \mathcal{M}$, for any subspace $V \subset T_m\mathcal{M}$, we define the geodesic subspace $\mathrm{Exp}_m V = \{\mathrm{Exp}_m(v) | v \in V\}$ where $\mathrm{Exp}_m : T_m\mathcal{M} \to \mathcal{M}$ is the Riemannian exponential at m. We define a projection operator on a subspace S of $\mathcal{M}$ by $\pi_S(x) = \mathrm{argmin}_{y \in S} \mathrm{d}(x, y)^2$ when it is correctly defined. This projection, defined by minimization, might be ill-defined unless we restrict ourselves to a neighborhood of m. Assuming it is well-defined, equation (3.2) can now be generalized to manifolds by using geodesic distances measured after projection on S:

$$\frac{\frac{1}{C-1}\sum_{c=1}^{C} \mathrm{d}(\pi_{\mathrm{Exp}_m(V)}(\mu), \pi_{\mathrm{Exp}_m(V)}(\mu_c))^2}{\frac{1}{N-C}\sum_{c=1}^{C}\sum_{i \in I_c} \mathrm{d}(\pi_{\mathrm{Exp}_m(V)}(\mu_c), \pi_{\mathrm{Exp}_m(V)}(x_i))^2}. \tag{3.3}$$

where μ (resp. μ_c) are the Fréchet means [43] of the observations (resp. of the observations of class c). Reduced rank LDA on the manifold becomes the problem of maximizing this with respect to $m \in \mathcal{M}$ and V linear subspace of $T_m\mathcal{M}$. V can be constructed in a forward fashion by a basis $\{v_1, ..., v_k\}$ where k is a chosen number of component. Note that an alternative generalization could propose to recompute the Fréchet means after the projection, which yields a criterion different than equation (3.3) since the Fréchet mean of the projection is in general not the projection of the Fréchet mean. This alternative generalization of equation (3.1) would be more expensive to compute, since the computation of the different Fréchet means of the projections would be required at every step of the optimization procedure.

3.2.1 Inference

In practice, it is hard to find a robust procedure which optimizes both m and V at the same time. We propose a greedy procedure: we first optimize jointly m and a first geodesic component $v_1 \in T_m\mathcal{M}$, and then add new components v_k one at a time. In the linear case, this procedure yields the exact same optimum. A theoretical discussion about the validity of this procedure in the nonlinear case will be part of further work.

Note that if closed-form expressions are available for Riemannian logarithms and exponentials, the proposed geometric GDA can be computed efficiently. This is the case for Kendall shape space, the sphere or the manifold of symmetric positive-definite matrices with affine-invariant metric for instance. We will provide results in the case of the sphere $\mathbb{S}^2$ in Section 3.5.1.

Note also that the optimization problem (3.3) might be ill-defined if some degeneration is observed in the data, for instance if all the data points lie on a single geodesic, as in the linear case when the between-class covariance matrix does not have full rank. The study of the conditions for this estimation procedure to be well-defined will not be conducted in this chapter.

3.2.2 Dimension reduction and classification

After estimation of m and V, one can project the observations onto V by taking the coordinates of $\pi_{\mathrm{Exp}_m(V)}(y)$ for each observation y. This gives a low-dimensional representation of the data, on a space in which classes differences are most pronounced. This representation has the same range of applications as dimension reduction with linear LDA.

Classification can then be done in one of two ways. First, directly in the low-dimensional space $\mathrm{Exp}_m(V)$ by comparison of test observations geodesic distances to the different classes centroids on $\mathrm{Exp}_m(V)$, in a fashion very similar to the classic LDA. Second, it can be done after projection of the data onto $V \subset T_m\mathcal{M}$ using any usual classifier, whose performances will in general be improved if the FDR (3.3) has been correctly optimized.

Unfortunately, when no closed-form expressions are available for Riemannian logarithms or exponentials, geometric GDA is intractable. To remedy this, we propose a generalization of the alternative formulation of LDA.

3.3 Probabilistic Geodesic Discriminant Analysis

In the linear case, the restricted Gaussian classifier formulation of LDA assumes each class is distributed along a normal distribution, with common covariance Σ. In this linear setting, the probability of an observation y, if it is of class c, is:

$$y|c = c \sim \mathcal{N}(y|\mu_c, \Sigma). \tag{3.4}$$

In [36], the authors show that maximimizing the likelihood of this model with a rank constraint on the means μ_c ($\mathrm{rank}(\mu_c)_{c=1,\ldots,C} < K$) is equivalent to projecting the observations onto the K first discriminant components found by maximization of the Fisher Discriminant Ratio (3.1), even when the within-class covariance matrix Σ is unknown.

There is no natural way to generalize equation (3.4) to manifold-valued data. In particular, it is hard to make sense of the homoscedasticity hypothesis in LDA since it involves comparing covariance matrices defined at different tangent spaces on the manifold. One possible generalization of equation (3.4) is to consider:

$$y|c = c \sim \mathrm{Exp}_m(d_c + \alpha) \tag{3.5}$$

where m is a point on the manifold and $\alpha \sim \mathcal{N}(0, \Sigma)$ is a normal distribution on the tangent space $T_m\mathcal{M}$. If the manifold is flat, this model is equivalent to (3.4), and the rank constraint can in theory be enforced on the vectors $d_c \in T_m\mathcal{M}$. The homoscedasticity is replaced with the assumption that, as seen from the tangent space to m, the logarithms of the observations of the different classes are distributed along normal distributions with the same covariance matrix Σ. This approach is still hardly tractable in practice. First, learning the model will require estimating the full within-class covariance matrix Σ. Second, the rank constraint is difficult to implement in practice. We therefore extend the model defined in [82], which is similar to LDA, to:

$$y_i | c \sim \mathcal{N}(\mathrm{Exp}_m(F\alpha_c + G\beta_i), \sigma) \tag{3.6}$$

where:

- $\mathcal{N}$ is a normal isotropic distribution on $\mathcal{M}$ with density $p(y, \mu, \sigma) = \frac{1}{D(\mu,\sigma)} e^{-\frac{1}{2\sigma} d^2(y,\mu)}$, as defined in [27]. It can also be taken to be a normal distribution on a larger space of observations, to ease computations, as used in the applications below,
- F is a s times $C-1$ matrix which can be seen as the between-class covariance matrix,
- G is a s times N_G matrix where $N_G \in \mathbb{N}$ is the selected number of intra-class components to estimate: it can be seen as the principal components of the within-class variations, as seen from $T_m\mathcal{M}$,
- For each class c, α_c in $\mathbb{R}^{C-1}$ contains the coordinates of the class c in the $C-1$-dimensional space represented in F,
- β_i in $\mathbb{R}^{N_G}$ is a hidden variable which contains the coordinates of the i-th observation within its class, in the N_G-dimensional space represented in G.

We put normal priors on α and β, and an automatic relevance determination prior on G as in [68]:

$$P(G;\gamma) = \prod_{i=1}^{N_G} \left(\frac{\gamma_i}{2\pi}\right)^{\frac{s}{2}} \exp\left(-\frac{\gamma_i}{2}\|G_i\|_2^2\right) \tag{3.7}$$

where $(\gamma_i)_{i=1,\dots,N_G}$ is a set of parameters which are estimated during the learning procedure and $(G_i)_{i=1,\dots,N_G}$ are the columns of G. This prior allows the automatic selection of a relevant number of dimensions in the within-class covariance structure. An alternative would be to use methods similar to [18, 79] which iteratively add dimensions to the optimized subspace.

Compared to the tangent LDA, which consists in performing an LDA after having projected the observations onto the tangent space to the Fréchet mean, the proposed method updates the within and between-class components with a constant feedback from the real geometry of the data. Besides, we allow the joint optimization of the point m and do not constrain it to be the Fréchet mean of the data, which may not be optimal in the perspective of class separation (in [39] the authors show it is not optimal in the case of exact PGA).

3.3.1 Inference

As in [110], the mode of the posterior distribution of the variables α and the optimal values of the parameters can be obtained as a maximum a posteriori using a gradient descent. In more details, we maximize $P(y_j|\theta, \beta)P(\theta)P(\beta)$ with respect to the parameters $\theta = (F, G, \alpha, m, \sigma, \gamma)$ and β. The computation of the gradient requires the differentiation of a function of a geodesic endpoint with respect to its initial conditions. It can be done by backward integration using the method described in [93].

This approach is tractable in a wide variety of situations:

- Even if there is no closed-form expression for Riemannian exponential, geodesics can still be computed through integration of the Hamiltonian system of equations, using only the inverse of the metric and its gradient, as shown in [20]. In that case, automatic differentiation is a competitive way to compute the gradients, as shown in [50].

- The normal distribution in equation (3.6) can be replaced with a normal distribution on a larger space which contains the observations e.g.a pixel-wise normal distribution for images, or a noise in $\mathbb{R}^3$ for $\mathbb{S}^2$. This saves the computation of the normalization constant of the Riemannian normal distribution and of geodesic distances. Since this distribution is used only to measure residuals, we believe it has a limited effect on the model.

- The estimation procedure can be parallelized among the different subjects, rendering it efficient even with large data sets.

3.3.2 Dimension reduction and classification

After estimation of the parameters of the model, it is possible to project an observation y onto the geodesic subspace defined by F by optimization of:

$$\delta \mapsto d(\mathrm{Exp}_m(F\delta), y)^2 \tag{3.8}$$

with respect to $\delta \in \mathbb{R}^{N_c - 1}$, which indicates the position of the observation y in the geodesic subspace $\mathrm{Exp}_m(F)$. Doing so yields a low-dimensional description of each data point. Classification can be performed after dimension reduction of the dataset. We will show results of this classification procedure in Section 3.5.

Classification can also be done using the probabilistic GDA model, by maximizing the likelihood of an observation with respect to the classes. For each unobserved y, using Bayes rule:

$$p(c = c_k|y) = \int p(y|c = c_k, \beta_k = \beta)p(\beta)p(c = c_k)\mathrm{d}\beta. \tag{3.9}$$

The integral over the hidden variable β corresponds to looking at the observation J through all its possible representations as an object of class c_k, where the representations have been learned through the matrix G. This integral is expensive to compute or approximate in most cases and we decide to settle for the mode:

$$p(c = c_k|y) \propto p(y|c = c_k, \beta^*)p(c = c_k) \tag{3.10}$$

where $\beta^* = \mathrm{argmax}_\beta p(y|c = c_k, \beta)$, which can be estimated via gradient descent.

The ability to compute the integral (3.9) would allow to evaluate the new observation as an element in the space quotiented by the different representations of the elements of the class c_k. Additionally, as described in [82], it would also allow to do one-shot learning i.e. being able to decide if a new observation is in the set of known classes or if it is more likely to belong to a yet unobserved class.

3.4 Probabilistic GDA for shape analysis.

The Probabilistic GDA introduced above can be applied in a variety of situations, we will focus on examples of applications in the case of shapes modelled using the LDDMM framework [75, 107]. We first introduce this framework, before rewriting the model (3.6) in this particular case.

3.4.1 Embedding shapes and images on a manifold

The LDDMM framework provides a way to compare shapes via the action of diffeomorphisms of the ambient space. Such diffeomorphisms are obtained by integration of the flow of a square integrable time-varying vector field. The parametrization of the diffeomorphisms then amounts to the parametrization of time-varying vector fields. In our approach, we use as in [23] a sparse description of vector fields:

$$X(x) = \sum_{i=1}^{p} k(x, q_i)p_i \tag{3.11}$$

where $p \in \mathbb{N}$ is fixed, $(q_i)_{i=1,\ldots,p}$ is a set of control points, $(p_i)_{i=1,\ldots,p}$ is a set of momenta and k is a Gaussian kernel of fixed width ρ. The space of such vector fields is a Reproducible

Kernel Hilbert Space (RKHS) K with

$$\langle X, X' \rangle_K = \sum_{i,j=1}^{p} k(q_i, q'_j) p_i^\top p'_j. \tag{3.12}$$

Given an initial vector field X of this form, one can show [23] that there is a unique time-varying vector field $X(t, \cdot)$ such that $X(0, \cdot) = X$ which minimizes $\int_0^1 \|X(t, \cdot)\|_K^2$. We call this the geodesic flow of the initial vector field. Considering only such geodesics, we get a parametrization of diffeomorphisms solely determined by the initial set of control points and momenta. Strictly speaking, the obtained set of diffeomophisms is not a Riemannian manifold. But when the set of initial control points c is fixed, we dispose of a mapping which enables us to describe each diffeomorphism by a tangent vector to the landmark manifold at c. We simply adapt the previously described GDA to this manifold and map back the results onto the diffeomorphism space. We denote $\Phi_{q,p} \cdot M$ the action of the diffeomorphism $\Phi_{q,p}$ parametrized by the initial control points and momenta q, p on the shape M. If M is a mesh embedded in $\mathbb{R}^n$, then $\Phi_{q,p}$ acts on the points of the meshes directly. If M is an image, $\Phi_{q,p} \cdot M = M \circ \Phi_{q,p}^{-1}$ where M is seen as an element of $L_2(\mathbb{R}^n; \mathbb{R})$ for some integer n.

3.4.2 A generative model

Let us assume that we have a collection of shapes $(y_k)_{i=1,\ldots,N}$ where $N \in \mathbb{N}$. We note n the dimension of the ambient space and p the number of control points. As described in equation (3.6), we assume that each shape y_k was generated with probability:

$$\frac{1}{(2\pi)^{\frac{\Lambda}{2}} \sigma^\Lambda} \exp\Big(-\frac{1}{2\sigma^2} \|\Phi_{q,F\alpha^{c_k}+G\beta_k} \cdot M - y_k\|_\Lambda^2\Big) \tag{3.13}$$

where:

- $\Lambda \in \mathbb{N}$ is the dimension of the observations *e.g.* number of voxels for the images, number of faces for varifolds. We embed those observed shapes in a Λ-dimensional space on which we define a norm $\| \cdot \|_\Lambda$ (L^2 for images, varifold norm for meshes as in [34]),
- M is a template shape,
- $\Phi_{q,F\alpha^{c_k}+G\beta_k}$ is the diffeomorphism obtained with the initial momenta $p_k = F\alpha^{c_k} + G\beta_k$ and control points q.

Note that we replaced the normal distribution on the manifold by a normal distribution on the set of shapes, that can be defined for images, varifolds or currents as shown in [34]. There are two reasons for this. First, the orbit of M under the action of the group of diffeomorphisms does not in general contain the observations, the idea being to describe shape variability with strong smoothing constraints on the shape structures. Second, even if we could use geodesic distances on the manifold of diffeomorphisms, the computation of this geodesic distance would be too expensive in general to make the inference of the model tractable.

For the inference, we estimate the mode of the logarithm of the posterior distribution, which writes, using Bayes rules and assuming that F, G, α and β are independent:

$$\begin{aligned} l(\theta) = \log(P(F, G, \alpha, \beta | y_k; \gamma, I, \sigma)) = -\Lambda \log(\sigma) \\ - \frac{1}{2\sigma^2} \|\Phi_{q, F\alpha^{c_k} + G\beta_k} \cdot M - y_k\|_\Lambda^2 - \frac{1}{2}\|\beta\|_2^2 \\ \sum_{i=1}^{N_G} \frac{\Lambda pn}{2} \log(\frac{\gamma_i}{2\pi}) - \sum_{i=1}^{N_G} \frac{\gamma_i}{2}\|G_i\|_2^2 - \frac{1}{2}\|\alpha\|_2^2 \end{aligned} \tag{3.14}$$

Derivating (3.14) yields the closed-form updates for σ and γ:

$$\gamma_i = \frac{\Lambda pn}{\|G_i\|_2^2}. \tag{3.15}$$

$$\sigma^2 = \frac{\sum_{k=1}^{N} \|(\Phi_{F\alpha^{c_k} + G\beta_k}) \cdot M - y_k\|_\Lambda^2}{\Lambda N} \tag{3.16}$$

The computation of the gradients with respect to the momenta p, the control points q and the template M can be done by backward integration of system of adjoint equations as detailed in [23, 110, 93] and propagated to α, β, F and G using the chain rule. The optimized functional is once again not convex in general. Algorithm 1 gives a pseudo-code for the estimation procedure. A complete code of the model is made available [1].

Algorithm 1 Probabilistic GDA inference on shapes

$F, G, \alpha, \beta, M, q \leftarrow$ Initialization from Tangent LDA
$\gamma, \sigma \leftarrow$ (3.15)(3.16): for initialization.
while no convergence **do**
    Compute $l(\theta)$
    Compute $\nabla_M l(\theta)$, $\nabla_p l(\theta)$, $\nabla_q l(\theta)$.
    Propagate to $\nabla_F l(\theta)$, $\nabla_G l(\theta)$, $\nabla_\alpha l(\theta)$, $\nabla_\beta l(\theta)$
    Update $(F, G, \alpha, \beta, M, q)$ by line search.
    $\gamma, \sigma \leftarrow$ (3.15)(3.16): closed-form update.
return $F, G, \alpha, \beta, M, q, \sigma, \gamma$

[1] A code for the model is available at www.deformetrica.org

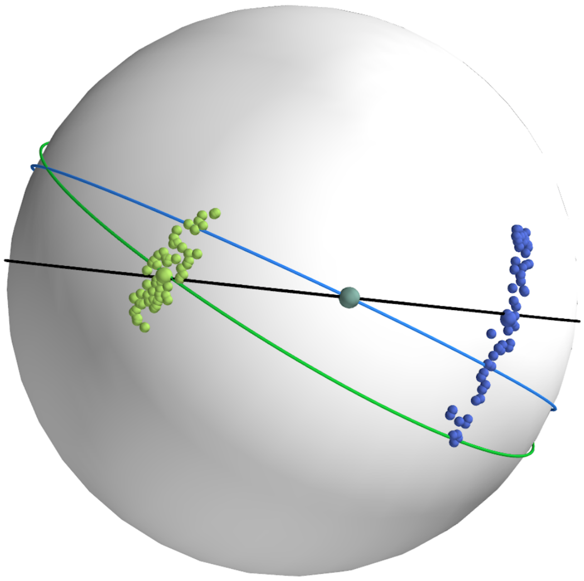


Figure 3.1: The points are labelled data. The black geodesic is the first component of the PGA method. The blue geodesic is obtained by geometric GDA with m set to the Fréchet mean. The green geodesic is obtained by geometric GDA with optimization of m.

3.5 Applications and Results

3.5.1 Geometric GDA on $\mathbb{S}^2$

We performed the optimization of the criterion given in equation (3.3) in the case of the sphere $\mathbb{S}^2$ with the metric induced from $\mathbb{R}^3$, on a synthetic set of points of two classes. Note that, whether we optimize the position of the point m on which the geodesic subspace is built or not, the optimization problem is in general not convex. We therefore perform multiple gradient descents with randomly chosen initial conditions, and select the final estimated values which give the optimum of the Fisher Discriminant Ratio. We compare three methods: an LDA performed in the tangent space to the Fréchet mean, a geometric GDA performed with a geodesic subspace set to the Fréchet mean (GDA) and a geometric GDA performed with the joint estimation of the geodesic subspace and of the point on which it is built (full GDA).

Figure 3.1 shows the estimated geodesics in the different cases of GDA, as well as the result of an exact PGA built by optimization of the explained variance on a geodesic subspace at the Fréchet mean. In each case, we measure the FDR after projection onto the first component found after optimization. Note that the FDR measured after projection assuming a linear structure differs from the FDR defined in equation (3.3). Indeed, the projection of the classes centroids is in general different from the class centroids of the projections, unlike in the linear case. We provide the values of the FDRs measured after projection and the FDRs measured in equation (3.3) in Table 3.1.

The geometric GDA outperforms an LDA performed in the tangent space to the

Method	Tangent LDA	GDA	full GDA
FDR of projection	495	514	647
FDR equation (3.3)	x	505	636

Table 3.1: Fisher Discriminant Ratios. Higher FDRs indicate a better class separation.

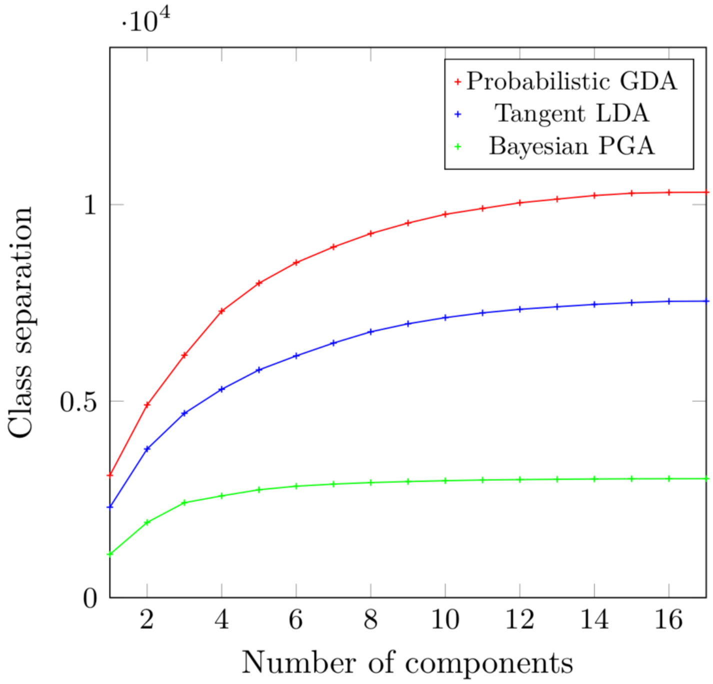


Figure 3.2: Class separation for each number of selected components, on the kimia-216 dataset.

Fréchet mean in terms of class separation, indicating that we may obtain better classification results in some situations. In addition, as mentioned in Section 3.2, the optimization of m in equation (3.3) allows a significant improvement.

3.5.2 Kimia-216

We used the setting described in Section 3.4 on shapes from the kimia-216 dataset. The kimia-216 dataset consists of 18 classes each containing 12 observations. We extracted the contour of the shapes on the images and modelled them as varifolds with a Gaussian kernel of width set at 13 (expressed in pixels of the original images). The number of points of the obtained shapes is not controlled and vary between 300 and 800. For each class, we randomly selected 9 observations that we rigidly aligned one to another. We proceeded to the estimation of the model described in equation (3.6), simultaneously estimating the matrices F and G, the vectors α for each class, the mode of β for each observation, as well as the template shape I and the set of control points. We set the kernel width ρ of the diffeomorphisms to 10 (in terms of pixels of the original images).

After estimation of the parameters of the model, we projected each training observation using the method described in Section 3.3.2. To measure the quality of the projection, we compute, in the low-dimensional space, the between-class covariance matrix B and the within class-covariance matrix W and compute the eigenvalues of $W^{-1}B$. Those eigenvalues measure the separation of the classes, and are equivalent to the FDR in the linear case. We compute those eigenvalues for the probabilistic GDA, the tangent LDA and the bayesian PGA with 17 components. Note that the Bayesian PGA is a special case of the model 3.6 when there is a single class. The results are provided in Figure 3.2. The probabilistic GDA outperforms both the tangent LDA and the Bayesian PGA in the separation it provides after projection of the data.

Then, we provide a plot of the two first components of each observations found using the probabilistic GDA, for visualization purposes, to be compared with the same components for tangent LDA and bayesian PGA, on Figure 3.3.

Finally, we investigated classification performances on the kimia-216 database, using the classification procedure described in Section 3.3.2. In details, for each test observation and each candidate class k, we evaluated the mode of the integral (3.10) by rigidly aligning the test observation to an element of the class k and performing a gradient descent on β with α set to α_k, the position of the class k in the space spanned by F. We take the class which gives the smallest residual after the descent. A 3-fold result of this classification procedure gives an average accuracy of 89%. Note that such low classification results compared to usual benchmarks [54] can be expected since this dataset is not well adapted to deformable models: the differences between the shapes occur both on small and large scales.

3.5.3 Brain structures in the course of Alzheimer's disease

From MRI images in the ADNI dataset, we segmented hippocampi from 125 normal controls and MCIc subjects (subjects who have or will convert to Alzheimer's disease) using Freesurfer [17]. We then ran the probabilistic GDA model four times on randomly extracted training set and test set in the data. Each run provided an estimation of a single discriminant geodesic component. Figure 3.4 shows an example of discriminant geodesic component.

After estimation on the training set, we projected both testing and training set onto the first geodesic component by optimization of (3.8). We then trained a logistic regression classifier on the projected, 1-dimensional, data. This is common practice after LDA dimension reduction, to learn the appropriate threshold for the classification and to correct for the strict homoscedasticity hypothesis of the model.

The AUC and accuracy scores are available in Table 3.2 and compared to a classification based on the hippocampi volumes on the exact same folds, as well as to other reference methods performing cross-sectional hippocampus-based classification of normal

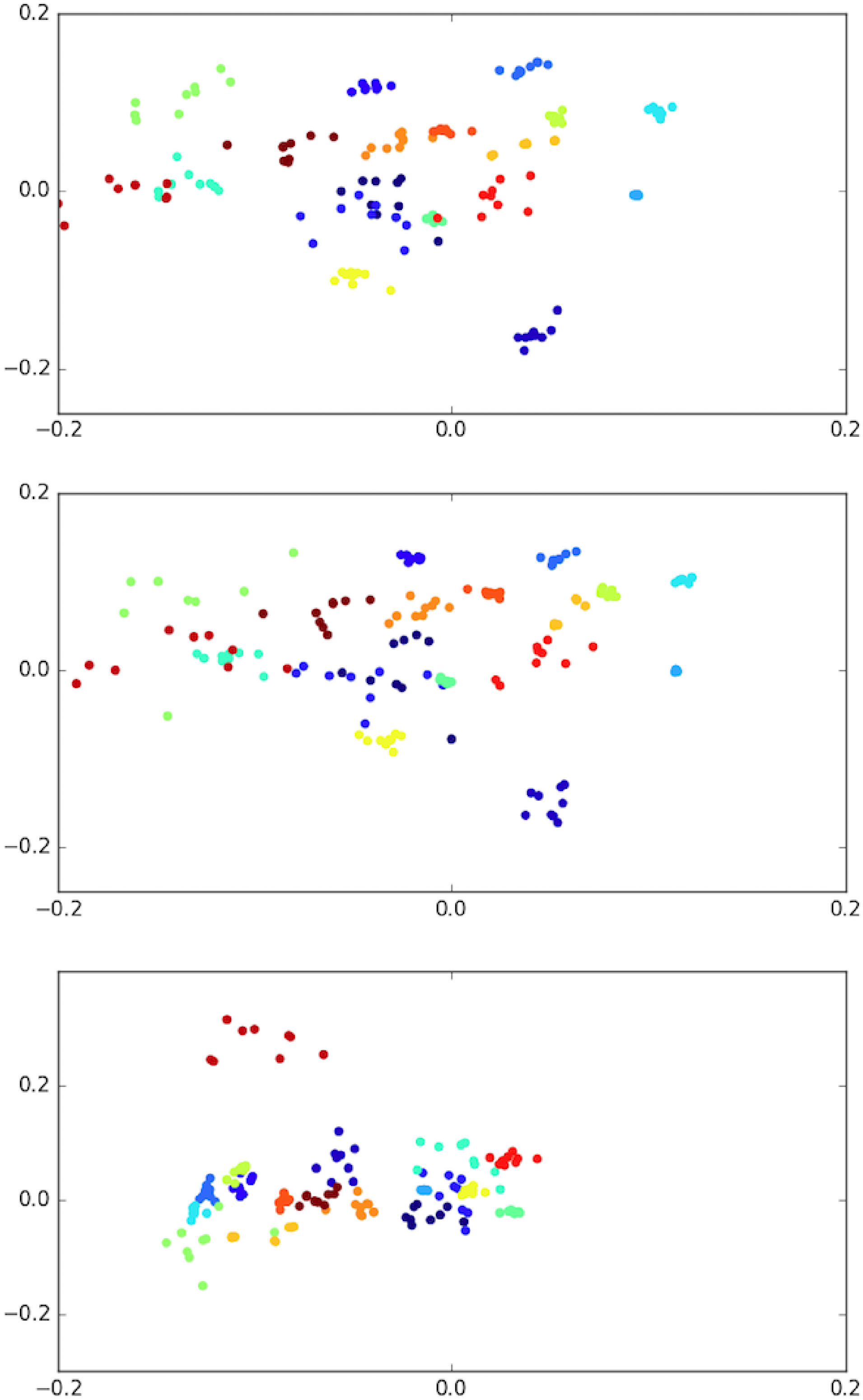


Figure 3.3: First two components of probabilistic GDA (top), tangent LDA (middle), Bayesian PGA (bottom), (arbitrary units).

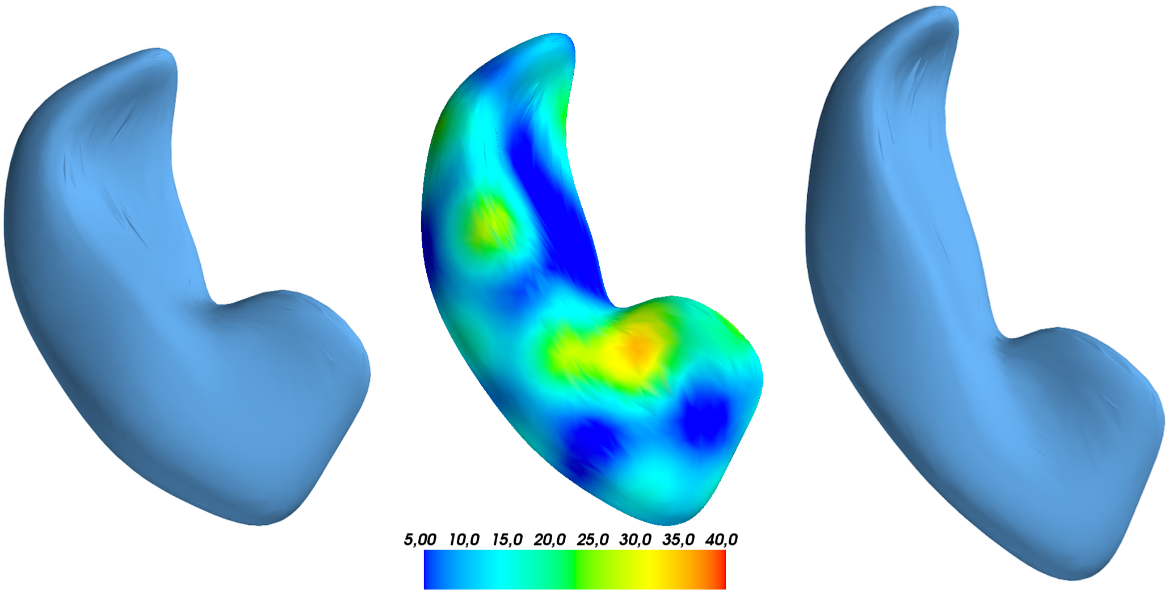


Figure 3.4: Three observations on the first geodesic discriminant component estimated from the ADNI dataset. From left to right, we follow the geodesic from normal controls to MCIc subjects. The middle observation is the estimated M template, colored with the initial velocity field norm.

	AUC (std)	Accuracy (std)
Tangent LDA	0.77 (0.06)	0.76 (0.07)
Probabilistic GDA	0.78 (0.07)	0.77 (0.08)
Volumes	0.68 (0.003)	0.56 (0.003)
Chupin et al. [13]	x	0.71
Cuignet et al. [16]	x	0.73

Table 3.2: AUC and accuracy scores at MCIc vs normal controls classification using only the hippocampus. Colors indicate shape labels.

controls versus MCIc subjects. Our method provides state-of-the-art accuracy and AUC results. Note that the problem of classifying MCIc versus normal controls is in general much better solved using whole T1 MRIs, which could be future work using the same proposed probabilistic GDA applied to full 3D images.

Our method in this case could be compared with [42] in which the authors do an analysis of hippocampi differences between Alzheimer's and normal controls modelling the shapes using the elastic shape framework [42]. However, their analysis of the differences is done after having performed a PCA on the tangent space to the Fréchet mean, and their approach requires a parametrization of the surfaces.

3.6 Conclusion

We propose generalizations of the different formulations of LDA. The geometric GDA constructs a geodesic subspace which maximizes the FDR as seen in this curved space, but is hard to compute in general. The probabilistic GDA, generalization of the Gaussian classifier formulation of LDA, is much more efficient to compute. We illustrated the methods with dimension reduction and classification tasks, with an example on a set of 3D shapes segmented from subjects with Alzheimer's disease where we reach state-of-the-art classification results.

Applications to data sets of different types would allow to best show the applicability of the method. Future work also includes improving the estimation procedure for the probabilistic model, to take full advantage of the hidden variable β, using for instance use a stochastic version of the EM algorithm [19]. In addition, several theoretical discussions could be conducted: to see when the π operation is well-defined, to study the consistency of the estimation and the identifiability of the model or to formulate criteria to identify and handle degenerate cases.

Part III

Riemannnian geometry learning

Introduction

So far, we built methods to work with manifold-valued data: to quickly compute the parallel transport on a given manifold or to perform the equivalent of Linear Discriminant Analysis on curved spaces. Doing so, we kept postulating that the Riemannian geometry on which we worked was given and fixed in advance. This is a very clear limitation of these methods. First, even when it is possible to coin a Riemannian manifold in the observation space, this geometry may be ill conceived and not suited for the particular task at stake. For instance, authors of [3, 89] equip sub-manifolds of the observation spaces with the metric induced from the ℓ^2 metric on the observations: we argue that this is an ad hoc choice which has no particular reason to be particularly relevant. The choice of Riemannian metric in particular is extremely important since most statistical methods on Riemannian manifolds rely heavily on the metric tensor –through distances, the volume form or geodesics– and therefore their performances depend on this tensor. This discussion is echoed in the LDDMM community, where the traditional hand-crafted way to generate diffeomorphisms from a well-chosen RKHS space is replaced by a data-driven diffeomorphic construction (see [8] for instance). Second, for complex spaces, such as images, it is difficult to come up with a relevant Riemannian manifold.

In [49], the authors propose to use a trained variational autoencoder to pull-back the induced metric on the parametrized surface by the decoder back to the latent space. Similar approaches are developed in [3]. We discuss the difference with our approach in Section 4.3.1.

In this part, we therefore tackle the issue of estimating a manifold and a Riemannian metric at the same time as we train a model to achieve a given statistical task. In Chapter III, we detail the motivation and propose a list of criteria that could be used to learn a Riemannian geometry. In the rest of the manuscript, we then propose to find when and how these criteria can be used to learn a Riemannian manifold. We provide basic existence results, estimation methods, and considerations regarding the uniqueness of the Riemannian manifolds which best optimize the given criteria. Each of the results will be further motivated by experiments.

In Chapter 4, we discuss when and how it is possible to estimate a Riemannian metric so that a given distribution is a Riemannian normal distribution on the obtained Riemannian manifold. In Chapter 5, we discuss when and how it is possible to estimate a Riemannian metric on $\mathbb{R}^d$ so that a given family of curves are geodesics.

Finally, we come back to the modelling of longitudinal disease progression and propose a way to learn a manifold so as to refine and extend the field of application of a model such as the one proposed in [86]. We first propose in Section 5.6 a somehow pedestrian approach for estimating a Riemannian metric along with the longitudinal model parameters. This approach, although mathematically principled, has a complexity which scales exponentially with the dimension of the observation space. We therefore relax the modelling in Chapter 6 and propose an alternative approach for Riemannian geometry learning which is much more efficient in high dimensions. This approach relies on deep neural networks. We show how the two approaches provide similar results when run on the same data. Finally, we extend the longitudinal model obtained to handle multimodal observations, even when there are missing modalities at some of the subject's visits.

In this Chapter, we start by giving formal definitions for Riemannian normal distributions in Section III. This allows us to formulate a list of criteria that can be used to constrain a Riemannian geometry learning problem. We motivate and list these criteria in Section III.

Riemannian normal distributions

In this Section, $(\mathcal{M}, g)$ is a d-dimensional Riemannian manifold. A random point X on $\mathcal{M}$ is an $\mathcal{M}$-valued random variable. Note that if $\pi : U \to \mathbb{R}^d$ is a coordinate chart with U an open subset of $\mathcal{M}$, then the $Y = \pi(X)$ is a random vector. We can now define probability densities for random points on $\mathcal{M}$.

Definition 1. *A random point X on $\mathcal{M}$ has a probability density function $p : \mathcal{M} \to \mathbb{R}$ if, for all measurable sets A on $\mathcal{M}$, we have:*

$$P(x \in A) = \int_A p(y) \mathrm{d}\mathcal{M}(y)$$

and

$$1 = P(\mathcal{M}) = \int_{\mathcal{M}} p(y) \mathrm{d}\mathcal{M}(y).$$

which means that X has a density for the measure $\mathrm{d}\mathcal{M}(y)$*.*

Remark. If X has a probability density function $p : \mathcal{M} \to \mathbb{R}$, then $Y = \pi(X)$ is a random vector with probability density $\rho(\pi(x)) = p(x)\sqrt{|g(\pi(x))|}$, where $|\cdot|$ is the determinant and g is the metric (see equation (8.6) in Appendix 8 for details). We will often work in coordinates in the Chapters which follow, heavily relying on this remark.

Geodesic completeness

To guarantee the good definition of random points, mean and covariance of random variables on a manifold, it is necessary to make regularity assumptions on the manifold. In [78], the authors assume that $(\mathcal{M}, g)$ is geodesically complete. By the Hopf-Rinow Theorem, this guarantees that the Riemannian logarithm is always well-defined and that the manifold, seen as a metric space, is complete. Note that in general, geodesic completeness is a required hypothesis when building statistical models on a manifold, since these models often need the existence of a length-minimizing geodesic connecting any two points.

Definition 2 (Normal distribution). *Let $\mu \in \mathcal{M}$ and Γ be a bilinear form on $T_\mu\mathcal{M}$. The normal distribution with mean μ and concentration matrix Γ on the manifold has probability density:*

$$p_{\mu,\Gamma}(x) = k \exp\left(-\log_\mu(x)^\top \Gamma \log_\mu(x)\right) \tag{3.17}$$

where

$$k = \left(\int_{\mathcal{M}} \exp\left(-\log_\mu(x)^\top \Gamma \log_\mu(x)\right) \mathrm{d}\mathcal{M}(x)\right)^{-1} \tag{3.18}$$

is the appropriate normalization constant.

As noted above, if we have a coordinate chart $\Phi : \mathcal{M} \to \mathbb{R}^d$ on the manifold in a neighborhood of μ, then the density of the random variable $Y = \Phi(X)$ expressed in this coordinate chart reads:

$$p_{\mu,\Gamma}(y) = k \exp\left(-\log_\mu(\Phi^{-1}(y))^\top \Gamma \log_\mu(\Phi^{-1}(y))\right) \sqrt{|g(y)|}$$

where all computations are done in coordinates and k can be re-expressed as:

$$k = \left(\int_{\mathbb{R}^d} \exp\left(-\log_\mu(\Phi^{-1}(y))^\top \Gamma \log_\mu(\Phi^{-1}(y))\right) \sqrt{|g(y)|}\mathrm{d}y^1 \dots dy^k\right)^{-1} \tag{3.19}$$

where all computations are done in coordinates.

In [111], the authors use a more restricted definition of Gaussian distributions on manifolds, by only considering concentration forms of the form $\Gamma = \tau g$ at μ with $\tau > 0$.

Motivations and criteria for Riemannian geometry learning

We now put the focus on learning Riemannian manifolds that are adapted to a certain statistical task. We list here a few of the constraints that can be expressed on a Riemannian manifold:

(C_1) An observed data set is distributed approximately along a normal distribution on $(\mathcal{M}, g)$ as defined in Section III. We will show that this task is, under some conditions, the one performed by Generative Adversarial Networks for instance. We give theoretical and practical considerations for this criterion in Chapter 4.

(C_2) The observed classes in the data are well separated using the Riemannian distance. In particular, we could optimize a criterion similar to the one in LDA (see e.g. [101]) to obtain a Riemannian metric inducing large distances on elements of different classes. This is what inspired works such as [101]. It could also be seen as an extension of the work detailed in Part 3 where we would not only learn a submanifold so that the classes are well separated, but the metric on this manifold too.

(C_3) An observed set of curves are geodesics on $(\mathcal{M}, g)$. We give in Chapter 5 some theoretical results about the existence of a metric so that a given set of curves are geodesics, detailed results for the toy example $\mathcal{M} = \mathbb{R}$ as well as experimental results.

(C_4) Any combination of the above: a notion of distances between observations, a list of curves which we want to be geodesics and/or a distribution that should be matched.

The question now is, for each case, whether it is possible to optimize a manifold and a Riemannian metric on this manifold so that the obtained geometry satisfies the criterion. In which case is there a unique Riemannian metric fulfilling the criterion ? In other cases, how constrained is a Riemannian metric which satisfies one of the criteria ?

Note that criterion (C_1) could alternatively be formulated using distributions which are not normal but uniform or Cauchy. We use this example as a baseline case to study in a simple environment how much it constraints a Riemannian manifold.

Note that this list of criteria is not exhaustive. In particular, we could imagine formulating a cost which depends on the curvature of the manifold, on its volume form, on the parallel transport etc. We do consider however that this list is general enough to provide interesting insights into the feasibility and applicability of Riemannian manifold learning in general.

Remark. In practice, if we were to optimize any of the given criteria, we would need a cost function to optimize, or alternatively a procedure ensuring we reach optimality of these criteria in some sense. For criterion (C_1), this could be achieved by measuring a distance –e.g. ℓ^2, KL or Wasserstein– between the target distribution f and the normal distribution on the manifold. For criterion (C_3), this could be achieved by computing the integrals of the ℓ^2 distances between geodesics on $\mathcal{M}$ and the target curves. From there, one can imagine being able to build a cost function which is a adapted to a combination of these criteria, and we will give such examples later.

Chapter 4

Riemannian metrics so as to be normally distributed

In this Chapter, we study the criterion (C_1), that is to know when it is possible to estimate a Riemannian manifold such that a given density becomes the density of a normal distribution on the manifold.

To start, note that the criterion (C_1) is not sufficient to fully characterize a Riemannian metric when the dimension of the manifold is larger than 1, as shown in [37] for instance. Informally, fixing this criterion only fixes –relative– distances to the mean as measured along geodesics connecting points to the mean. Consequently, any 'radial' transformations does not modify the density of the distribution.

We start with detailed results for the toy example $\mathcal{M} = \mathbb{R}$, where analytical computations are possible. Then, we expose a simpler version of the criterion where we look at manifolds which are diffeomorphic to $\mathbb{R}^d$ for $d \in \mathbb{N}$ and equip them with the push-forward of the Euclidean metric η on $\mathbb{R}^d$. We show that formulated this way, the criterion (C_1) corresponds to the task accomplished by a Generative Adversarial Network (GAN) [33].

4.1 The toy example $\mathcal{M} = \mathbb{R}$

Here we consider the case $\mathcal{M} = \mathbb{R}$. We find all the random variables with probability density on $\mathbb{R}$ for which there exists a metric g such that its density is the density of a normal distribution on $(\mathbb{R}, g)$. We show uniqueness of such metrics –up to a multiplicative constant– and exhibit analytical formulae to compute them.

We will use the following Lemma:

Lemma 5. *Let $u : \mathbb{R} \to \mathbb{R}$ be a smooth function. Then we have:*

$$\sqrt{\frac{\pi}{2}} erf\left(\frac{u(x)}{\sqrt{2}}\right)' = u'(x) \exp\left(-\frac{1}{2}u^2(x)\right) \tag{4.1}$$

where $erf : \mathbb{R} \mapsto]-1, 1[$ is the error function. We also have:

$$(erf^{-1}(x))' = \frac{\sqrt{\pi}}{2} \exp\left(erf^{-1}(x)^2\right). \tag{4.2}$$

We now give necessary conditions for a function f to be the density of a normal distribution on $(\mathbb{R}, g)$ for some metric g:

Proposition 5. *Let f be the density function of a Riemannian normal distribution with mean μ and concentration matrix $\gamma > 0$ in the canonical chart on $(\mathbb{R}, g)$. We assume that $(\mathbb{R}, g)$ is geodesically complete. Then:*

i. $f(x) > 0$ *for all* $x \in \mathbb{R}$.

ii. $\int_{-\infty}^{\infty} f(t)\mathrm{d}t = 1$,

iii. $\lim_{x\to\pm\infty} \int_{\mu}^{x} \frac{f(t)}{f(\mu)}\mathrm{d}t = \pm\sqrt{\frac{\pi}{2\gamma}}$. *(In particular, μ is the median of the distribution.)*

Proof. We have, for all $x \in \mathbb{R}$:

$$f(x) = k \exp\left(-\frac{\gamma}{2}(\log_\mu x)^2\right)\sqrt{g(x)}$$

where $k^{-1} = \int_{-\infty}^{\infty} \exp\left(-\frac{\gamma}{2}(\log_\mu x)^2\right)\sqrt{g(x)}\mathrm{d}x$. So $f(x) > 0$ for all x which shows i. Then:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_{-\infty}^{\infty} k \exp\left(-\frac{\gamma}{2}(\log_\mu x)^2\right)\sqrt{g(x)}\mathrm{d}x = \frac{k}{k} = 1$$

which shows ii. For $x \in \mathbb{R}$, we have $f(\mu) = k\sqrt{g(\mu)}$ and:

$$\frac{f(t)}{f(\mu)} = \sqrt{\frac{g(t)}{g(\mu)}} \exp\left(-\frac{\gamma}{2}(\log_\mu x)^2\right) = \frac{1}{\sqrt{\gamma}}\sqrt{\gamma\frac{g(t)}{g(\mu)}} \exp\left(-\frac{1}{2}\left(\int_\mu^t \sqrt{\gamma\frac{g(s)}{g(\mu)}}\mathrm{d}s\right)^2\right).$$

We now use (4.1) with $u(t) = \int_\mu^t \sqrt{\gamma\frac{g(s)}{g(\mu)}}\mathrm{d}s$ to get:

$$\frac{f(t)}{f(\mu)} = \sqrt{\frac{\pi}{2\gamma}}\mathrm{erf}\left(\frac{1}{\sqrt{2}}\int_\mu^t \sqrt{\gamma\frac{g(s)}{g(\mu)}}\mathrm{d}s\right)'.$$

Now integrating from μ to x, we get:

$$\int_\mu^x \frac{f(t)}{f(\mu)}\mathrm{d}t = \sqrt{\frac{\pi}{2\gamma}}\mathrm{erf}\left(\sqrt{\frac{\gamma}{2}}\int_\mu^x \sqrt{\frac{g(t)}{g(\mu)}}\mathrm{d}t\right) - 0.$$

Now, since $(\mathbb{R}, g)$ is geodesically complete, then the integral on the right-hand side goes to $\pm\infty$ as $x \to \pm\infty$. This implies:

$$\lim_{x\to\pm\infty}\int_\mu^x \frac{f(t)}{f(\mu)}\mathrm{d}t = \pm\sqrt{\frac{\pi}{2\gamma}}$$

which shows iii. $\square$

We now prove that these conditions are sufficient.

Theorem 2. *Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth positive function. We assume:*

i. $f(x) > 0$ *for all* $x \in \mathbb{R}$.

ii. $\int_{-\infty}^{\infty} f(t)\mathrm{d}t = 1,$

Then there exists a smooth Riemannian metric g on $\mathbb{R}$, $\mu \in \mathbb{R}$ and a concentration matrix γ such that:

- $(\mathbb{R}, g)$ *is geodesically complete.*
- f *is the density of a normal distribution with mean μ and concentration matrix γ on $(\mathbb{R}, g)$.*

This metric is unique up to a global rescaling ($g \mapsto \alpha g$ with $\alpha > 0$).

Proof. **Existence** Let $\alpha \in \mathbb{R}, \mu \in \mathbb{R}+$ be such that: $\lim_{x\to\pm\infty} \int_\mu^x \frac{f(t)}{f(\mu)}\mathrm{d}t = \pm\sqrt{\frac{\pi}{2\gamma}}$ (possible using the hypothesis i). We can then define for all $x \in \mathbb{R}$:

$$\boxed{g(x) = \left[\frac{2}{\sqrt{\pi}}\frac{f(x)}{f(\mu)}(\mathrm{erf}^{-1})'\left(\sqrt{\frac{2\gamma}{\pi}}\int_\mu^x \frac{f(t)}{f(\mu)}\mathrm{d}t\right)\right]^2} \tag{4.3}$$

where $\mu \in \mathbb{R}$. g is smooth by composition and positive using i: g is a Riemannian metric on $\mathcal{M}$. We also have $\sqrt{g(\mu)} = \frac{2}{\sqrt{\pi}}(\mathrm{erf}^{-1})'(0) = 1$. Then:

$$\int_\mu^x \sqrt{g(t)}\mathrm{d}t = \sqrt{\frac{2}{\gamma}}\left[\mathrm{erf}^{-1}\left(\sqrt{\frac{2\gamma}{\pi}}\int_\mu^x \frac{f(t)}{f(\mu)}\mathrm{d}t\right) - \mathrm{erf}^{-1}(0)\right] = \sqrt{\frac{2}{\gamma}}\mathrm{erf}^{-1}\left(\sqrt{\frac{2\gamma}{\pi}}\int_\mu^x \frac{f(t)}{f(\mu)}\mathrm{d}t\right)$$

which implies

$$\sqrt{\frac{2\gamma}{\pi}}\int_\mu^x \frac{f(t)}{f(\mu)}\mathrm{d}t = \mathrm{erf}\left(\sqrt{\frac{\gamma}{2}}\int_\mu^x \sqrt{\frac{g(t)}{g(\mu)}}\mathrm{d}t\right)$$

We now use (4.1) to get:

$$\frac{f(t)}{f(\mu)} = \sqrt{\frac{g(t)}{g(\mu)}}\exp\left(-\frac{\gamma}{2}\left(\int_\mu^t \sqrt{\frac{g(s)}{g(\mu)}}\mathrm{d}s\right)^2\right) = \frac{\gamma}{\sqrt{g(\mu)}}\exp\left(-\frac{1}{2}\left(\int_\mu^t \sqrt{\frac{g(s)}{g(\mu)}}\mathrm{d}s\right)^2\right)\sqrt{g(t)}$$

so that finally:

$$f(t) = \frac{f(\mu)}{g(\mu)}\exp\left(-\frac{\gamma}{2}\left(\int_\mu^t \sqrt{\frac{g(s)}{g(\mu)}}\mathrm{d}s\right)^2\right)\sqrt{g(t)}.$$

We now set $k = \frac{f(\mu)}{\sqrt{g(\mu)}}$. Since f verifies ii, we have $k^{-1} = \int_{\mathbb{R}} \exp\left(-\frac{\gamma}{2}\left(\int_{\mu}^{t}\sqrt{\frac{g(s)}{g(\mu)}}\mathrm{d}s\right)^2\right)\sqrt{g(t)}\mathrm{d}t$. Finally, $(\mathbb{R}, g)$ is geodesically complete thanks to ii and Proposition 15. This shows existence.

Uniqueness Let g be a Riemannian metric on $\mathbb{R}$ and let $\mu \in \mathbb{R}$ and $\gamma > 0$ such that f is the density of a normal distribution on $(\mathbb{R}, g)$ with mean μ and concentration matrix γ. First, using ii, we obtain that γ and μ are as defined above. Then: $f(\mu) = k\sqrt{g(\mu)}$ and by definition, we have:

$$\frac{f(x)}{f(\mu)} = \frac{1}{\sqrt{\gamma}}\sqrt{\gamma\frac{g(x)}{g(\mu)}}\exp\left(-\frac{1}{2}\left(\int_{\mu}^{x}\sqrt{\gamma\frac{g(t)}{g(\mu)}}\mathrm{d}t\right)^2\right).$$

We use (4.1) again to get:

$$\frac{f(x)}{f(\mu)} = \sqrt{\frac{\pi}{2\gamma}}\left(\mathrm{erf}\left(\sqrt{\frac{\gamma}{2}}\left(\int_{\mu}^{x}\sqrt{\frac{g(t)}{g(\mu)}}\mathrm{d}t\right)\right)\right)'$$

so that

$$\int_{\mu}^{x}\frac{f(t)}{f(\mu)}\mathrm{d}t = \sqrt{\frac{\pi}{2\gamma}}\mathrm{erf}\left(\sqrt{\frac{\gamma}{2}}\int_{\mu}^{x}\sqrt{\frac{g(t)}{g(\mu)}}\mathrm{d}t\right)$$

and

$$\sqrt{\frac{\gamma}{2}}\left(\int_{\mu}^{x}\sqrt{\frac{g(t)}{g(\mu)}}\mathrm{d}t\right) = \mathrm{erf}^{-1}\left(\sqrt{\frac{2\gamma}{\pi}}\int_{\mu}^{x}\frac{f(t)}{f(\mu)}\mathrm{d}t\right)$$

which implies

$$\sqrt{g(x)} = \sqrt{g(\mu)}\frac{2}{\sqrt{\pi}}\frac{f(x)}{f(\mu)}(\mathrm{erf}^{-1})'\left(\sqrt{\frac{2\gamma}{\pi}}\int_{\mu}^{x}\frac{f(t)}{f(\mu)}\mathrm{d}t\right).$$

Hence g is proportional to the one found in equation (4.3). Therefore, any metric which is so that f is a normal distribution on $(\mathbb{R}, g)$ is proportional to the metric defined in (4.3). This show uniqueness up to a global rescaling. □

To summarize, given a smooth positive function f which integrates to 1, there exists a unique –up to a global rescaling– geodesically complete Riemannian metric on $\mathbb{R}$, unique $\gamma > 0$ and $\mu \in \mathbb{R}$ such that f is the density of a normal distribution on $(\mathbb{R}, g)$ with mean μ and concentration matrix γ. μ is the median of the distribution, while γ can be easily identified as $2\pi f(\mu)^2$. Note that in the previous Theorem, we showed a solution with $g(\mu) = 1$. It is always possible to set $\gamma = 1$ instead and use the metric with $g(\mu) = \gamma$. Up to this re-parametrization, this metric is unique.

Diffeomorphic formulation

We now look into how the case where $\mathcal{M}$ is the image of $\mathbb{R}$ by a diffeomorphism. It can then be equipped with the push-forward of the Euclidean metric on $\mathbb{R}$. Does it allow to recover any form of densities as above ?

Let $\Phi : \mathbb{R} \to \mathbb{R}$ be a diffeomorphism. We consider the push-forward of a Riemannian metric η by Φ i.e. the pull-back of η by Φ^{-1}. We have:

$$\Phi_*(\eta)(p) = \left(\Phi^{-1}\right)^* (\eta)(p) = \left(\Phi^{-1\prime}\right)^2 (p)\eta(\Phi^{-1}(p)). \tag{4.4}$$

We show in Appendix 8.6 that any metric g on $\mathbb{R}$ such that $(\mathbb{R}, g)$ is geodesically complete can be written as the pull-back of the Euclidean metric by a well-chosen diffeomorphism of $\mathbb{R}$. Using Theorem 2, this implies that for any smooth positive function f, f can be seen as the image of a normal Euclidean density on $\mathbb{R}$ by a diffeomorphism of $\mathbb{R}$. Therefore, to estimate a metric which optimizes criterion (C_1), it is equivalent to estimate a diffeomorphism of $\mathbb{R}$ which optimizes this criterion. Besides, looking at equation (4.3) where g is defined as the squared of a derivative, there is a straightforward correspondence between the diffeomorphism and the metric. A diffeomorphism Φ such that $\Psi_*(\eta)$ is the optimum of criterion (C_1) in this case is given as:

$$\Phi^{-1}(x) = \sqrt{\frac{2}{\gamma}}\mathrm{erf}^{-1}\left(\sqrt{\frac{2\gamma}{\mu}}\frac{\int_\mu^x f(t)\mathrm{d}t}{f(\mu)}\right). \tag{4.5}$$

Note that we have $\Phi^{-1}(\mu) = 0$.

4.2 Experiments

In this section, we set target densities f on $\mathbb{R}$ which obey the hypothesis of Theorem 2. We then compute, for each of these target densities

- The Riemannian metric on $\mathbb{R}$ which is such that f is the density of a normal distribution on $(\mathbb{R}, g)$.
- The diffeomorphism Ψ which is such that f is the density of a normal distribution on $(\mathbb{R}, \Psi_*(\eta))$ where η is the Euclidean metric on $\mathbb{R}$.

The computation of the Riemannian metric g is done using equation (4.3).[1] In practice, the computations are fine as long as the cumulative distribution function corresponding to f is accessible in closed-form. Indeed, if it is accessible, a simple binary search allows to find the median μ of the distribution in an efficient manner, and the integral in equation

[1] Code for these experiments: `https://gitlab.com/maxime.louis.x2012/one_dimensional_distributions`.

(4.3) can be efficiently computed. If this cumulative distribution function is unavailable, then integrals of the form $\int_\mu^x f(t)\mathrm{d}t$ needs to be approximated numerically. In preliminary experiments, this proved to be too unstable for correct estimations. The estimation of the diffeomorphism is done using equation (4.5) and is as difficult to compute as the metric. Note that we set a metric equal to 1 for all these experiments.

Figure 4.1 shows the results on 4 different target densities. In each case, we display the computed metric g, the corresponding density of the normal distribution on $(\mathbb{R}, g)$ and the diffeomorphism Ψ which is such that $\Psi_*(\eta) = g$. When the target distribution is a normal distribution, we recover the Euclidean metric. Note also that there is no issue in writing a bi-modal or multimodal distribution as a normal distribution for some Riemannian metric.

Note also that there is a competition effect between the local measure $\sqrt{g(t)}$ and the logarithms. Namely, a large local measure allocates high probabilities to a given area, but a large area between the mean and a given point tends to diminish this probability, since it increases the geodesic distance.

This is particularly clear for the bi-modal case, where the increase in logarithm is compensated by a very large increase in metric around the modes.

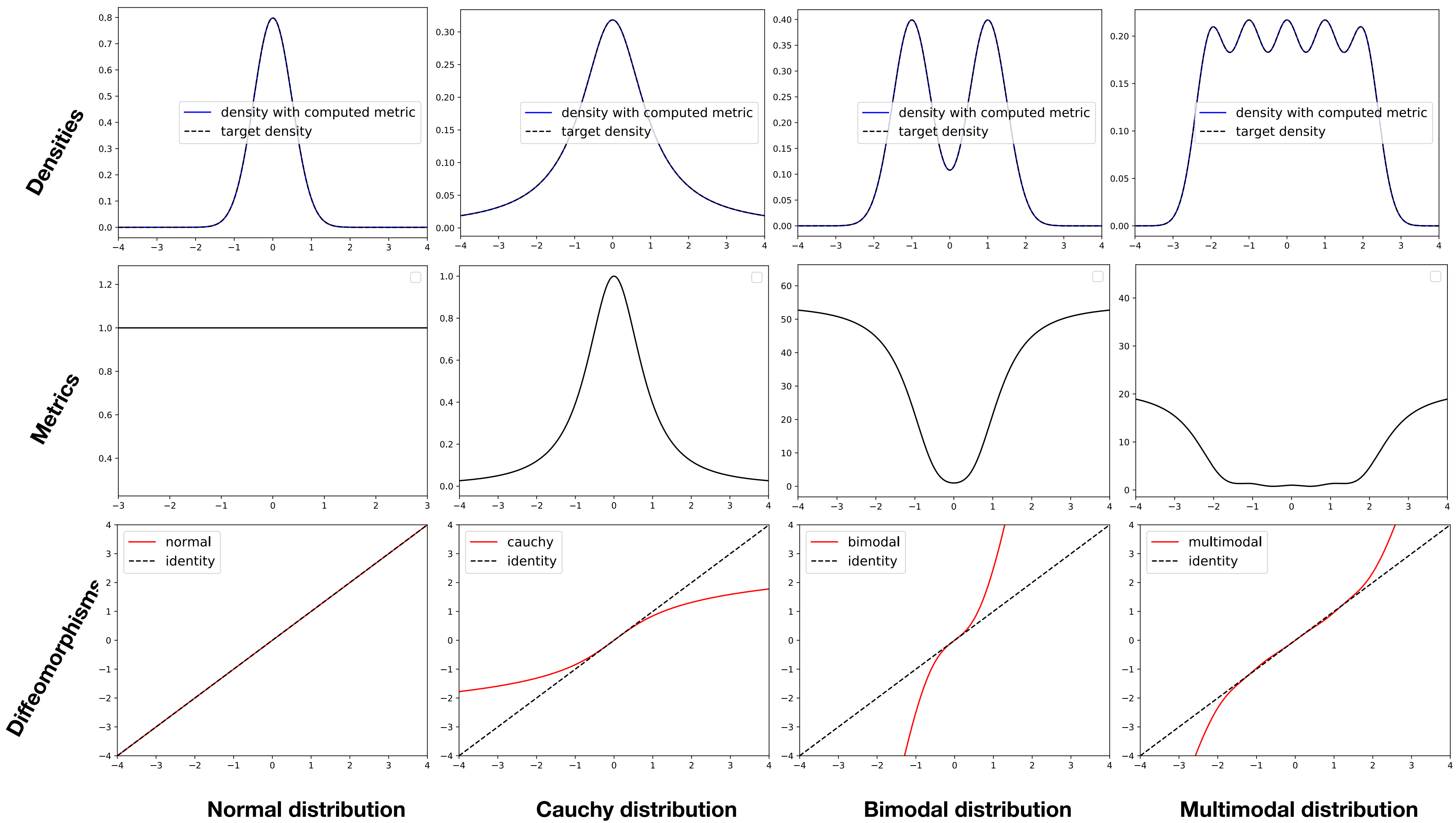


Figure 4.1: For four different target distributions (Normal, Cauchy, a combination of two normals and a combination of five normals) we plot the metric g such that the target distribution is the density of a normal distribution on $\mathcal{M}$, the corresponding densities and a diffeomorphism Ψ such that g is obtained from push-forward of the Euclidean metric via Ψ.

Estimation from samples

Suppose we are given samples $(x_i)_{i=1,\dots,N}$, with $N \in \mathbb{N}$, from a distribution with density f. We approach f using a kernel methods [29]:

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1,\dots,N} K\left(\frac{x - x_i}{h}\right) \tag{4.6}$$

where K is a radial positive kernel and h is a chosen width. We choose a simple approach with the heuristic $h = 1.06 * \hat{\sigma} N^{-\frac{1}{5}}$ (see [29]) and using a Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$. We get a closed form expression for the cumulative distribution function:

$$\begin{aligned}
C(x) &= \frac{1}{Nh} \sum_{i=1,\dots,N} \int_{-\infty}^{x} K\left(\frac{t - x_i}{h}\right) \mathrm{d}t &(4.7)\\
&= \frac{1}{Nh\sqrt{2\pi}} \sum_{i=1,\dots,N} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{t - x_i}{h}\right)^2\right) \mathrm{d}t &(4.8)\\
&= \frac{1}{N\sqrt{\pi}} \sum_{i=1,\dots,N} \int_{-\infty}^{\frac{x-x_i}{h\sqrt{2}}} \exp\left(-u^2\right) \mathrm{d}u &(4.9)\\
&= \frac{1}{2} + \frac{1}{2N} \sum_{i=1}^{N} \operatorname{erf}\left(\frac{x - x_i}{\sqrt{2}h}\right) &(4.10)
\end{aligned}$$

which allows us to compute efficiently the estimated metric $\hat{g}$ and $\hat{\gamma}$ as detailed above. The median $\hat{\mu}$ of the density $\hat{f}$ can be approximated efficiently by binary search. This allows the estimation from a sample collection, when the density is not available. Note that the task accomplished here is, as noted above, similar to the task accomplished by GANs.

Figure 4.2 shows an example of metric estimation done this way, on the hippocampus volumes from the ADNI database (previously normalized to zero-mean and unit variance).

4.3 Shedding light on Generative Adversarial Networks

Let $d, D \in \mathbb{N}$ with $d \leq S$. Let $\Psi : \mathbb{R}^d \to \mathbb{R}^D$ be a smooth diffeomorphism on its image. We now look into the density of a normal distribution on $\mathcal{M} := \Psi(\mathbb{R}^d)$ equipped with the metric $g := \Psi_*(\eta)$. Let $\mu = \Psi(z_\mu) \in \Psi(\mathbb{R}^d)$ and Γ be a bilinear form on $T_\mu\mathcal{M}$. Let $x = \Psi(z_x) \in \Psi(\mathbb{R}^d)$. $(\mathcal{M}, g)$ is geodesically complete. The density of the normal distribution with mean μ and concentration Γ is :

$$p_{\mu,\Gamma}(x) = k \exp\left(-\log_\mu(x)^\top \Gamma \log_\mu(x)\right).$$

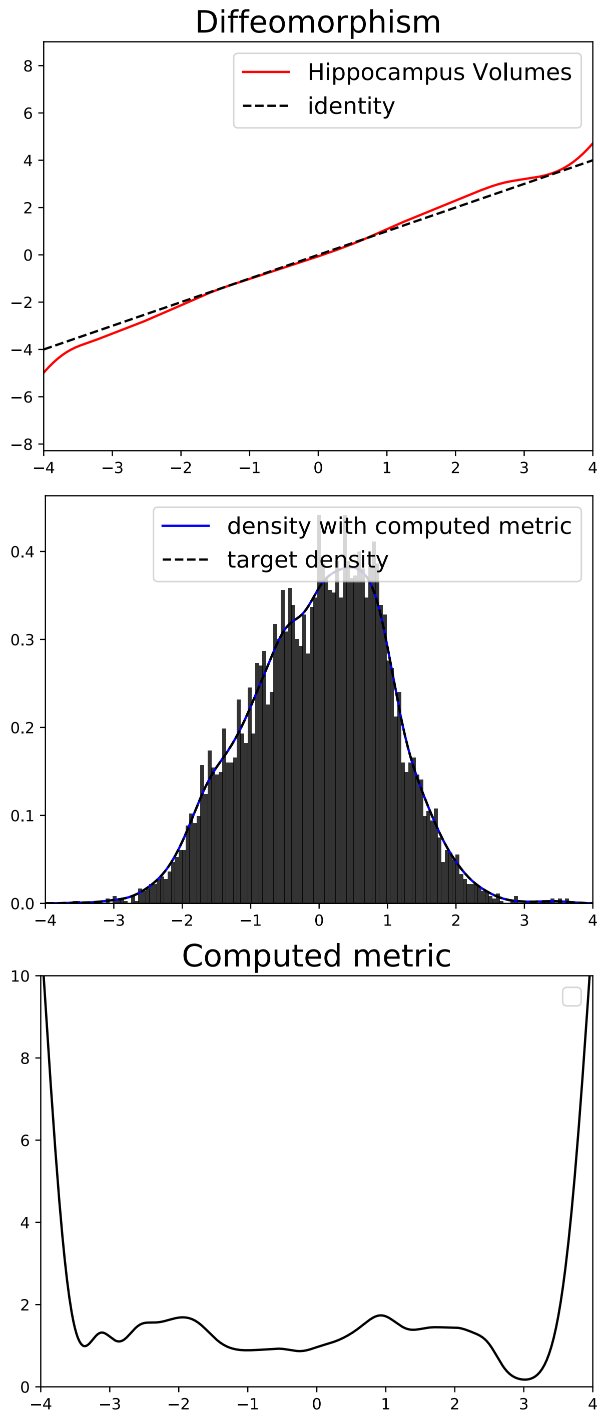


Figure 4.2: Estimation of a Riemannian metric from samples, using a kernel density estimation (4.7).

We have $\log_\mu(x) = \Psi_*(z_x - z_\mu)$ so that:

$$p_{\mu,\Gamma}(x) = k \exp\left(-\Psi_*(z_x - z_\mu)^\top \Gamma \Psi_*(z_x - z_\mu)\right).$$

In particular, if $\Gamma = g$, then:

$$p_{\mu,\Gamma}(x) = k \exp\left(-\|z_x - z_\mu\|^2\right).$$

In coordinates, this can be rewritten as:

$$p_{\mu,\Gamma}(z) = k \exp\left(-\|z - z_\mu\|^2|\right) \sqrt{|g(z)|}. \tag{4.11}$$

Now, we have $\sqrt{|g(z)|} = 1$ in the coordinate chart defined by Ψ^{-1} by definition. Therefore we obtain:

$$p_{\mu,\Gamma(z)} = k \exp\left(-\|z - z_\mu\|^2\right). \tag{4.12}$$

This means that, as one could expect, a normal distribution with mean μ and concentration matrix g on a manifold which is defined as the image of $\mathbb{R}^d$ by a diffeomorphism has for the density the image of the density of the usual normal distribution on $\mathbb{R}^d$ with mean $\Psi^{-1}(\mu)$ and unit covariance.

This does shed some light on the task undertaken by Generative Adversarial Networks [33], where the generator attempts to map a normal distribution (or a uniform distribution) onto an observed distribution of data. Our interpretation here, provided that this generator is injective, is that this task is a Riemannian geometry learning task, where the Riemannian metric is learned so that a normal distribution on it is close to the data distribution. The optimization of the criterion (C_1) in this case is done by minimizing an estimation of the Kullback-Leibler divergence between the normal on the manifold and the data [33] or by minimizing an estimation of the Wasserstein distance between these distributions [1].

Generative adversarial networks

In [33], the authors propose a setup to generate realistic data from a sample of data in $\mathcal{Y}$. If the data is distributed along the density π_{data}, they offer to optimize a network $g : \mathbb{R}^d \mapsto \mathcal{Y}$ –the generator– such that the image of samples from a fixed simple distribution in $\mathbb{R}^d$ –typically a uniform or a Gaussian– are likely elements of the distribution π_{data}. To do so, they use an extra network –the discriminator– which is trained to distinguish between real samples from the data and fake samples generated by the generator. Both the generator and the discriminator are trained simultaneously in a competitive fashion.

Note that the injectivity condition that we enforce on the generative network to obtain this interpretation can in principle be enforced using neural networks such as the ones described in [41].

4.3.1 Push-forward versus pull-back metric

A Generative Adversarial Network is an example of deep generative model Ψ, which maps a low-dimensional latent space $\mathcal{Z}$ to a larger observation space $\mathcal{Y}$. The analysis of such generative models under the scope of differential geometry is a blooming domain. In [89] for instance, the authors find simple conditions so $\Psi(\mathcal{Z})$ is locally a submanifold of $\mathcal{Y}$ when ψ is a deep neural network. Then, they show how to pull-back an observation space metric –such as the ℓ_2 distance on images– back to the latent space $\mathcal{Z}$, and how Riemannian exponentials, logarithms and parallel transport can be computed on the latent space with this geometry. In [3, 37, 103], the authors follow this procedure to derive further procedures allowed by this modelling, such as clustering.

While this procedure of pulling-back an ambient metric from the observation space back to the latent space has the advantage of removing some of the arbitrariness from the generative model latent space structure, as discussed in [37], we argue here that the choice of metric in the observation space is ad hoc. In most cases, this metric will be chosen to be the Euclidean distance, which may very well be inadapted to capture differences of interest between images. At best, it can be a handcrafted metric containing some knowledge about the nature of the observations. On the other hand, as shown above, when considering $\mathcal{Z}$ equipped with the Euclidean metric, the obtained structure of the observations placed in $\mathcal{Z}$ is not completely arbitrary, since they must follow a normal distribution. Pushing-forward this metric into the observation space then yields a Riemannian manifold which is also not completely arbitrary: the observations are normally distributed – which corresponds to a minimal level of prior knowledge used– and radial distances to the mean are relevant. Of course, in every other direction, the obtained distances are arbitrary. In any case, one can imagine that by imposing a much stronger structure on the latent space, such as with criteria (C_3) or (C_4), one can get a Riemannian metric on $\Psi(\mathcal{Y})$ which is best adapted to a targeted statistical task.

This is what we propose in the following Chapters. In Chapter 5, we show results and experiments when learning a metric so that a set of observed curves are geodesics. In Chapter 5 and 6, we propose experiments to learn a Riemannian geometry which is adapted to model disease progression. In these approaches, we place stronger constraints on the set of metrics which optimize the criteria.

Chapter 5

Riemannian metrics for geodesicity

In this Chapter, we provide results around criterion (C_3), which aims at finding a metric so that a given set of curves are all geodesics. For simplicity, we will work in most cases with a fixed manifold, and we study when there exists a Riemannian metric on this manifold so that a given set of curves are geodesics for this metric.

The main problems associated with this criterion are:

- When does there exist a Riemannian manifold such that a set of curves are all geodesics on this manifold ?
- When is such a Riemannian manifold unique ?

We start by a definition:

Definition 3. *Let $\mathcal{M}$ be a Riemannian manifold and let g be Riemannian metric on $\mathcal{M}$. We say that g is geodesically rigid if all other metrics $\tilde{g}$ which have the same unparametrized geodesics as g are proportional to g.*

So that the second question becomes: what metrics are geodesically rigid ?

We start by dealing with the toy example $\mathcal{M} = \mathbb{R}$ in Section 5.1. We exhibit all curves which are geodesics for a given metric and provide explicit formulae to compute the corresponding metric.

We then reproduce in Section 5.2 a result found in [71] showing that a Riemannian metric is almost always geodesically rigid. This indicates that criterion (C_3) is sufficiently strong to fully constrain a Riemannian metric –at least when all geodesics are considered – in the sense that all optimal metrics for this criterion are almost always proportional.

In Section 5.3 we prove a simple general existence result for injective or cyclic smooth curves on a smooth manifold: this is the main result of this Chapter. This acts as complementary motivation, indicating that in a variety of situations there exists metrics which are optimal for criterion (C_3).

Finally, Section 5.5 shows some experimental results using a parametric family of metrics on $\mathbb{R}^d$ introduced in Section 5.4.

5.1 The toy example $\mathcal{M} = \mathbb{R}$

In this section, we discuss the existence and the uniqueness of a metric g for which a smooth curve $\gamma : \mathbb{R} \to \mathbb{R}$ is a geodesic on $(\mathcal{M}, g)$.

We start with a result about uniqueness.

Proposition 6. *Let $g_1, g_2 : \mathcal{M} \to \mathbb{R}+^*$ be two smooth positive functions such that $(\mathbb{R}, g_1)$ and $(\mathbb{R}, g_2)$ are geodesically complete. We set $p \in \mathcal{M}$ and define:*

$$F_i(u) = \int_p^u \sqrt{g_i(t)}\mathrm{d}t, \ \ i = 1, 2$$

As shown in the Lemma 15 in the Appendix, since $(\mathcal{M}, g_1)$ and $(\mathcal{M}, g_2)$ are geodesically complete, then F_1 and F_2 are diffeomorphism of $\mathbb{R}$.

If there exists $a, b \in \mathbb{R}$ with $a \neq 0$ such that for all $t \in \mathbb{R}$ $F_1^{-1}(t) = F_2^{-1}(at + b)$ then g_1 and g_2 are proportional. Equivalently, if $(\mathbb{R}, g_1)$ and $(\mathbb{R}, g_2)$ share a non-trivial geodesic, then g_1 and g_2 are proportional. This means that all metrics on $\mathbb{R}$ are geodesically rigid.

Proof. Let $a, b \in \mathbb{R}$ with $a \neq 0$ such that $F_1^{-1}(t) = F_2^{-1}(at + b)$. Note that we have, for all $u \in \mathbb{R}$, $F_1'(u) = \sqrt{g_2(u)}$ and $(F_1^{-1})'(u) = \frac{1}{F_1'(F_1^{-1}(u))}$. Let $t \in \mathbb{R}$, we have $F_1(F_2^{-1}(at+b)) = t$. Deriving this identity yields:

$$\begin{aligned} 1 &= \frac{d\left(F_2^{-1}(at+b)\right)}{\mathrm{d}t}\sqrt{g_1(F_2^{-1}(at+b))} \\ &= \frac{a}{F_2'(F_2^{-1}(at+b))}\sqrt{g_1(F_2^{-1}(at+b))} \\ &= \frac{a}{\sqrt{g_2(F_2^{-1}(at+b))}}\sqrt{g_1(F_2^{-1}(at+b))}. \end{aligned}$$

F_2^{-1} is onto $\mathbb{R}$ which implies that for all $t \in \mathcal{M}$ we have $a^2 g_1 = g_2$.

Now let us assume that $(\mathbb{R}, g_1)$, $(\mathbb{R}, g_2)$ share a non-trivial geodesic $\gamma : \mathbb{R} \to \mathbb{R}$. Using Lemma 6, there exists $a_1, b_1, a_2, b_2 \in \mathbb{R}$ such that for all $t \in \mathbb{R}$ $\gamma(t) = F_1^{-1}(a_1 t + b_1) = F_2^{-1}(a_2 t + b_2)$ with $a_1 \neq 0$ and $a_2 \neq 0$ since γ is non-trivial. This means that for $u \in \mathbb{R}$ we have $F_1^{-1}(u) = F_2^{-1}\left(\frac{a_2}{a_1}u + b_2 - b_1\frac{a_2}{a_1}\right)$ which implies that g_1 and g_2 are proportional. □

Now, we prove a sufficient condition for the existence of a metric such that γ is a geodesic on $(\mathbb{R}, g)$.

Proposition 7. *Let $\gamma : \mathbb{R} \to \mathbb{R}$ be a smooth strictly monotonic function such that $\gamma(t) \xrightarrow[t \to \pm\infty]{} \pm\infty$. Then there exists a metric g such that $(\mathbb{R}, g)$ is geodesically complete and such that γ is a geodesic on $(\mathbb{R}, g)$.*

Proof. We define $g : \mathbb{R} \mapsto \mathbb{R}$ by $g(t) = \frac{1}{\gamma'^2(\gamma^{-1}(t))}$. This is well-defined since $\gamma'(t) > 0$ for all $t \in \mathbb{R}$. g is positive and smooth so that it is a Riemannian metric on $\mathbb{R}$.

We now show that $(\mathbb{R}, g)$ is geodesically complete. We define $F(t) = \int_0^t \sqrt{g(u)}\mathrm{d}u$. By Proposition 15, it is enough to show that $\lim_{t\to\pm\infty} F(t) = \pm\infty$. We have $F(\gamma(t))' = \gamma'(t)F'(\gamma(t)) = \gamma'(t)\sqrt{\frac{1}{\gamma'^2(t)}} = \mathrm{sign}(\gamma'(t)) = \mathrm{sign}(\gamma'(0))$, so that we have $F(\gamma(t)) = \mathrm{sign}(\gamma'(0))t$. Therefore, we have $\lim_{t\to\pm\infty} F(t) = \pm\infty$.

We now show that γ is a geodesic on $(\mathbb{R}, g)$. We have $F(\gamma(t)) = \mathrm{sign}(\gamma'(0))t$. Besides, F is a diffeomorphism of $\mathbb{R}$ and we have $\gamma(t) = F^{-1}(\mathrm{sign}(\gamma'(0)t)$. Using Lemma 6 this shows that γ is a geodesic on $(\mathbb{R}, g)$. □

Finally, we show that this condition is necessary.

Proposition 8. *Let g be a metric on $\mathbb{R}$ such that $(\mathbb{R}, g)$ is geodesically complete. Then any geodesic γ is monotonic and is such that $\lim_{t\to\pm\infty} \gamma(t) = \pm\infty$.*

Proof. Monotony comes from Lemma 6. The second property comes from Lemma 6 combined with Proposition 15. □

Therefore, a single geodesic on $\mathbb{R}$ equipped with a geodesically complete Riemannian metric contains enough information to identify this metric. This solves the problem of identifying a metric for criterion (C_3) when $\mathcal{M} = \mathbb{R}$. Note that the proof is constructive: given a curve which satisfies the shown conditions, it is possible to compute exactly a Riemannian metric which makes it a geodesic. This Riemannian metric is unique up to a rescaling constant.

5.2 Geodesically rigid metrics

In this section, we reproduce a result from [71]. In the paper, the authors discuss whether it is possible to reconstruct a metric given its non-parametrized geodesics. Their goal is to know whether from telescope observations of free falling objects one can reconstruct the space-time metric. Free falling objects follow unparametrized geodesics for the space-time metric, and it is possible to follow the trajectory of very distant objects for instance using Super Nova explosions occurring in distant galaxies. We have a very different motivation: the identification of a Riemannian geometry of interest for a targeted statistical task.

Here, the paper gives a more interesting result in the perspective of identifiability of metric learning machine learning. The main result is the following:

Theorem 3 (Matveev 2012)**.** *Let $n \geq 4$. A generic Riemannian metric on $\mathbb{R}^n$ is geodesically rigid i.e. the space of geodesically rigid Riemannian metrics is open and dense for the $\mathcal{C}^2$-topology in the space of Riemannian metrics.*

We reproduce a sketch of the proof here for $n = 4$, it can be adapted to larger dimensions. The proof consists in three steps:

- Construction of a geodesically rigid metric $\overline{g}$ defined in a neighborhood of $x_0 \in \mathcal{M}$.

- Proof that any metric $\tilde{g}$ is arbitrarily close to a geodesically rigid metric $\overline{g}_t$, done by adding $\varepsilon\overline{g}$ to g (openness).
- Proof that all metrics close enough to g_t are geodesically rigid (density).

Proof. First, the authors construct a geodesically rigid metric $\overline{g}$ on $\mathbb{R}^n$. They pick $x_0 \in \mathbb{R}^4$, and choose a metric $\overline{g}$ which is the identity at x_0 and such that its curvature tensor R_{ijkl} at x_0 is explicitly given by:

$$R_{ijkl} = h_{ik}h_{jl} - h_{il}h_{jk} + H_{ik}H_{jl} - H_{il}H_{jk}$$

where $h = \operatorname{diag}(1, 2, -1, 0)$ and $H = \operatorname{diag}(0, 0, 1, 1)$. The existence of such a metric is known in the literature.

Now the key ingredient is to use the fact that $\overline{g}$ satisfy the equation:

$$ng^{a(i}W^{j)} = g^{ab}W^{(i}_{ab[l}\delta^{j)}_{k]} \tag{5.1}$$

which comes by definition of the projective Weyl tensor W (see e.g. equation (19) in[71]). Now, every metric which is geodesically equivalent to $\overline{g}$ has the same Weyl tensor as $\overline{g}$. Hence, any metric g which is geodesically equivalent to $\overline{g}$ is a solution of the equation (5.1) at x_0. One can show, using the prescribed forms of $\overline{g}$ –and hence W– at x_0 that this system is exactly of rank 9 and has 10 unknowns –the components of g at x_0. Therefore, all solutions to this system are proportional to $g(x_0)$. Now, the system of equations (5.1) is of rank 9 in a neighborhood of x_0, since it is the biggest dimension of a non-zero minor determinant of g. Therefore, in a neighborhood of x_0, g is proportional to $\overline{g}$. Hence, g and $\overline{g}$ are conformally equivalent. A final Theorem by Weyl proves that conformally equivalent 4-dimensional metrics are proportional. Using the exact same arguments as before, we then get that $\overline{g}$ is geodesically rigid.

Let $\tilde{g}$ be an arbitrary metric in a neighborhood of x_0. Let $\varepsilon > 0$. The authors consider $g_t = (1-t)\tilde{g} + t\overline{g}$. The system (5.1) is of rank 9 for g_t for t sufficiently close to 1, as shown above. Since the coefficients of the system are algebraic expressions in t and in the coefficients of $\overline{g}$ and $\tilde{g}$, then the system is of rank 9 for almost all $t \in [0, 1]$. We select t which is such that the rank of the system (5.1) for g_t is 9 and such that g_t is at distance less than ε from $\tilde{g}$ for the $\mathcal{C}^2$-topology[1]. Then g_t is geodesically rigid.

Finally, as shown above, there exists $\varepsilon' > 0$ such that any metric which is ε'-close to g_t is geodesically rigid. Hence a generic metric is geodesically rigid. $\square$

The proof can be adapted in higher dimensions.

This result is an excellent motivation for tasks which rely on the criterion (C_3). Indeed, this means that the set of geodesics of a given metric is enough to identify this metric (up

[1]It is important to consider this topology since the system of equations (5.1) has coefficients which are functions of the first and second derivatives of the metric.

to a global rescaling). This could be an ingredient in demonstrating the identifiability of some statistical models which offer to use this criterion.

5.3 A more general existence result

We prove the following Theorem establishing a simple condition for a set of curves on a manifold $\mathcal{M}$ to be geodesics for a certain metric.

Theorem 4. *Let $\mathcal{M}$ be a smooth manifold and $(\gamma_i)_{i\in\{1,\dots,n\}}$ be a family of smooth regular injective curves on $\mathcal{M}$ which do not intersect. There exists a Riemannian metric g such that γ_i is a geodesic on $(\mathcal{M}, g)$ for all $i \in \{1,\dots,n\}$.*

The proof first construct a tubular neighborhood of each geodesic. Then, each tubular neighborhood is mapped onto a normal bundle on the segment $[0,1]$, which can be equipped with a metric such that its 0-section is a geodesic. Finally, we stitch the obtained metrics for each geodesic using a partition of unity.

Proof. Open neighborhood of the curves. Let $i \in \{1,\dots,n\}$. Since γ_i is injective and regular, $\gamma_i([0,1])$ is a submanifold of $\mathcal{M}$. By the tubular neighborhood Theorem, there exists a vector bundle E_i on $\gamma_i([0,1])$ and a diffeomorphism Φ_i from a neighborhood U_i of the 0-section in E_i to a neighborhood V_i of $\gamma_i([0,1])$ in $\mathcal{M}$ such that $\Phi_i \circ 0_{E_i} = \iota_i$ where ι_i is the embedding of $\gamma_i([0,1])$ in $\mathcal{M}$. Without loss of generality, we can suppose $\gamma_j([0,1]) \cap U_i = \emptyset$ for all $j \neq i$.

Riemannian metric on the neighborhoods. To do so, we use the fact that E_i is diffeomorphic to the normal bundle on the segment $[0,1] \in \mathbb{R}$ which is trivial and which we denote N. Let $\Psi_i : E_i \to N$ be such a diffeomorphism. Now N can be equipped with a Riemannian metric h_i such that $[0,1]$ is a geodesic. Using Ψ_i and Φ_i, we can push-forward the metric h_i to get a metric g_i on U which is so that γ_i is a geodesic on (U_i, g_i).

Stitching the metrics with a partition of unity. For each $i \in \{1,\dots,n\}$, pick an open subset V_i such that $V_i \subset \overline{V_i} \subset U_i$ and which contains $\gamma_i([0,1])$, and set $O = \mathcal{M} \setminus (\cup_i \overline{V_i})$. O is open so that $\mathcal{C} = \{O, U_1, \dots, U_n\}$ is an open cover of $\mathcal{M}$. O can be equipped with a metric g_O (there always exists a Riemannian metric on a smooth manifold). Finally, we use a partition of the unity $\rho_O, \rho_1, \dots, \rho_n$ on $\mathcal{C}$ and set $g = \rho_O g_O + \sum_i \rho_i g_i$. g is a Riemannian metric on $\mathcal{M}$ as a positive combination of Riemannian metrics. Each γ_i is a geodesic on $(\mathcal{M}, g)$ by construction. □ □

This result can be easily extended to include smooth curves $\gamma : [0,1] \to \mathcal{M}$ which are injective on $[0,1[$, verify $\gamma(0) = \gamma(1)$ and $\gamma'(0) = \gamma'(1)$ by considering tubular neighborhoods which are this time diffeomorphic to the normal bundle on the circle. It can also easily be extended to the case when there is an infinite number of non-intersecting curves.

This result indicates that there are metrics which are optimal for criterion (C_3) in a wide variety of situations.

5.4 A parametric family of Riemannian metrics on $\mathbb{R}^d$

For the experiments which follow, we explain here how to build a parametric Riemannian metric on $\mathbb{R}^d$ with $d \in \mathbb{N}$.

5.4.1 Cholesky decomposition

We first recall the Cholesky decomposition of positive definite matrices.

Proposition 9. *Let $n \in \mathbb{N}$. Let $\Sigma \in M_n(\mathbb{R})$ be a positive definite matrix. Then there exists $U \in M_n(\mathbb{R})$ upper triangular such that $\Sigma = U^\top U$. There exists a unique such U with positive diagonal coefficients. Conversely, for any U upper triangular with positive diagonal coefficients, $U^\top U$ is a positive-definite matrix.*

Thus, finding a parametrization for positive definite matrices amounts to having upper triangular matrices with positive diagonal coefficients. This turns out to be advantageous from a computational perspective, since any gradient-based can be used by computation of the gradient with respect to the upper triangular matrices parameters directly. This allows to keep positive definite matrices throughout the estimation.

5.4.2 Building a Riemannian metric

Let $d, s \in \mathbb{N}$ and $\rho > 0$. Using Proposition 9, the parametrization of a Riemannian metric on $\mathbb{R}^d$ can be done using upper triangular matrices with positive diagonal coefficients. To make this parametrization space-dependent, we fix a set of points $(x_i)_{i=1,\dots,s}$ in $\mathbb{R}^d$, and let $(U_i)_{i=1,\dots,s}$ be a set of upper triangular matrices with positive diagonal coefficients. For any $q \in \mathbb{R}^d$, we define:

$$g_U^{-1}(q) = \sum_{i=1}^{s} U_i^\top U_i \exp\left(-\frac{\|x_i - q\|^2}{\rho^2}\right) + I_D \tag{5.2}$$

g_U^{-1} is everywhere positive definite –even if the $U_i - s$ all have zero diagonal coefficients. Note that far from the x_i-s, g_U converges to the identity. Besides, note that g^{-1} remains positive definite if the U_i's all have non-negative diagonal entries. This is an important fact, which indicates that when estimating the U_i in an optimization procedure, it will be easy to maintain a well-defined metric on the whole space $\mathbb{R}^d$.

We choose to parametrize the inverse of the metric and not the metric itself to allow an easier integration of the geodesic Hamiltonian equations and an easier use of the Fanning scheme developed in Part I.

Hamiltonian equations. We can compute explicitly the Hamiltonian and the Hamiltonian equations corresponding to this inverse metric, at $q, p \in \mathbb{R}^d$:

$$H(p,q) = p^\top g_U^{-1}(q) p = \frac{1}{2}\left(\sum_{i=1} p^\top U_i^\top U_i p \exp(-\frac{\|x_i - q\|^2}{\rho^2}) + p^\top p\right)$$

so that:

$$\frac{\partial H(p,q)}{\partial p} = \sum_{i=1}^{s} U_i^\top U_i p \exp(-\frac{\|x_i - q\|^2}{\rho^2}) + p$$

and

$$\begin{aligned}
\frac{\partial H(p,q)}{\partial q} &= \frac{1}{2}\sum_{i=1}^{s} p^\top U_i^\top U_i p \frac{\partial \exp(-\frac{\|x_i-q\|^2}{\rho^2})}{\partial q} \\
&= \frac{1}{2}\sum_{i=1}^{s} p^\top U_i^\top U_i p \frac{\partial \left(-\frac{\|x_i-q\|^2}{\rho^2}\right)}{\partial q} \exp(-\frac{\|x_i - q\|^2}{\rho^2}) \\
&= -\frac{1}{\rho^2}\sum_{i=1}^{s} p^\top U_i^\top U_i p \,(q_k - x_{ik}) \exp(-\frac{\|x_i - q\|^2}{\rho^2}).
\end{aligned}$$

Using these formulae, it is possible to integrate the Hamiltonian equations (1.5) and compute geodesics for any given initial condition. An implementation of this Riemannian metric allowing geodesic computation, parallel transport and automatic differentiation on any cost function with respect to the metric parameters is available here: `https://gitlab.com/maxime.louis.x2012/parametric_riemannian_metric`.

Figure 5.1 shows one-dimensional geodesics obtained using this procedure, as well as the corresponding inverse metrics, which in this case are simply positive functions. In this example, the interpolation points are regularly spaced between 0 and 1, so that the inverse metric goes to 1 as x goes to $\pm\infty$. Therefore, when a geodesic starts to go lower than 0 or higher than 1, its progression gets close to a linear progression. When $d = 1$, we have the following:

Proposition 10. *Let $(x_i)_{i=1,\dots,s}$, $(U_i)_{i=1,\dots,s} > 0$ and let g_U be the corresponding Riemannian metric on $\mathbb{R}$. Then $(\mathbb{R}, g_U)$ is a geodesically complete Riemannian manifold.*

Proof. The result comes from Proposition 15. In our case, first $x \mapsto \int_0^x \sqrt{g(t)}\mathrm{d}t$ is well-defined and smooth on $\mathbb{R}$ and it is clear that $\int_0^x \sqrt{g(t)}\mathrm{d}t \xrightarrow[x\to-\infty]{} -\infty$ and $\int_0^x \sqrt{g(t)}\mathrm{d}t \xrightarrow[x\to+\infty]{} +\infty$. □

This guarantees that $(\mathbb{R}, g_U)$ is geodesically complete. The result actually holds for any $d \in \mathbb{N}$:

Proposition 11. *Let $d \in \mathbb{N}$, let $(x_i)_{i=1,\dots,s}$ in $\mathbb{R}^d$ and let $(U_i)_{i=1,\dots,s} \in \mathcal{M}_d(\mathbb{R})$ be a set of upper triangular matrices with non-negative diagonal coefficients. Then $(\mathcal{M}, \mathbb{R}^d)$ is geodesically complete.*

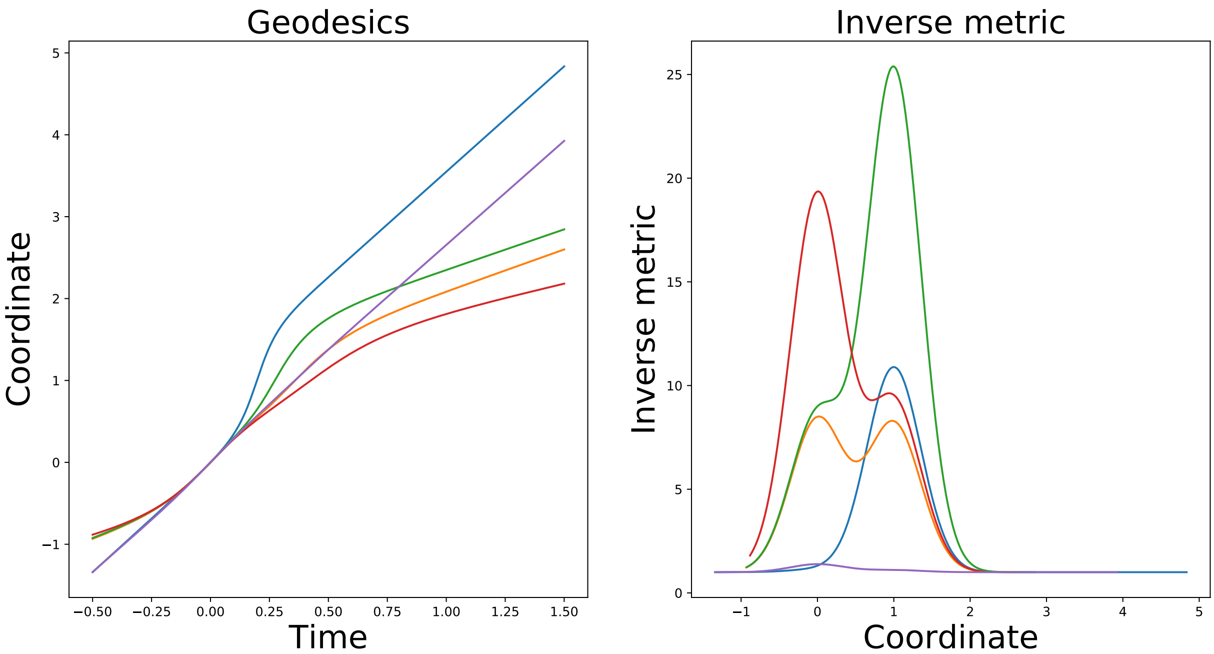


Figure 5.1: Geodesics with same initial conditions but varying metrics.

Proof. Towards a contradiction, let $\gamma :]a, b[\to \mathcal{M}$ be a geodesic which is non-extendible geodesic at b. The proof consists in two steps.

γ is bounded: Let $t_0 \in]a, b[$. We have, for any t in $]a, b[$:

$$\|\dot{\gamma}(t)\|_2^2 \leq \|\dot{\gamma}(t)\|_2^2 + \sum_{i=1}^{s} \dot{\gamma}(t)^\top U_i^\top U_i \dot{\gamma}(t) \exp\left(-\frac{\|x_i - \gamma(t)\|^2}{\rho^2}\right) = \|\dot{\gamma}(t)\|_{g_U}^2 = \|\dot{\gamma}(t_0)\|_{g_U}^2$$

Therefore, we have $\|\gamma(t) - \gamma(t_0)\|_2 \leq \|\dot{\gamma}(t_0)\|_{g_U}(t - t_0)$ for all $t \in]a, b[$ so that γ remains within a compact subset K of $\mathbb{R}^d$.

γ can be extended: We now consider a sequence $(t_n) \in]a, b[$ such that $t_n \xrightarrow{n \to \infty} b$. The set $I = \{(t_n, \dot{\gamma}(t_n)) | t_n \in]a, b[\}$ is compact. Besides, for any $N \in \mathbb{N}$, there exists $\varepsilon > 0$ such that the geodesic can be extended to $]t_N - \varepsilon, t_N + \varepsilon[$ by Cauchy-Lipschitz Theorem. Therefore, there exists a uniform $\varepsilon > 0$ such that for all $N \in \mathbb{N}$, the geodesic can be extended around t_N to $]t_N - \varepsilon, t_N + \varepsilon[$. This implies that γ can be extended to $]a, b + \frac{\varepsilon}{2}[$ at least, which is a contradiction. □

Figure 5.2 illustrates the same procedure done in two dimensions, for varying parameters ρ. The figure shows both geodesics on $\mathbb{R}^2$ and the corresponding metrics, where at each point, the symmetric positive-definite matrix which is the metric is represented by an ellipse. Lower values of ρ allow for more irregular metrics. Figure 5.3 shows geodesics when the Riemannian metric is diagonal with linearly increasing eigenvalues in the x-direction.

These explicit formulae allow efficient computations of geodesics on the Riemannian manifold $(\mathcal{M}, \mathbb{R}^D)$. The obtained geodesic points can then be differentiated with respect

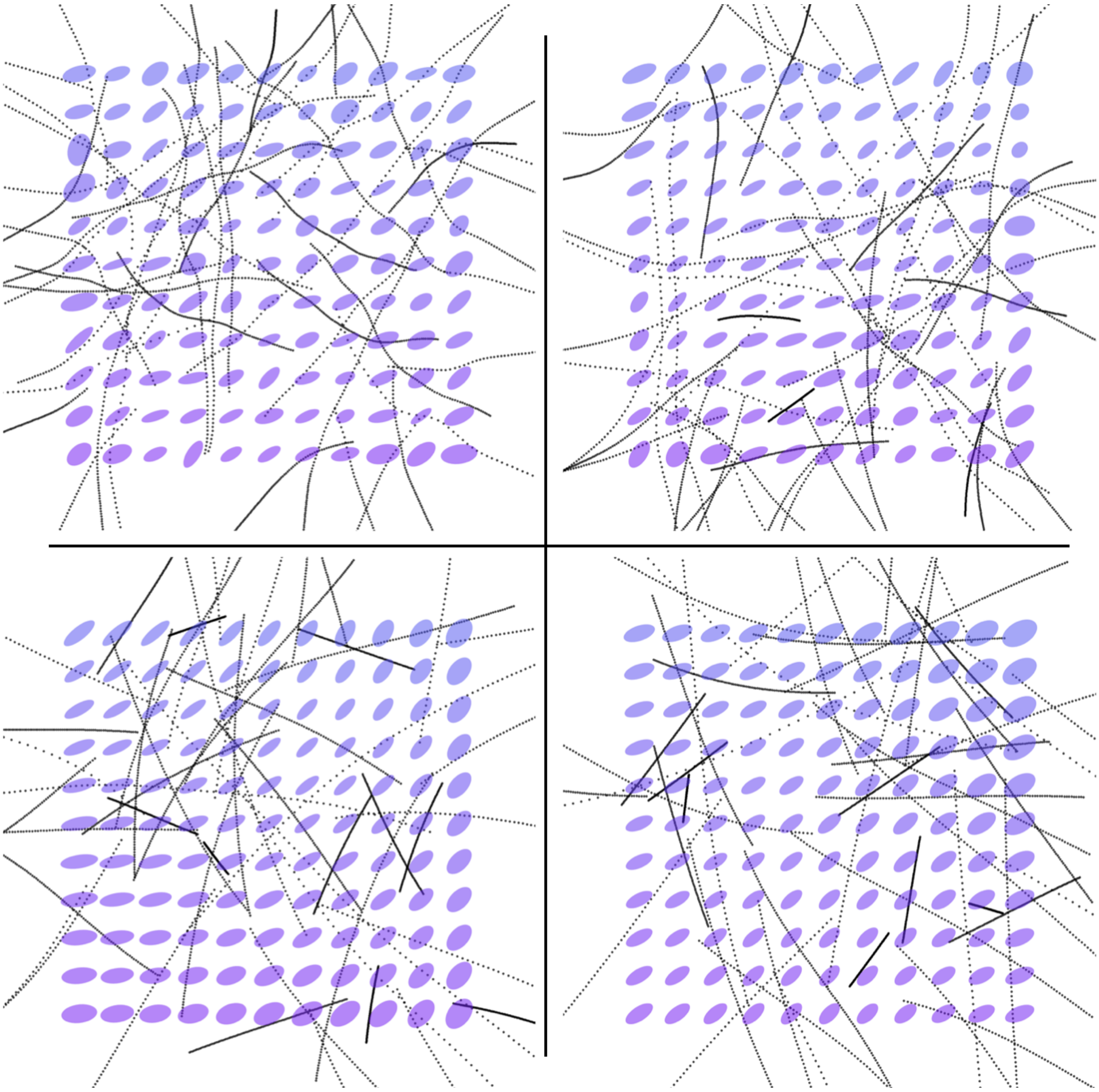


Figure 5.2: Examples of metrics and geodesics corresponding to the metric described in Section 5.4, for $\rho = 0.05, 0.1, 0.3, 0.5$, from left to right, top to bottom. Geodesics tend to follow shortest directions which mean they tend to be 'orthogonal' to the observed ellipses. Lower values of ρ allow for more irregular metrics. The geodesics are affinely parametrized, and proper time is shown using the dotted representation of the trajectories.

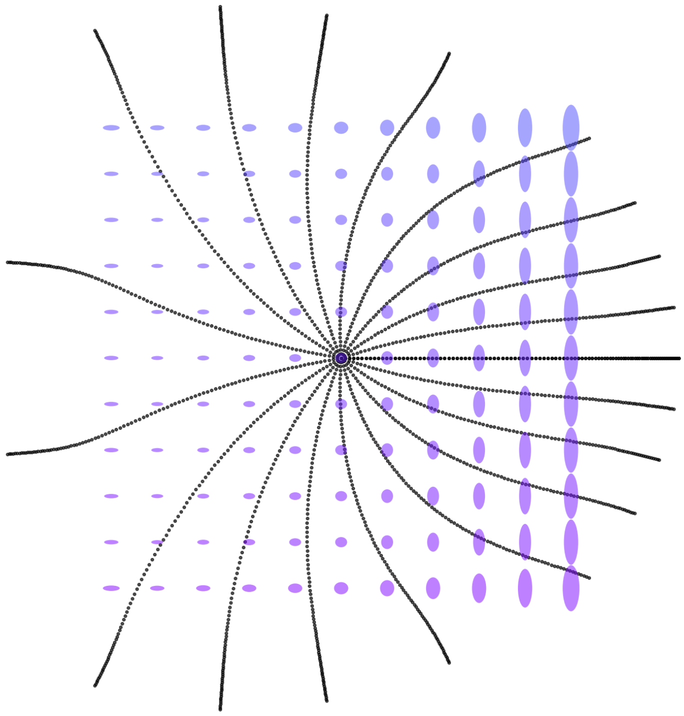


Figure 5.3: Geodesics, all starting form the center point and going outwards with isotropic initial conditions. The Riemannian metric prescribed here is with an eigenvalue in the y direction which increases linearly with x. When going in the x-direction, the cost of a displacement in the y direction becomes larger, so that geodesics are bent in the x direction.

to the metric coefficients $(U_i)_i$ using automatic differentiation. This will prove to be a memory intensive operation, since the graph of computation gets large: the gradient has to be back-propagated through each integration step of the geodesic equation for instance. The Fanning scheme can be used on this metric in an efficient way since the inverse metric and its spatial derivatives are available in closed-form. All together, this allows for a functional implementation of a parametric Riemannian metric.

We will show in the next section examples of use of this parametrization to optimize a cost function inspired from criterion (C_3).

5.5 Experiments

To illustrate the results of this Chapter, we conducted a series of experiments using the parametric family of metrics defined in Section 5.4. We evaluated the ability to recover a given parametric metric from its geodesics.

So let $d \in \mathbb{N}$, $(x_i)_{i=1,\dots,s}$ in $\mathbb{R}^d$, and let $(U_i^t)_{i=1,\dots,s}$ be upper triangular matrices. We denote g_U the corresponding metric as defined in equation (5.2). We compute geodesics $\gamma_j : [0,1] \to \mathbb{R}^d$ for $j \in \{1,\dots,n\}$ with $n \in \mathbb{N}$ from this metric, with initial positions scattered within $]0,1[^d$ and initial velocities regularly spanning the possible directions in $\mathbb{R}^d$. Our goal from there is to recover the parameters U from the geodesics γ_j. To do so we minimize the cost function by gradient descent:

$$c(V_i) = \sum_{j=1}^{n} \int_{t=0}^{1} \|\gamma_j(t) - \mathrm{Exp}^V_{\gamma(0)}\left(\dot{\gamma}_j(0)t\right)\|_2^2 \mathrm{d}t \tag{5.3}$$

where V is a set of upper triangular matrices and Exp^V denotes the Riemannian exponential on $(\mathcal{M}, g_V)$. This cost function correspond to a measurement of the ℓ^2 distance between affinely parametrized geodesics from the metrics, g_U and g_V. The hope is that when a large number of geodesics are used in the cost function, then V will converge towards U during the estimation. In practice of course, we resort to a discretized version of this cost:

$$c(V_i) = \sum_{j=1}^{n} \frac{1}{n_d} \sum_{k=0}^{n_d} \|\gamma_j(\frac{k}{n_d}) - \mathrm{Exp}^V_{\gamma(0)}\left(\dot{\gamma}_j(0)\frac{k}{n_d}\right)\|_2^2 \tag{5.4}$$

where $n_d \in \mathbb{N}$ is the chosen number of discretization steps.

In all the cases we ran, the cost function does decrease steadily towards 0. But not in all cases does the recovered metric converges to the fixed target metric. It is difficult in general to draw conclusion regarding the number of geodesics to be used, or how to best place them to optimize the recovered metric. However, there is a notable improvement of the estimated metrics as n increases. Figure 5.4 shows an example of a run in $\mathbb{R}^2$, where 1000 geodesics are used for the estimation.

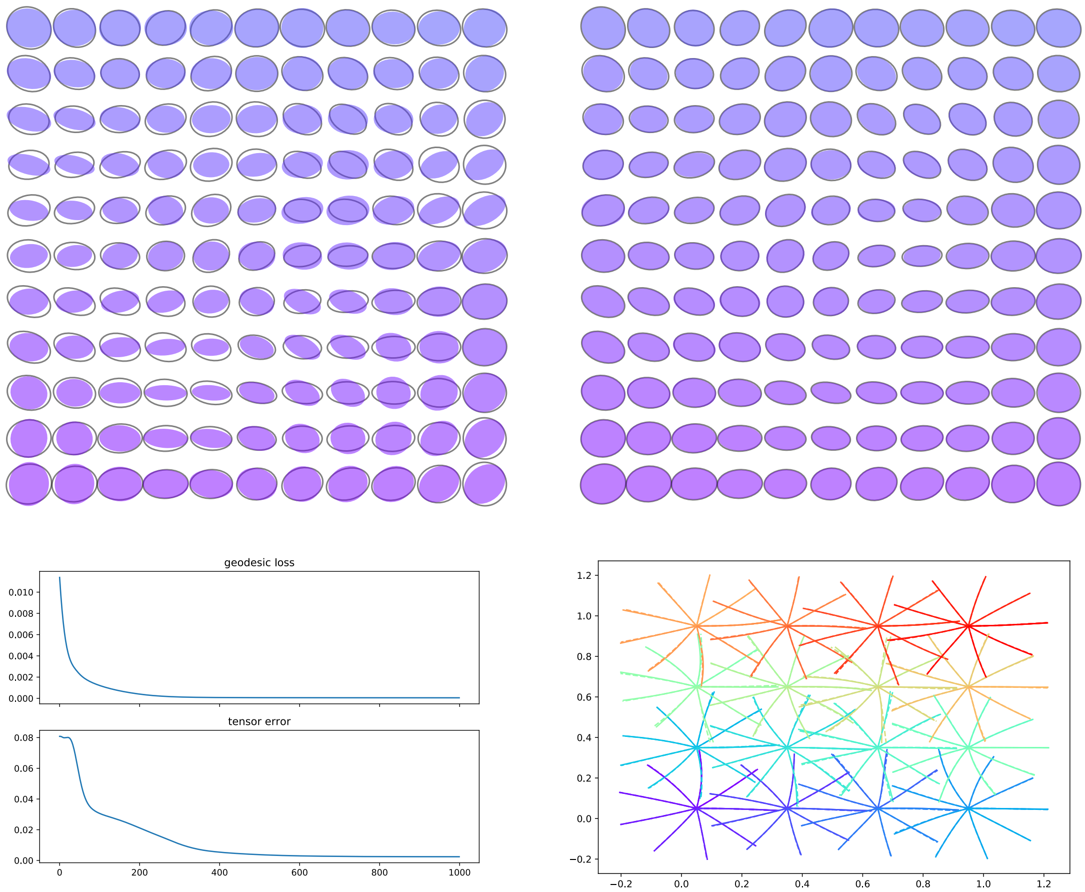


Figure 5.4: Estimation of a Riemannian metric so as to make a family of curves geodesic. **Top left**: target metric g_U (ellipses black lines) and initial random metric g_V (purple ellipses). **Top right**: target metric g_U (ellipses black lines) and estimated metric g_V after 1000 iterations of gradient descent on V (purple ellipses). **Bottom left**: geodesic reconstruction error (5.3) and tensor error (integral of the affine-invariant distance between the two metrics over $[0,1]^2$) during the estimation. **Bottom right**: the geodesics used for the estimation.

5.6 Parametric Riemannian metric learning for longitudinal disease modelling

In this section, we use a stronger criterion than criterion (C_3). Instead of asking for the geodesicity of a family of trajectories, we ask that these trajectories have a particular distribution on the manifold. Namely, we ask that they all be parallel trajectories to a common geodesic on the manifold. This can be seen as an extension of [86] where we allow the Riemannian metric to change so that exp-paralleled trajectories to a common reference geodesic do correspond best to observed patterns of progressions. It addresses one of the fundamental limitations of [86] and in general of methods optimizing linear mixed-effect models on manifolds: the need of prior knowledge about the geometry of the data and of its patterns of progression.

We are given a longitudinal data set $(t_{ij}, y_{ij})_{i=1,\dots,N, j=1,\dots,n_i}$ where the $t_{ij} \in \mathbb{R}$ are observation times and y_{ij} are observations. For $i \in \{1, \dots, N\}$, $(y_{ij})_{j=1,\dots,n_i}$ is the set of observations of the subject. We adapt a similar approach to [86] but we aim at estimating the Riemannian metric jointly with the other model parameters.

5.6.1 Original modelling

In [86], the authors assume that observations (y_{ij}) lie on a Riemannian manifold $(\mathcal{M}, g)$. They then suppose that each subject follows a trajectory which is defined by exp-parallelization of a geodesic γ i.e.:

$$y_i(t) = \mathrm{Exp}_{\gamma(\varphi_i(t))}\left(\Pi_{\gamma(t_0),\gamma(\varphi_i(t))}(w_i)\right) \tag{5.5}$$

where:

- t_0 is a reference time.
- w_i is an element of $T_{\gamma(t_0)}\mathcal{M}$.
- $\Pi_{\gamma(s)\gamma(s')}$ denotes the parallel transport along γ from $\gamma(s)$ to $\gamma(s')$.
- Exp is the Riemannian exponential.
- $\Phi_i(t)$ is the re-parametrized time for the individual: $\Phi_i(t) = e^{\eta_i}(t - \tau_i + t_0)$, allowing varying paces of progression and ages of onset for different subjects.
- γ is a reference geodesic given by $\gamma(t_0) = p_0$ and $\gamma'(t_0) = v_0$.

This shares similarities with the work already presented in Part I. As mentioned above, the parallel transport operation can be computed in an efficient way using the Fanning scheme.

The authors then formulate a mixed-effect model, where the random parameters are $(\eta_i, \tau_i, w_i)_i$ and the fixed effects are (v_0, t_0, p_0). Particular care is taken to model the prior for each variable correctly, we refer to the original paper for these details, which are not of relevance in this Chapter.

5.6.2 Proposed extension

We propose here to include the Riemannian metric g in the parameters using the parametric metrics described in equation (5.2). Doing so is straightforward, at least in principle, using the parametric metrics defined in Section 5.4, and it simply leads to an additional parameter U_i to be estimated. The rest of the modelling is unchanged and our aim is still to proceed with a Maximum A Posteriori estimation of the parameters.

The model formulated this way does not belong to the exponential family, and we resort to a modification of the original MCMC-SAEM [19] algorithm used in [88]. Note that an alternative for the estimation would be to optimize the ELBO bound exhibited in [47].

The inference consists in finding the Maximum A Posteriori (MAP) of a directed probabilistic model with latent variables u. The E step requires the computation of integrals of the form $\int_u \log\left(p(y|u,\theta)\right) p(u,\theta_k)\mathrm{d}u$ which are intractable in our case, so we resort the the Stochastic Approximation EM (SAEM) [19] which alternates:

- *Simulation.* For each observation y_i, generate u_i, a realization of the hidden variable under the posterior density $p(u|y_i,\theta_k)$.
- *Approximation.* Update $Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k(\sum_i p(y_i|u_i,\theta) - Q_{k-1}(\theta))$
- *Maximization.* Set $\theta_{k+1} = \mathrm{argmax}_\theta Q(\theta)$.

where during L burn-in iterations $\gamma_k = 1$ and then for all $k > L$, $0 \leq \gamma_k \leq 1$, $\sum_{i=1}^{\infty} \gamma_k = \infty$ and $\sum_{i=1}^{\infty} \gamma_k^2 < \infty$. Once again, this procedure is intractable in our case, since the maximization step cannot be computed at a reasonable cost. We therefore replace the Approximation step by simply setting $Q(\theta) = \sum_i p(y_i|u_i,\theta)$, which can be optimized by stochastic gradient descent. This amounts to keeping only the burn-in phase of the SAEM ($\gamma_k = 1$) which is, as we noted empirically, the most important phase with respect to space exploration of the individual variables u_i.

Simulation-Expectation To perform the Simulation step, we use the symmetric Hasting-Metropolis sampler [72], a Markov Chain Monte Carlo method. We run 25 iterations of the Markov Chain for each simulation, to limit samples correlation.

Maximization The maximization can be performed by stochastic gradient descent on Q with respect to all the parameters which do not have a closed-form update, and by closed-form update for the noise variance. We run ten epochs of gradient descent at each

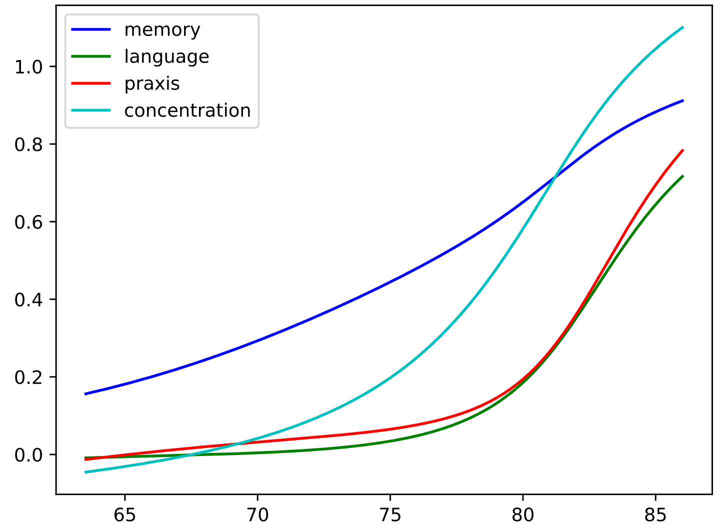


Figure 5.5: Average scenario of disease progression with the learned metric g_U. x axis is standardized time, y axis is normalized cognitive score.

maximization, using Adam [46]. This gradient descent step is reasonnably fast thanks to the multi-threaded and gpu possibilities offered by PyTorch.

5.6.3 Results

We provide here the results of the estimation of this model on the same data as in [86]: the cognitive scores measurements on MCIc subjects from the ADNI database. The goal here is to show that without prior knowledge on the patterns of progression of the score, our proposed longitudinal model still recovers a relevant scenario of progression which for instance respects the known ordering of the symptoms onset.

The average scenario –the geodesic $t \mapsto \mathrm{Exp}_{p0}(t\dot{\gamma}(v_0))$– is shown on Figure 5.5. This scenario is to be compared with Figure 2 in [86] which we reproduce on Figure 5.6. First, we recover the same ordering of the symptoms onset, which were already known in the literature. Second, we do recover patterns of progression which are similar to the ones *postulated* in [86]. However, our formulation offers much more flexibility for the mean scenario. For instance, our obtained scenario allows for intersections between the curves for each score, which is forbidden in [86].

The drawback of this method is that it requires repeated computations of Riemannian exponentials: directly and through the use of the Fanning scheme. As the dimension of the observation space $\mathbb{R}^d$ increases, the number of points x_i required to build a large family of parametric metrics in $\mathbb{R}^d$ increases exponentially, and so does the complexity of the integration of the Hamiltonian equations. In the next Chapter, we propose an alternative method, relying on deep learning techniques, to perform this Riemannian metric learning

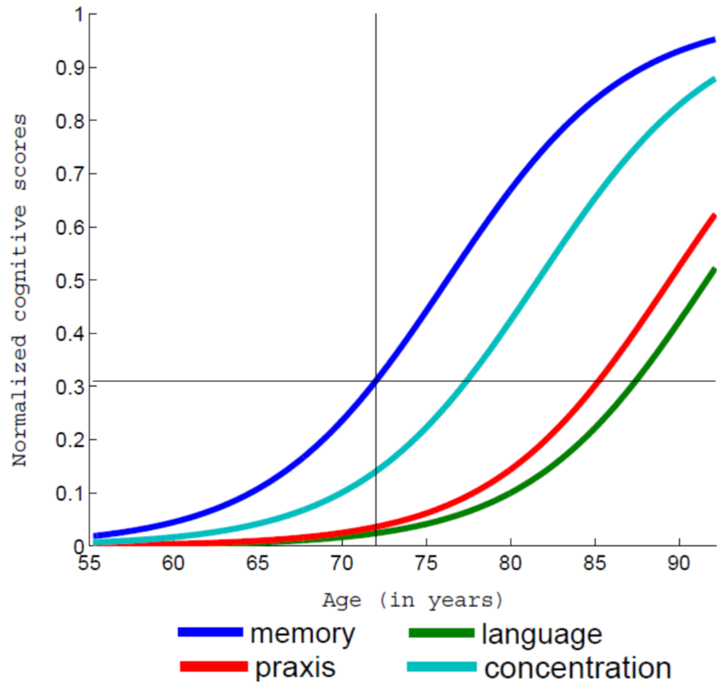


Figure 5.6: Average scenario of disease progression with fixed metric, reproduction from [86].

task in high-dimensional spaces.

Chapter 6

Disease modelling using deep neural networks [66]

This chapter is a reproduction of [66].

6.1 Introduction

The analysis of the longitudinal aspects of a disease is key to understand its progression as well as to design prognosis and early diagnostic tools. Indeed, the time dynamic of a disease is more informative than static observations of the symptoms, especially for neuro-degenerative diseases whose progression span years with early subtle changes. More specifically, we tackle in this paper the issue of disease modelling: we aim at building a time continuous reference of disease progression and at providing a low-dimensional representation of each subject encoding the subject's position with respect to this reference. This task must be achieved using longitudinal datasets, which contain repeated observations of clinical measurements or medical images of subjects over time. In particular, we aim at being able to achieve this longitudinal modelling even when dealing with very high-dimensional data.

The framework of Riemannian geometry is well suited for the analysis of longitudinal trajectories. It allows for principled approaches which respect the nature of the data - e.g. explicit constraints- and can embody some a priori knowledge. When a Riemannian manifold of interest is identified for a given type of data, it is possible to formulate generative models of progression on this manifold directly. In [86, 48, 7], the authors propose a mixed-effect model which assumes that each subject follows a trajectory which is parallel to a reference geodesic on a Riemannian manifold. In [90], a similar approach is constructed with a hierarchical model of geodesic progressions. All these approaches make use of a predefined Riemannian geometry on the space of observations.

A main limitation of these approaches is therefore the need to know this Riemannian manifold. It may be possible to coin a relevant Riemannian manifold in low-dimensional cases and with expert knowledge, but it is nearly impossible in the more general case of high-dimensional data or when multiple modalities are present. Designing a Riemannian metric is in particular very challenging, as the space of Riemannian metrics on a manifold is vastly large and complex. A first possibility, popular in the literature, is to equip a submanifold of the observation space with the metric induced from a metric on the whole space of observations –e.g. ℓ^2 on images. However, we argue here that this choice of larger metric is arbitrary and has no reason to be of particular relevance for the analysis at hand. Another possibility is to consider the space of observations as a product of simple manifolds, each equipped with a Riemannian metric. This is only possible in particular cases, and even so, the product structure constrains the shapes of the geodesics which need to be geodesics on each coordinate. Other constructions of Riemannian metrics exist in special cases, but there is no simple general procedure. Hence, there is a need for data-driven metric estimation.

There are a few Riemannian metric learning approaches do exist in the litterature. In [38], the authors propose to learn a Riemannian metric which is defined by interpolation of a finite set of tensors, and they optimize the tensors so as to separate a set of data according to known labels. This procedure is intractable as the dimension of the data increases. In [2] and in [92], the authors estimate a Riemannian metric so that an observed set of data maximizes the likelihood of a generative model. Their approaches use simple forms for the metric. Finally, in [53], the authors show how to use transformation of the observation space to pull-back a metric from a given space back to the observation space, and give a density criterion for the obtained metric and the data.

We propose in this chapter a new approach to learn a smooth manifold and a Riemannian metric which are adapted to the modelling of disease progression. We describe each subject as following a straight line trajectory parallel to a reference trajectory in a low-dimensional latent space $\mathcal{Z}$, which is mapped onto a submanifold of the observation space using a deep neural network Ψ, as seen in [89]. Using the mapping Ψ, the straight line trajectories are mapped onto geodesics of the manifold $\Psi(\mathcal{Z})$ equipped with the push-forward of the Euclidean metric on $\mathcal{Z}$. After inference, the neural network parametrizes a manifold which is close to the set of observations and a Riemannian metric on this manifold which is such that subjects follow geodesics on the obtained Riemannian manifold, which are all parallel to a common reference geodesic in the sense of [86]. This construction is based on Theorem 4 in Chapter 5 giving mild conditions under which there exists a Riemannian metric such that a family of curves are geodesics. Additionally, this particular construction of a Riemannian geometry allows very fast computations of Riemannian operations, since all of them can be done in closed form in $\mathcal{Z}$.

Section 6.2 describes the Riemannian structure considered, the model as well as the inference procedure. Section 6.3 shows the results on various features extracted from the ADNI data base [40] and illustrates the advantages of the method compared to the use of predefined simple Riemannian geometries.

6.2 Propagation model and deep generative models

6.2.1 Push-forward of a Riemannian metric

We explain here how to parametrize a family of Riemannian manifolds. We use deep neural networks, which we view as non-linear mappings, since they have the advantage of being flexible and computationally efficient.

Let $\Psi_w : \mathbb{R}^d \to \mathbb{R}^D$ be a neural network function with weights w, where $d, D \in \mathbb{N}$ with $d < D$. It is shown in [89] that if the transfer functions of the neural network are smooth monotonic and the weight matrices of each layer are of full rank, then Ψ is differentiable and its differential is of full rank d everywhere. Consequently, $\Psi(\mathbb{R}^d) = \mathcal{M}_w$ is locally a d-dimensional smooth submanifold of the space $\mathbb{R}^D$. It is only locally a submanifold: M_w could have self intersections since Ψ_w is in general not one-to-one. Note that using architectures as in [41] would ensure by construction the injectivity of Ψ_w.

A Riemannian metric on a smooth manifold is a smoothly varying inner product on the manifold tangent space. Let g be a metric on $\mathbb{R}^d$. The push-forward of g on $\mathcal{M}_w$ is defined by, for any smooth vector fields X, Y on $\Psi_w(\mathbb{R}^d)$:

$$\Psi_w^*(g)(X, Y) := g((\Psi_w)_*(X), (\Psi_w)_*(Y))$$

where $(\Psi_w)_*(X)$ and $(\Psi_w)_*(Y)$ are the pull-back of X and Y on $\mathbb{R}^d$ defined by $(\Psi_w)_*(X)(f) = X(f \circ \Psi_w^{-1})$ for any smooth function $f : \mathbb{R}^d \to \mathbb{R}$, and where Ψ_w^{-1} is a local inverse of Ψ_w, which exists by the local inversion theorem.

By definition, Ψ_w is an isometry mapping a geodesic in $(\mathbb{R}^d, g)$ onto a geodesic in $(\mathcal{M}_w, \Psi_w^*(g))$. Note that the function Ψ_w parametrizes both a submanifold $\mathcal{M}_w$ of the space of observations and a metric $\Psi_w^*(g)$ on this submanifold. In particular, there may be weights w_1, w_2 for which the manifolds $\mathcal{M}_{w_1}, \mathcal{M}_{w_2}$ are the same, but the metrics $\Psi_{w_1}^*(g), \Psi_{w_2}^*(g)$ are different.

In what follows, we denote $g_w = \Psi_w^*(g)$ the push-forward of the Euclidean metric g. Since $(\mathbb{R}^d, g)$ is flat, so is $(\mathcal{M}_w, g_w)$. This neither means that $\mathcal{M}_w$ is flat for the induced metric from the Euclidean metric on $\mathbb{R}^D$ nor that the obtained manifold is Euclidean (ruled surfaces like hyperbolic paraboloid are flat still non Euclidean). A study of the variety of Riemannian manifolds obtained under this form would allow to better understand how vast or limiting this construction is.

6.2.2 Model for longitudinal progression

We denote here $(y_{ij}, t_{ij})_{j=1,\dots,n_i}$ the observations and ages of the subject i, for $i \in \{1, \dots, N\}$ where $N \in \mathbb{N}$ is the number of subjects and $n_i \in \mathbb{N}$ is the number of observation of the i-th subject. The observations lie in a D-dimensional space $\mathcal{Y}$. We model each individual as following a straight trajectory in $\mathbb{Z} = \mathbb{R}^d$ with $d \in \mathbb{N}$:

$$l_i(t) = e^{\eta_i}(t - \tau_i)\vec{e}_1 + \sum_{j=2}^{d} b_i^j \vec{e}_j \tag{6.1}$$

where $(\vec{e}_1, \ldots, \vec{e}_d)$ is the canonical basis of $\mathbb{R}^d$.

With this writing, on average, the subjects follow a trajectory in the latent space given by the direction $\vec{e}_1$. To account for inter-subject differences in patterns of progression, each subject follows a parallel to this direction in the direction $\sum_{j=2}^{d} b_i^j \vec{e}_j$. Finally, we re-parametrize the velocity of the subjects in the $\vec{e}_1$ direction using η_i which encodes for the pace of progression and τ_i which is a time shift. This writing is so that $l_i(\tau_i)$ is in $\mathrm{Span}(e_2, \ldots, e_d)$, the set of all possible states at the time τ_i. Hence, after inference, all the subjects progression should be aligned with similar values at $t = \tau_i$. We denote, for each subject i, $\varphi_i = (\eta_i, \tau_i, b_i^2, \ldots, b_i^d) \in \mathbb{R}^{d+1}$. φ_i is a low-dimensional vector which encodes the progression of the subject.

As shown above, we map $\mathbb{Z}$ to $\mathcal{Y}$ using a deep neural network Ψ_w. The subject reconstructed trajectories $t \mapsto y_i(t) = \Psi_w(l_i(t))$ are geodesics in the submanifold $(\mathcal{M}_w, g_w)$. The geodesics are parallel in the sense of [86] and [90]. Note that the apparently constrained form of latent space trajectories (6.1) is not restrictive: the family of functions parametrized by the neural network Ψ_w allows to curve and move the image of the latent trajectories in the observation space, and for example to align the direction $\vec{e}_1$ with any direction in $\mathcal{Y}$.

6.2.3 Encoding the latent variables

To predict the variables φ for a given subject, we use a recurrent neural-network (RNN), which is to be thought as an encoder network. As noted in [15, 30], the use of a recurrent network allows to work with sequences which have variable lengths. This is a significant advantage given the heterogeneity of the number of visits in medical longitudinals studies. In practice, the observations of the subject are not regularly spaced in time. To allow the network to adapt to this, we provide the ages of the visit at each update of the RNN.

We use an Elman network, which has a hidden state $h \in \mathbb{R}^H$ with $H \in \mathbb{N}$, initialized to $h_0 = 0$ and updated along the sequence according to $h_k = \rho_h(W_h y_{ik} + U_h h_{k-1} + b_h)$ and the final value predicted by the network after a sequence of length $f \in \mathbb{N}$ is $\varphi = W_\varphi \rho_\varphi(h_f) + b_z$ where ρ_φ and ρ_h are activation functions and $W_h, U_h, W_\varphi, b_h, b_\varphi$ are the weights and biases of the network. We denote $\theta = (W_h, U_h, W_\varphi, b_\varphi)$ the parameters of the encoder.

When working with scalar data, we use this architecture directly. When working with images, we first use a convolutional neural network to extract relevant features from the images which are then fed to the RNN. In this case, both the convolutional network and the recurrent network are trained jointly by backpropagation. Figure 6.1 summarizes the

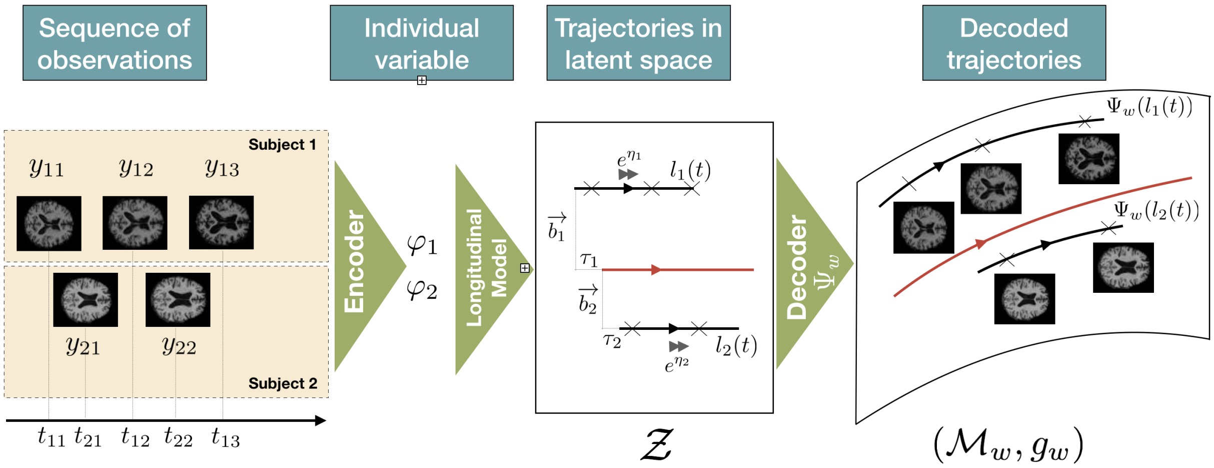


Figure 6.1: The observed sequences are encoded into latent trajectories, which are then decoded into geodesics on a submanifold of the observation space.

whole procedure.

6.2.4 Regularization

To impose some structure in the latent space $\mathbb{Z}$, we impose a regularization on the individual variables $\varphi_i = (\eta_i, \tau_i, b_i^2, \ldots, b_i^d)$. The regularization cost used is:

$$r(\eta, \tau, b^2, \ldots, b^d) = \frac{(\eta - \overline{\eta})^2}{\sigma_\eta^2} + \frac{(\tau - \overline{\tau})^2}{\sigma_\tau^2} + \sum_{j=2}^{d} (b^j)^2 \tag{6.2}$$

This regularization requires the individual variables η and τ to be close to mean values $\overline{\tau}, \overline{\eta}$. The parameters $\overline{\eta}$, $\overline{\tau}$ are estimated during the inference. $\sigma_\eta > 0$ is fixed but the estimation of $\overline{\eta}$ allows to adjust the typical variation of η with respect to the mean pace $\overline{\eta}$, while the neural network Ψ_w adjusts accordingly the actual velocity in the observation space in the $\vec{e}_1$ direction. σ_τ is set to the empirical standard deviation of the time distribution $(t_{ij})_{ij}$, meaning that we expect the delays between subjects to be of the same order as the standard deviation of the visit ages.

6.2.5 Cost function and inference

Overall, we optimize the cost function:

$$c(\theta, w, \overline{\eta}, \overline{\tau}) = \sum_i \sum_j \frac{\|y_i(t_{ij}) - y_{ij}\|_2^2}{\sigma^2} + \sum_i r(\varphi_i) \tag{6.3}$$

where $\sigma > 0$ is a parameter controlling the trade-off reconstruction/regularity.

The first term contains the requirements that the geometry $(\mathcal{M}_w, g_w)$ be adapted to the observed progressions since it requires geodesics $y_i(t)$ to be good reconstructions of

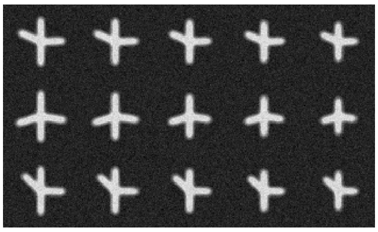


Figure 6.2: Each row represents a synthetic subject.

the individual trajectories. As shown in the Appendix, there exists solutions to problems of this kind: under mild conditions there exists a metric on a Riemannian manifold that the subjects' progressions are geodesics. But this is only a partial constraint: there is a whole class of metrics which have geodesics in common (see [71] for the analysis of metrics which have a given family of trajectories as geodesics).

We infer the parameters of the model by stochastic gradient descent using the Adam optimizer [46]. After each batch of subjects B, we balance regularity and reconstruction by imposing a closed-form update on σ:

$$\sigma^2 = \frac{1}{N_B D} \sum_{i \in B} \sum_{j=1}^{N_i} \|\Psi_w\left(l_i(t_{ij})\right) - y_{ij}\|_2^2 \tag{6.4}$$

where $N_B = \sum_{i \in B} N_i$ is the total number of observations in the batch b. This automatic update of the trade-off criterion σ is inspired from Bayesian generative models which estimate the variance of the residual noise, as in e.g. [86, 111].

6.3 Experimental results

The neural network architectures and the source code for these experiments is available at `https://gitlab.com/maxime.louis.x2012/unsupervised_geometric_longitudinal`, tag IPMI 2019. For all experiments, the ages of the subjects are first normalized to have zero mean and unit variance. This allows the positions in the latent space to remain close to 0. We set $\sigma_\eta = 0.5$ and initialize $\overline{\eta}$ to 0.

6.3.1 On a synthetic set of images

To validate the proposed methodology, we first perform a set of experiments on a synthetic data set. We generate 64×64 gray level images of a white cross on a black background. Each cross is described by the arms length and angles. We prescribe a mean scenario of progression for the arm lengths and sample the arm angles for each subject from a zero-centered normal distribution. Figure 6.2 shows subjects generated using this procedure. Note that with this setting, an image varies smoothly with respect to the arms lengths

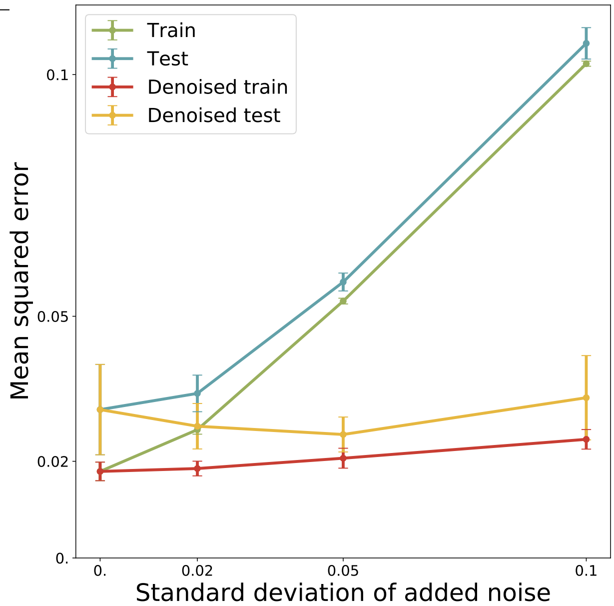


Figure 6.3: Reconstruction error on train and test sets and on denoised train and test sets, unseen during training.

and angles and hence the whole set of generated images is a smooth manifold.

We generate 10 training sets of 150 subjects and 10 test sets of 50 subjects. The number of observation of each subject is randomly sampled from a Poisson distribution with mean 5. The times at which the subject are observed are equally spaced within a randomly selected time window. We add different level of white noise on the images. We then run the inference on the 10 training sets for each level of noise. We set the dimension of the latent space $\mathbb{Z}$ to 3 for all the experiments.

For each run, we estimate the reconstruction error on both training set and test set, as well as the reconstruction error to the de-noised images, which were not seen during training. Results are shown on Figure 6.3. The model generalizes well to unseen data and successfully denoises the images, with a reconstruction error on the denoised images which does not vary with the scale of the added white noise. This means that the generated manifold of images is close to the manifold on which the original images lie. Besides, as shown on Figure 6.4, the scenario of progression along the $\vec{e}_1$ direction is well captured, while orthogonal directions $\vec{e}_2, \vec{e}_3$ allow to change the arm positions. Finally, we compare the individual variables (b_i^2, b_i^3) to the known arms angles which were used to generate the images. Figure 6.5 shows the results: the latent space is structured in a way that is faithful to the original arm angles.

6.3.2 On cognitive scores

We use the cognitive scores grading the subjects memory, praxis, language and concentration from the ADNI database as in [86]. Each score is renormalized to vary between 0 and 1, with large values indicating poor performances for the task. Overall, the data

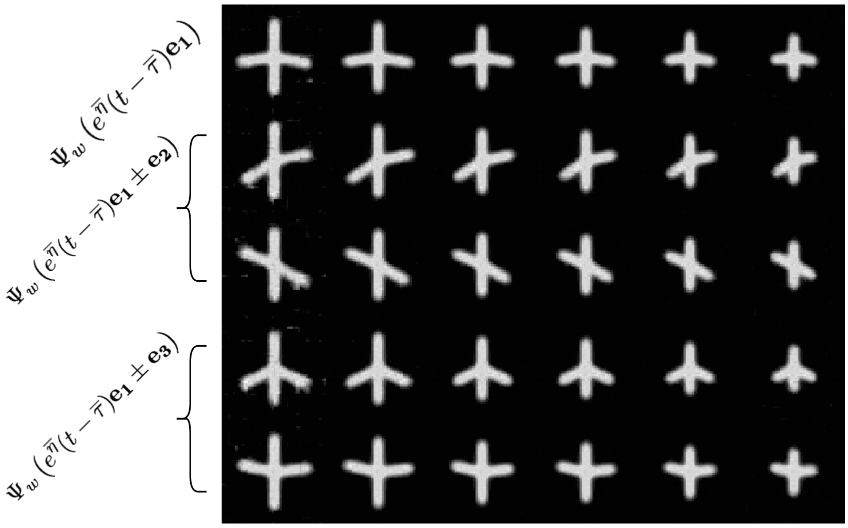


Figure 6.4: First row is $t \mapsto \Psi_\theta(\vec{e}_1 t)$. Following rows are $t \mapsto \Psi_\theta(\vec{e}_1 t + \vec{e}_i)$ for $i \in \{2,3\}$. These parallel directions of progression show the same arm length reduction with different arm positions.

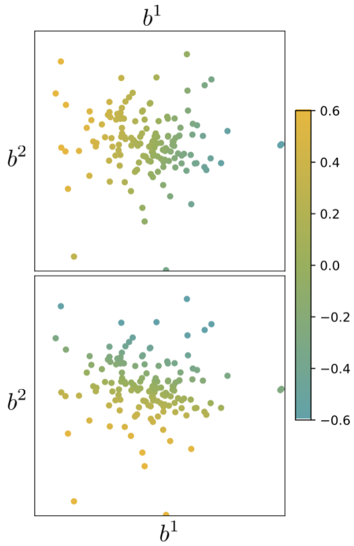


Figure 6.5: Individual variables b_i^1 and b_i^2 colored by left (top) and right (bottom) arm angle.

set consists of 223 subjects observed 6 times on average over 2.9 years. We perform a 10-fold estimation of the model. The measured mean squared reconstruction error is $0.079 \pm 1.1\text{e}-3$ on the train sets, while it is of $0.085 \pm 1.5\text{e}-3$ for the test sets. Both are close to the uncertainty in the estimation of these cognitive scores [95]. This illustrates the ability of the model to generalize to unseen observations. First, this indicates that M_w is a submanifold which is close to all the observations. Second, it indicates how relevant the learned Riemannian metric is, since unobserved subject trajectories are very close to geodesics on the Riemannian manifold.

Figure 6.6 shows obtained average trajectories $t \mapsto \Psi_w(e^{\overline{\eta}} e_1(t-\overline{\tau}))$ for a 10-fold estimation of the model on the data set, with $\mathcal{Z}$ dimension set to 2. All of these trajectories are brought back to a common time reference frame using the estimated $\overline{\tau}$ and $\overline{\eta}$. All average trajectories are very similar, underlining the stability of the proposed method. Note that despite the small average observation time of the subjects, the method proposed here allows to robustly obtain a mean scenario of progression over 30 years. Hence, despite all the flexibility that is provided through the different neural networks and the individual parameters, the model still exhibits a low variance. Besides, the obtained average trajectory here should be compared to 5.5 obtained using the parametric metric 5.4. The proximity of the results is encouraging for the approach developed here.

We compare the results to the mean trajectory estimated by the model [86], which is shown in Figure 6.8. Both cases recover the expected order of onset of the different cognitive symptoms in the disease progression. Note that with our model the progression of the concentration score is much faster than the progression of the memory score, although it starts later: this type of behaviour is not allowed with the model described in [86] where the family of possible trajectories is much narrower. Indeed, because it is difficult to craft a relevant Riemannian metric for the data at hand, the authors modelled the data as lying on a product Riemannian manifold. In this case, a geodesic on the product manifold is a product of geodesics of each manifold. This strongly restricts the type of dynamics spanned by the model and hence gives it a high bias.

The use of the product manifold also has an impact on the parallel variations around the mean scenario: they can only delay and slow/accelerate one of the component with respect to another, as shown on Figure 6.8. Figure 6.7 illustrates the parallel variations $\Psi(\overline{\alpha}e_0(t-\overline{\tau})+\vec{e}_1)$ one can observe with the proposed model. The variation is less trivial since complex interactions between each features are possible. In particular, the concentration score varies more in the early stages of the disease than in the late stages.

The individual variables φ. To show that the individual variables φ_i did capture information regarding the disease progression, we compare the distribution of the τ_i between subjects who have at least one APOE4 allele of the APOE gene -an allele known to be a implicated in Alzheimer's disease- and subjects which have no APOE4 allele of this gene. We perform a Mann-Whitney test on the distributions to see if they differ. For all folds, a p-value lower than 5% is obtained. For all folds, carriers have a larger τ meaning that they have an earlier disease onset than non-carriers. This is in accordance with [6]. Similarly, women have on average an earlier disease onset for all folds, in accordance with [52].

A closer look at the geometry

We look at the obtained Riemannian geometry by computing the latent position best mapped onto each of the observation by Ψ_w. We then plot the obtained latent positions

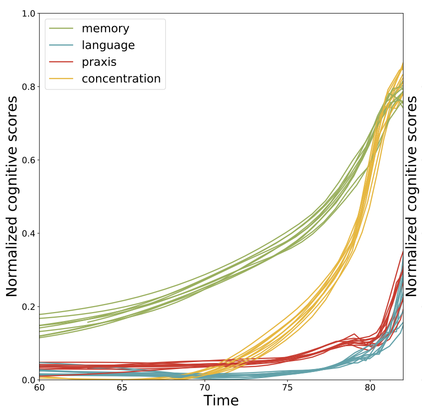


Figure 6.6: Learned main progression of the cognitive scores, for the 10-fold estimation.

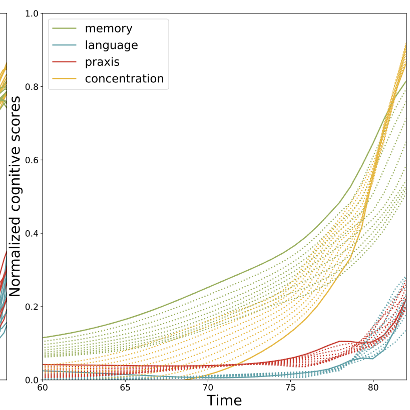


Figure 6.7: Mean geodesic of progression and parallel variations $t \mapsto \Psi(e^{\overline{\eta}}e_0(t-\overline{\tau})+\lambda\vec{e}_1)$ for varying λ.

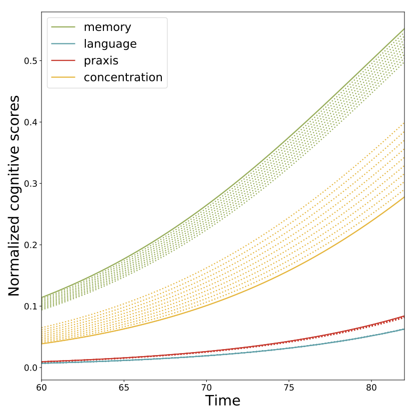


Figure 6.8: Mean geodesic of progression and parallel variations for the model with user-defined metric.

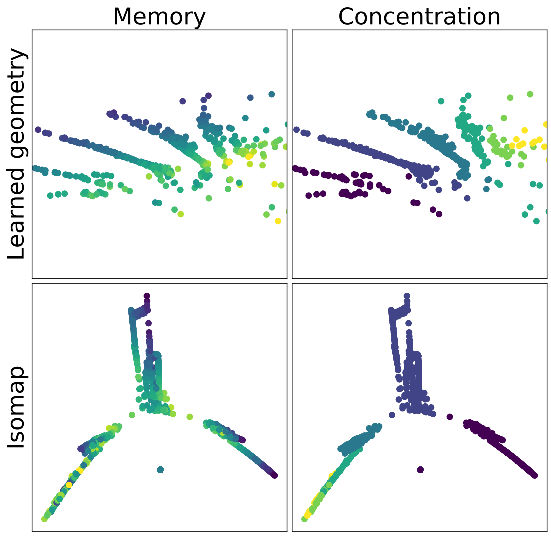


Figure 6.9: Top (resp. Bottom) latent positions (resp. Isomap embedding) of the observations colored by memory and concentration score.

to look at the structure of the learned Riemannian manifold. We compare the obtained structure with a visualisation of the structure induced by the ℓ^2 on the space of observations produced by Isomap [96]. Isomap is a manifold learning technique which attempts to reconstruct in low dimensions the geodesic distances computed from a set of observations. The results are shown in Figure 6.9. The geometry obtained after inference is clearly much more suited for disease progression modelling. Indeed, the $\vec{e}_1$ direction does correspond to typical increases in the different scores. The induced metric is not as adapted. This highlights the relevance of the learned geometry for disease modelling.

6.3.3 On anatomical MRIs

We propose here an estimation of the model on 3D MRIs preprocessed from the ADNI database, to check the behaviour of the proposed method in high dimension. We select subjects which ultimately convert to Alzheimer's disease. We obtain 325 subjects and a total of 1996 MRIs which we affinely align to the Colin-27 template and resample to 64x64x64. We run a 5-fold estimation of the model with $\dim \mathcal{Z} = 10$, using a GPU backend. We obtain a train mean squared error of $0.002 \pm 1e-5$ and a test mean squared error of $0.0024 \pm 3e-5$. Figure 6.10 shows one of the learned mean trajectory.

Once the model is estimated, we compare the distributions of the pace of progressions η_i between the individuals who have at least one APOE4 allele of the APOE gene and the individuals who have no APOE4 allele. For all 5 folds, the distributions of the paces of progression significantly differ, with p-values lower than 5% and in each case, the APOE4 carriers have a greater pace of progression, in accordance with [6]. The same analysis

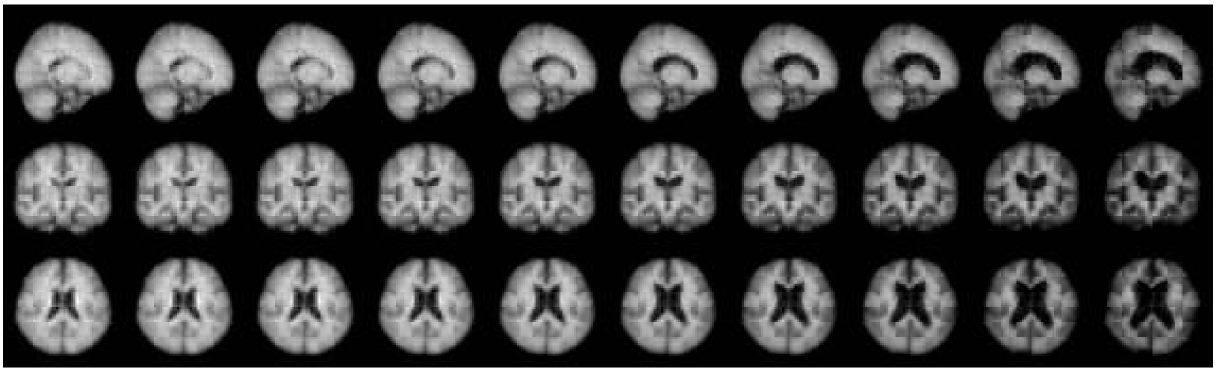


Figure 6.10: $t \mapsto \Psi_\theta(e_0 t)$ on the MRI dataset. The growth of the ventricles, caracteristic of aging and Alzheimer's disease is clearly visible.

between the individuals who have two APOE4 allele versus the individuals which have at maximum one APOE4 allele shows a significative difference for all folds for the τ variable: the APOE4 carriers have an earlier disease onset, as shown in [52]. This analysis further shows the value of the individual variables φ_i learned for each subject.

6.4 Conclusion

We presented a way to perform Riemannian geometry learning in the context of longitudinal trajectory analysis. We showed that we could build a local Riemannian submanifold of the space of observations which is so that each subject follows a geodesic parallel to a main geodesic of progression on this manifold. We illustrated how the encoding of each subject into these trajectories is informative of the disease progression. The latent space $\mathbb{Z}$ built in this process is a low-dimensional description of the disease progression. There are several possible continuations of this work. First, there is the possibility to conduct the same analysis on multiple modalities of data simultaneously. Then, after estimation, the latent space $\mathbb{Z}$ could be useful to perform classification and prediction tasks.

Chapter 7

Longitudinal auto-encoder for multimodal disease progression modeling [14]

This chapter is a reproduction of [14]. It is an extension of the previous chapter to multivariate data. We drop the Riemannian interpretation in this case.

7.1 Introduction

The longitudinal pattern of progression of a disease contains more information than a static observation. Leveraging this information is a key problem in machine learning for healthcare, complicated by to the nature of clinical data sets. These data sets may contain very heterogeneous observations from various modalities of subjects at multiple time points, such as clinical scores, imaging and biological samples. They include missing values, often by design: not all modalities are observed at each visit. Besides, the number of observations and their time spacing vary between subjects. For these reasons, the analysis of multiple modalities and their time dynamic at once is a challenging task.

Linear mixed effect model estimated via EM and their extension to the non-linear case [51, 56] were developed for the analysis of uni-modal longitudinal data. More recently, recurrent auto-encoder [84, 94] offer a way to encode trajectories into a low-dimensional embedding, allowing to perform unsupervised clustering of the trajectories [24]. Riemannian geometry based approaches such as [86, 66] offer ways to learn sub-manifolds of the observation space with a system of coordinate adapted to the progression of the modality observed in the data.

On the other hand, various unsupervised methods exist to fuse information from multiple modalities but from a single time snapshot. In [12, 77], the authors propose to learn a common embedding for multiple modalities auto-encoding, merging the information from all modalities and allowing the generation of missing modalities. In [76], unsupervised features are learned from heterogeneous health data as a dimensionality reduction method before machine learning tasks.

In [104], combining time and multi-modal approaches, the authors propose a setting for multi-modal time-series embedding. But their design does not handle missing modalities, common in clinical data sets. Besides, the fusion of the information from the different modalities is done at each time step and not on the progression pattern globally, thus decreasing the importance of the dynamics of each modality in the encoding.

To address these limitations, we propose a new setting for longitudinal multi-modal encoding. We extend to the multi-modal case the approach of [66]. Each modality is first separately encoded using a recurrent neural network. A fusion network is then used to merge the obtained representations into a unique representation, which describes the multi-modal trajectory of the subject as a time-parametrized linear trajectory in a latent space $\mathcal{Z}$. Then, this trajectory is decoded using a different neural network for each modality, which generates continuously varying trajectories of data changes. This setting allows to handle multiple modalities even when not all of them are observed at each visit and it can handle any number of visits and any time spacing between the visits. Finally, extrapolation in the latent space allows for prediction of the future of each modality and we show on a synthetic data set and on the ADNI database using cognitive scores and MRI jointly that the predictive power is enhanced by the fusion of each modality embeddings.

In section 7.2 we explain the proposed model, in section 7.3 we present experimental results highlighting the stability of the method on synthetic and real data sets and we show how the information from one modality that contributes to the encoding allows to refine prediction of the future of another modality.

7.2 Methods

We use a longitudinal data set which contains repeated observations of subjects, where the observations at each time point contain a various combination of modalities among $M \in \mathbb{N}$ modalities. For any subject $i \in \{1, \ldots, N\}$ where $N \in \mathbb{N}$ and for any modality $m \in \{1, \ldots, M\}$, we have a sequence $(y_{ij}^m, t_{ij}^m)_{j=1,\ldots,n_i^m}$ of observations y_{ij}^m of observed at times t_{ij}^m.

7.2.1 Decoding : Non linear mixed effect model

We set $d \in \mathbb{N}$ and consider a d-dimensional latent space $\mathcal{Z} = \mathbb{R}^d$ and its canonical basis $(\vec{e}_i)_{i=1,\ldots,d}$. Then, in the spirit of random slopes and intercepts models, we consider trajectories in $\mathbb{Z}$ of the form $l(t) = e^{\eta}(t - \tau)\vec{e}_1 + \sum_{i=2}^{d} \lambda^i \vec{e}_i$ where $\eta, \tau, , \lambda_2, \ldots, \lambda_d \in \mathbb{R}$ are random variables. These trajectories progress in the $\vec{e}_1$ direction and are translated in any direction orthogonal to $\vec{e}_1$, so that the λs play the role of random intercepts. η controls the pace of progression while τ allows for a time shift between the trajectories. We consider that the i-th subject follows a trajectory of this form with parameters $\varphi_i = \left(\eta_i, \tau_i, \lambda_i^2, \ldots, \lambda_i^d\right)$.

For each considered modality m, we consider a nonlinear mapping Ψ_{w_m} which maps $\mathbb{Z}$ on a subspace of the m-th modality observation space. This transports the mixed-effect model formulated in $\mathbb{Z}$ into the corresponding observation spaces. Note that the apparent rigidity of the family of trajectories considered in $\mathbb{Z}$ is not restrictive provided the mappings Ψ_{w_m} are flexible enough. In practice, the Ψ_{w_m} are neural networks, deconvolutional for images and fully connected for scalars. The right half of Figure 7.1 illustrates the procedure. Overall, this setting can be viewed as a non-linear mixed-effect model where the random effects are the φ_i's and the fixed effects are the parameters of the mappings Ψ_{w_m}.

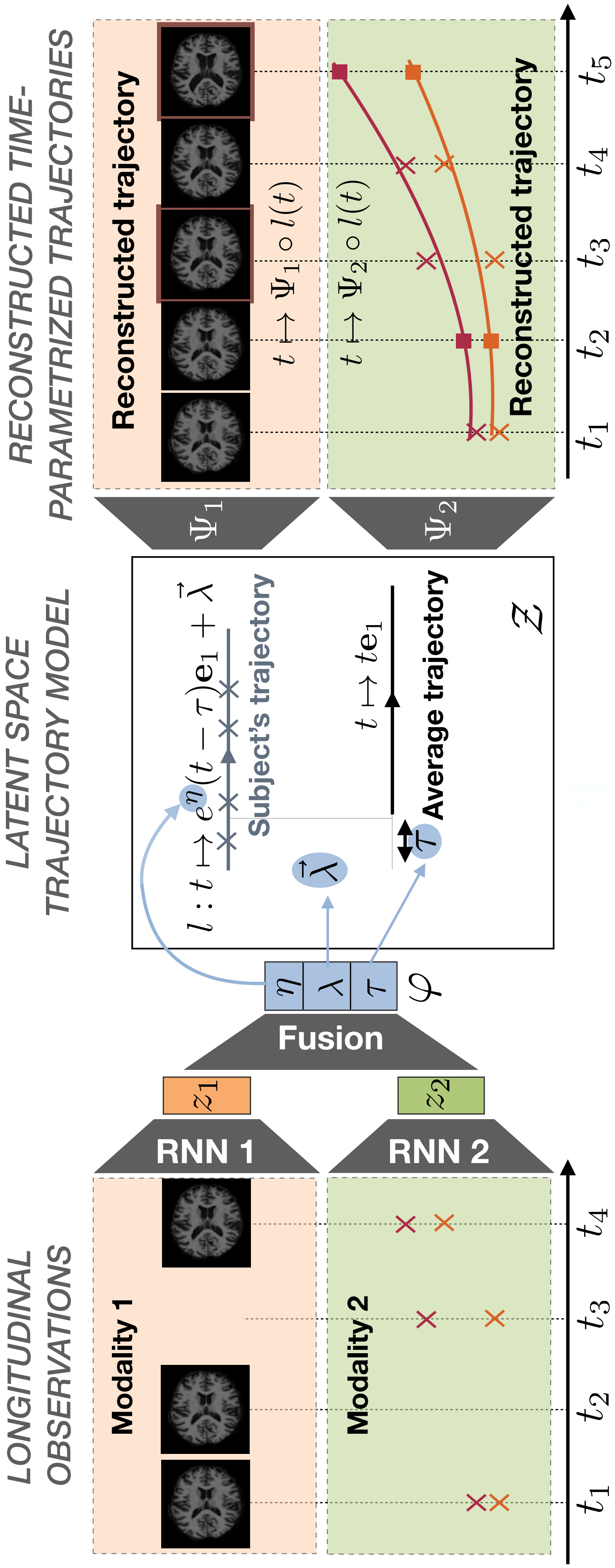


Figure 7.1: Description of the proposed longitudinal auto-encoder.

7.2.2 Encoding

Individual parameters φ_i are estimated via the use of an encoder network. More precisely, each modality is first processed by a dedicated Recurrent Neural Network (RNN), to get modality-wise representations. To correct for the varying spacings between the observations, we provide to the RNN the visit times, previously normalized to zero-mean and unit variance.

We then concatenate the obtained representations, and use a fully-connected network to merge the representations. The given architecture alllows fast inference for new subjects, and is trainable end to end. Besides, the fusion operation is learned so as to produce a single vector which contains the most information about the reconstruction of the whole sequences of all the modalities. The left part of Figure 7.1 illustrates the procedure.

7.2.3 Regularization, cost function and optimization

To enforce some structure in the latent space and in the family of trajectories obtained, we set the following regularization on the individual variable Φ_i: $r(\eta, \tau, (\lambda_i)_{i=2,\dots,d}) = \eta^2 + \tau^2 + \sum_{j=2}^{d} (\lambda^j)^2$. This regularization models the η variable to be distributed along a zero-centered normal distribution, which allows the pace of progression to vary typically between 0.2 and 5. times the mean velocity. The τ variable is regularized the same way. This regularization is not arbitrary: during each run, the observation times t_{ij}^m are rescaled to zero-mean unit variance, and thus τ can handle delays between subjects of order the standard deviation of the observation ages.

Overall, the optimized cost function for one subject is the regularization cost added to the ℓ^2 reconstruction cost summed over all modalities:

$$C\left((w_m)_m, \eta, \tau, (\lambda_i)_i\right) = r\left(\eta, \tau, (\lambda_i)_i\right) + \sum_m \frac{1}{\sigma_m^2} \sum_{j=1}^{n_i^m} \|y_{ij}^m - \Psi_{w_m}(l_i(t_{ij}^m))\|_2^2 \tag{7.1}$$

where the $(\sigma_m)_m$ are trade-off parameters between each modality and the regularization. We set an automatic update rule for these parameters after each batch by setting them to the empirical quadratic errors in reconstruction for the modality over the batch. The estimation is achieved by stochastic gradient descent with the Adam optimizer [46] and a batch size of 32 subjects. The Decoders are either fully connected or de-convolution networks depending on the kind of modality considered, with standard architectures. The encoders are either Elman networks or Elman networks working on features extracted using a convolution network in the case of images. All networks are trained end to end using back-propagation and the PyTorch library. A complete code to reproduce these experiments will be released upon publication of the paper.

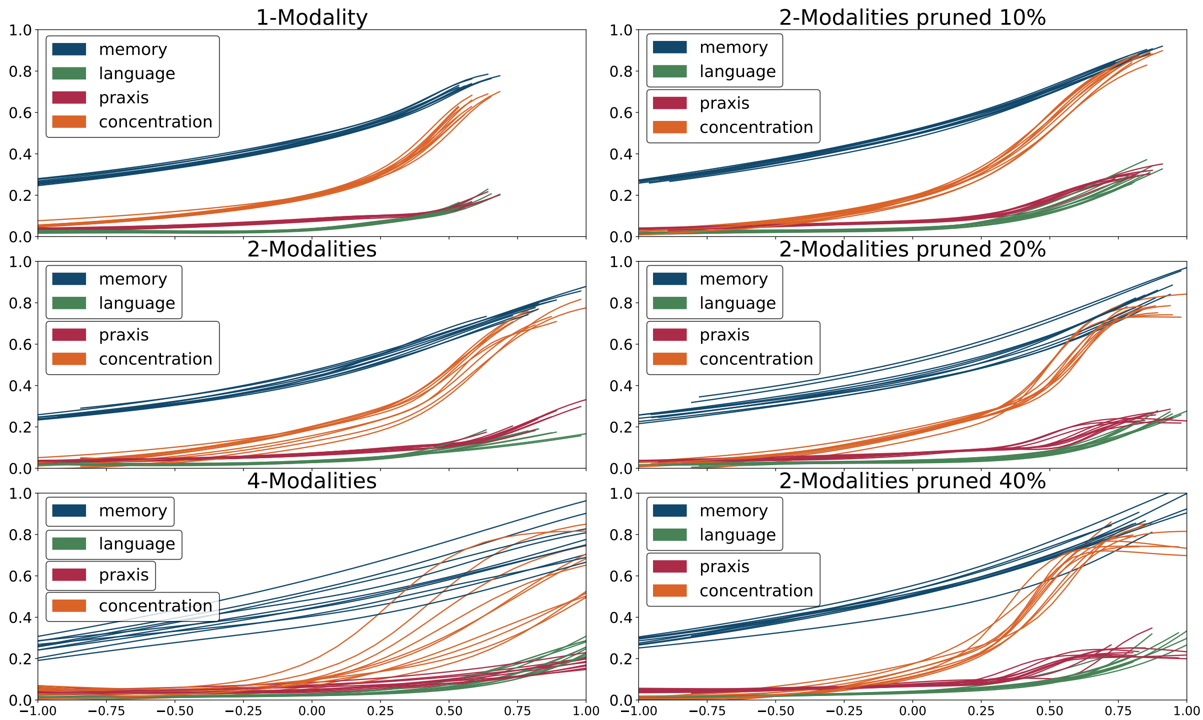


Figure 7.2: Left: average trajectories for the 10 folds, with increasing partitioning of the input features. Right: average trajectories for the 10 folds, with increasing pruning of the praxis+concentration modality.

7.3 Experimental results

7.3.1 Cognitive scores: proof of concept

As in [86], we apply our model on repeated measurement of 4 normalized cognitive scores extracted from the ADNI cohort, respectively associated with memory, language, praxis and concentration. We include the 248 MCI-converter subjects, followed for an average of 3 years, over 6 visits. We conduct 2 experiments in order to assess the robustness of the method, and report estimated average trajectories in Figure 7.2, as well as individual reconstruction errors in Table.7.1, computed from a patient-wise 10-fold cross validation.

First, we apply our model on an increasing partitioning of input feature. We consider 3 cases: selecting all scores at once as one modality, selecting separately memory+language and praxis+concentration as two modalities, and selecting each one separately. We note the overall good stability of the average model over multiple multi-modal architectures, with stability decreasing in the 4-modalities scenario, arguing for a concatenation of the consistent features.

In our second experiment, we assess the robustness of the model with the number of visits per subjects. To this end we consider the 2-modalities scenario, and perform a pruning of the data set, removing an increasing number of visits of the second modality, i.e. praxis+concentration per subjects. Data sets are obtained from pruning frequencies of respectively 10%, 20% and 40%. Here we also observe an overall good stability of the average trajectory over pruning frequency.

	Partitioning			Pruning		
	1-mod	2-mod	4-mod	2-mod 10%	2-mod 20%	2-mod 40%
Train (x10^{-3})	6.7	3.8 / 9.7	21.1 / 2.2 / 5.6 / 5.3	4.9 / 11.3	4.1 / 11.5	4.5 / 14.6
Test (x10^{-3})	7.8	5.1 / 10.6	24.7 / 3.3 / 7.1 / 5.2	4.9 / 11.7	5.0 / 11.9	5.4 / 15.5

Table 7.1: Mean 10-fold reconstruction error for the 2 cognitive scores experiments for each modality respectively

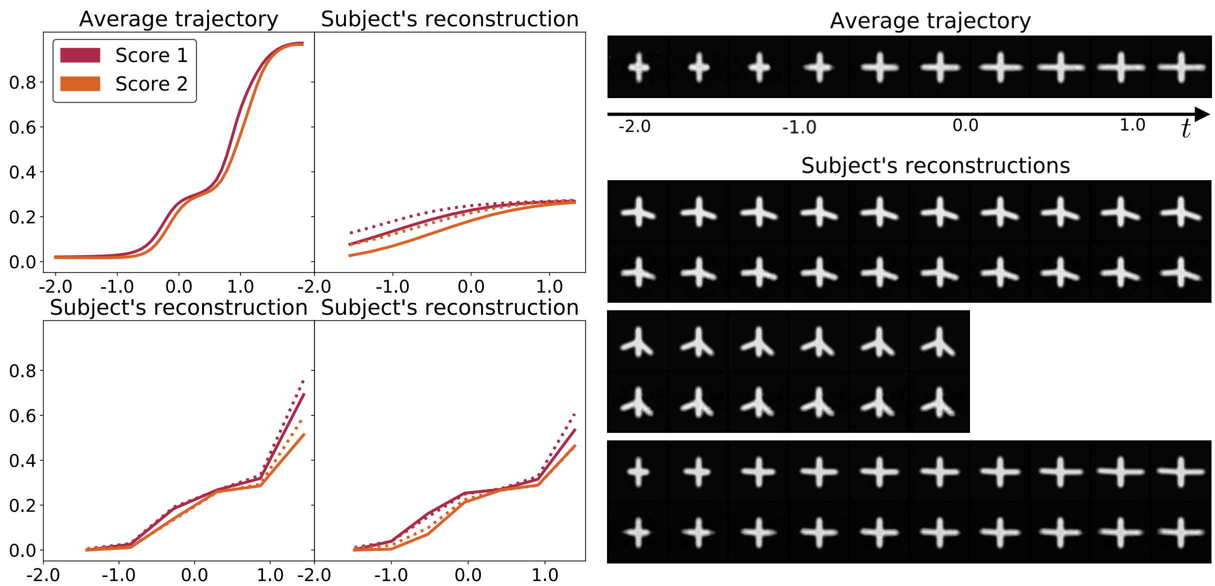


Figure 7.3: Left: average trajectory and reconstruction examples for the scalar data. Right: average trajectory and some reconstructions for the image data. Top rows are real data, bottom rows are reconstructions.

7.3.2 A synthetic data set

To test the proposed setup in realistic conditions, we generate a synthetic multi-modal data set comprising 300 subjects observed 7 times on average. The first modality is a 2D image of a cross, with varying arm lengths and angles while the second modality consists of two scores with a sigmoid-like growth. We set a time reparametrization function s with parameters a_1, a_2 defined by: $s_{a,b}(t) = t + a\text{sign}(t)t^2 + bt^3$. To generate an individual, we sample two sets of parameters $(a_k, b_k)_{k=1,2}$. These serve to reparametrize a scenario of score increase: the k-th score for the subject at time t is given by $\sigma \circ s_{a_k,b_k}$ where σ is the sigmoid function. Then, the arms lengths L_1, L_2 for the images of the subject at time t are given by $L_1 = \sigma \circ s_{(a_2-a_1)+\varepsilon_{a1},(b_2-b_1)+\varepsilon_{b1}}$, $L_2 = \sigma \circ s_{(a_2+a_1)+\varepsilon_{a2},(b_2+b_1)+\varepsilon_{b2}}$ where the ε are samples from a zero-mean normal distribution and constant with time. Finally, the arm angles are sampled along a normal distribution but are not informative of the synthetic disease process. This design is so that the images contain, in an intricate way, information about the progression of the scores materialized through the a_1, a_2, b_1, b_2 variables. The two modalities are different noisy facets of a common underlying process.

We perform a patient-wise 10-fold estimation of the model this data set. Figure 7.3 shows the obtained average trajectory for the first fold, as well as the reconstructions of some subjects images and scores observations. We evaluate and average for all folds the test and train reconstruction errors. For the cross, the test error is $2.0\ 10^{-8} \pm 8.\ 10^{-9}$ while the train error is $1.7\ 10^{-8} \pm 3.9\ 10^{-9}$. For the scores, the test error is $7.\ 10^{-3} \pm 3.\ 10^{-3}$ while the train error is $7.\ 10^{-3} \pm 3.\ 10^{-3}$. This shows that the model generalizes well to unseen data.

We use the trained model to predict the future scores on the test data. We do so by decoding the extrapolation of the latent trajectory encoded by the model. We repeat this experiment by gradually removing the last observations of the image modality, to look at the impact of this modality on the predictive power of the model. Figure 7.4 shows the experimental setup and the results. As the time span of the observed images shrinks, the prediction deteriorates: when more image data is available, the score prediction is more accurate. This shows the ability of the model to find a relevant common representation for the progressions of the different modalities.

7.3.3 Application to Alzheimer's disease future image prediction

On the 248 patients of section 7.3.1, we apply the same model on the 217 that have at least 1 MRI observation, leading to a total of 1199 cognitive scores measurements and 1441 MRIs. We work on both the MRI images and the cognitive scores. The MRI images are rigidly aligned and sub-sampled to 64^3 resolution. Note that the subjects do not have both the MRI and the cognitive scores measurements at each visit.

Figure 7.5 shows one of the estimated average trajectory for the MRI modality. We evaluate and average for all folds the test and train reconstruction errors on both modalities. For the MRI, the test error is $2.5\ 10^{-3} \pm 6.\ 10^{-5}$ while the train error is $2.4\ 10^{-3} \pm 2.\ 10^{-5}$. For the scores, the test error is $2.2\ 10^{-2} \pm 3.\ 10^{-3}$ while the train error is $1.7\ 10^{-3} \pm 6.\ 10^{-4}$. This shows that the model generalizes well to unseen data.

We then perform the same prediction task as in the previous section: we attempt to predict the future MRI from past data, using a variable amount of score data in the past. Figure 7.5 shows the prediction errors for different time horizon. Once again, the errors

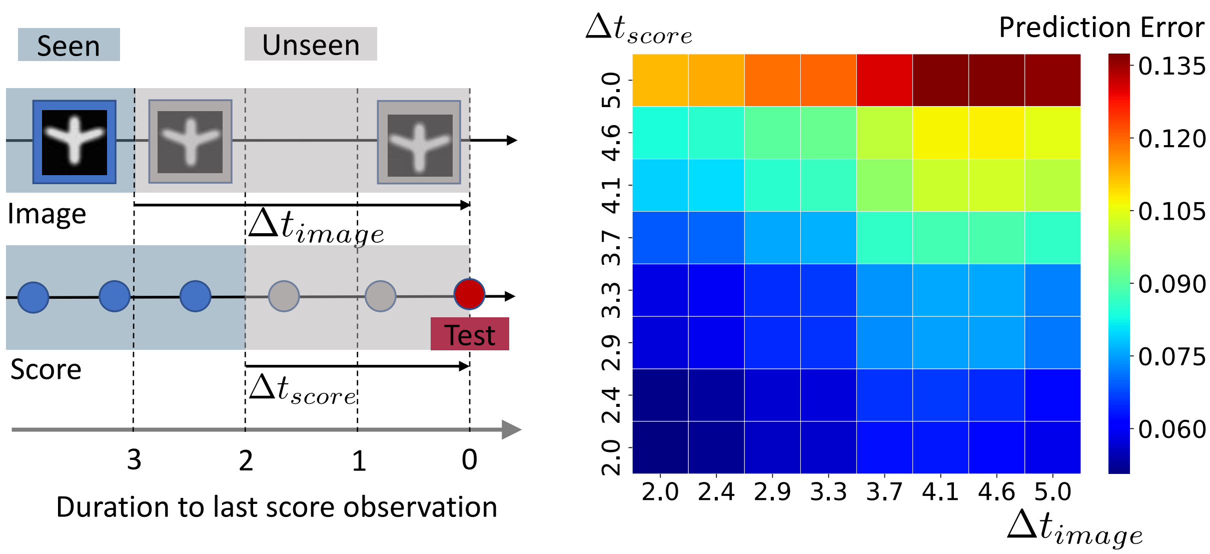


Figure 7.4: Left: description of the prediction setup. Right: the MRI prediction errors.

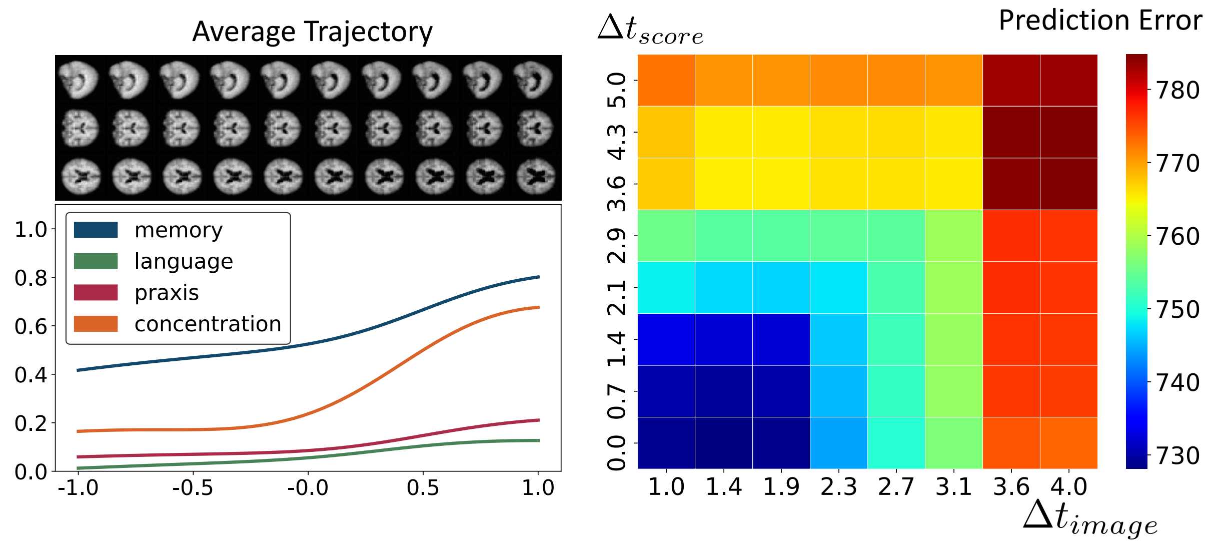


Figure 7.5: Left: average trajectory. Right: prediction error, in the same setup as in Section 7.3.2

increase as we feed the model with less cognitive scores measurements. This shows that the model captures information contained in the cognitive scores progression to refine the MRI prediction.

7.4 Conclusion and perspectives

We extended on a deep auto-encoder architecture with a mixed effect latent space to propose a practical framework for modeling multi-modal longitudinal data, trainable end-to-end. This allows for analysis of heterogeneous longitudinal data sets, deriving a model-wise average trajectory, as well as condensed patient representations. We study its robustness toward modalities partitioning and data set pruning and illustrate its utility in both synthetic and real scenarios. In the future we plan to model the progression of more modalities at once.

Part IV

Conclusion and perspectives

In this PhD, we proposed new tools for the analysis of manifold-valued data: a numerical scheme to compute the parallel transport along geodesics and a generalization of Linear Discriminant Analysis to manifold-valued data. Then, we tackled the problem of learning a manifold and a Riemannian metric on this manifold so as to optimize a given criterion. We first studied, theoretically and experimentally, simple criteria constraining Riemannian geometries. Building from this, we proposed a new way to embed longitudinal trajectories into a low-dimensional Riemannian sub-manifold of the observation space.

There are several paths to continue this work. First, there remains a lot of unknowns in Riemannian metric learning. Our results for the metrics which are optimal for some criteria are only partial, and it is in general difficult to measure exactly what can actually be learned without ambiguity, and where there remains a sort of gauge freedom on the optimal metrics. Besides, the final proposed model for longitudinal modeling relies on neural networks whose behavior is hard to constrain. In particular, it is difficult to enforce injectivity of such generative networks in practice and this is a threat to the principled Riemannian geometric framework that we use. In addition, the Riemannian manifolds that are learned with the model described in Chapter 6 are flat. It would be interesting to understand exactly how limiting that is.

Second, in the first part, we relied heavily on the LDDMM framework to model the observed shapes and images. The classical construction of families of diffeomorphisms using this framework is manual: the dynamics of the shapes is hand-prescribed by the user, and in general the family of diffeomorphisms considered remain high-dimensional. There is a need there to adapt this framework to the data, and for instance to learn a parametric family of diffeomorphisms best adapted to the data. This can be seen as a particular instance of Riemannian metric and manifold learning for shape analysis through diffeomorphisms. First steps in this direction have been undertaken in [8].

Finally, the longitudinal model developed in Chapter 6 could be further studied and applied to new data sets. In particular, we would need to look further into its ability to predict the progression of incoming subjects. Similarly, we would like to extend this model to the case when the observed population is heterogeneous in disease status, to get closer to real-world applications.

Part V

Appendix

Chapter 8

Riemannian geometry

We recall the notions of Riemannian geometry that are needed to follow the discussions in this thesis. [70] is a reference for this content.

8.1 Manifold, tangent vector, metric.

Definition 4. *A $\mathcal{C}^k$-differentiable manifold is a set $\mathcal{M}$ which is Haussdorf, second-countable, and such that there exists $n \in \mathbb{N}^*$ and a set of $(U_\alpha, \Phi_\alpha)_\alpha$ where $(U_\alpha)_\alpha$ is an open cover of $\mathcal{M}$ and for all α, $\Phi_\alpha : U_\alpha \to \Phi(U_\alpha) \subset \mathbb{R}^n$ is a homeomorphism. We also assume that for all α, β, $\Phi_\alpha \circ \Phi_\beta^{-1}$ is of class $\mathcal{C}^k$.*

In what follows, $\mathcal{M}$ is a smooth-differentiable manifold of dimension $n \in \mathbb{N}^*$. The definition of continuity and differentiability of functions defined on a manifold and of manifold-valued functions f is treated by looking at functions of the form $f \circ \Phi_\alpha$ or $\Phi_\alpha \circ f$.

Definition 5. *Let $f : \mathcal{M} \to \mathbb{R}$. We say that f is of class $\mathcal{C}^l$ with $l \in \mathbb{N}$ if for all α, $f \circ \Phi_\alpha^{-1}$ is of class $\mathcal{C}^l$.*

We use a similar definition for differentiable curves $\gamma : \mathbb{R} \to \mathcal{M}$.

Definition 6. *Let $p \in \mathcal{M}$. A tangent vector X at p is an equivalence class of differentiable curves γ with $\gamma(0) = p$ defined by $\gamma_1 = \gamma_2$ iff $\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma_1(t))|_0 = \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma_2(t))|_0$ for all differentiable f defined in a neighborhood of p. The set of tangent vector to p, denoted $T_p\mathcal{M}$ is a n-dimensional vector space. The set of smooth vector fields of $\mathcal{M}$ is the set of smoothly varying tangent vector i.e. a smooth section of the tangent bundle of $\mathcal{M}$. If X is a smooth vector field on $\mathcal{M}$, then we define:*

$$X(f) = \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))|_0 \tag{8.1}$$

which does not depend on the choice of γ in the equivalence class.

Let $\mathcal{M}$ be a differential manifold of dimension k. Let Φ, Φ' be two coordinates charts on an open subset U of $\mathcal{M}$. We denote $\Phi(x) = (x^1, \ldots, x^n)$ and $\Phi'(x) = (x'^1, \ldots, x'^n)$ for $x \in U$. Let $p \in U$. Let now $f : U \to \mathbb{R}$ be a smooth function. We define $F : \Phi(U) \subset \mathbb{R}^k \mapsto \mathbb{R}$ by $F(x) = f \circ \Phi^{-1}(x)$. We use Einstein notations.

Vector coordinates. We define, for $i \in \{1, \dots, k\}$, $\left(\frac{\partial}{\partial x^i}\right)_p \in T_p\mathcal{M}$ by:

$$\left(\frac{\partial}{\partial x^i}\right)_p (f) = \left(\frac{\partial F}{\partial x^i}\right)_\Phi (p)$$

The family $\left(\left(\frac{\partial}{\partial x^i}\right)_p\right)_{i=1,\dots,k}$ is a basis of $T_p\mathcal{M}$. So any $X \in T_p\mathcal{M}$ can be written uniquely $X = X^i \left(\frac{\partial}{\partial x^i}\right)_p$ where the X^i are the coordinates of X in the coordinate chart Φ.

Vectors change of coordinates. We have:

$$\frac{\partial}{\partial x^i}_p(f) = \frac{\partial f \circ \Phi^{-1}}{\partial x^i}(\Phi(p)) = \frac{\partial (f \circ \Phi'^{-1} \circ \Phi' \circ \Phi^{-1})}{\partial x^i}(\Phi(p))$$

We now let $F' = f \circ \Phi^{-1}$, which is a function of the coordinates x'. Now using the chain rule:

$$\frac{\partial}{\partial x^i}_p(f) = \frac{\partial (F' \circ \Phi' \circ \Phi^{-1})}{\partial x^i}(\Phi(p)) = \left(\frac{\partial x'^j}{\partial x^i}\right)_{\Phi(p)} \left(\frac{\partial F'(x')}{\partial x'^j}\right)_{\Phi'(p)}.$$

So that:

$$\left(\frac{\partial}{\partial x^i}\right)_p = \left(\frac{\partial x'^j}{\partial x^i}\right)_{\Phi(p)} \left(\frac{\partial}{\partial x'^j}\right)_p. \tag{8.2}$$

This is the formula for changing vector coordinate charts. If $X = X^i \left(\frac{\partial}{\partial x^i}\right)_p \in T_p\mathcal{M}$, we have:

$$X = X^i \left(\frac{\partial}{\partial x^i}\right)_p = X^i \left(\frac{\partial x'^j}{\partial x^i}\right)_{\Phi(p)} \left(\frac{\partial}{\partial x'^j}\right)_p.$$

so that $X'^j(p) = X^i(p) \left(\frac{\partial x'^j}{\partial x^i}\right)_{\Phi(p)}$: the coordinates of X in Φ multiplied by the Jacobian of the coordinate transformation.

Covectors. For $i \in \{1, \dots, k\}$, we define the linear form $(\mathrm{d}x^i)_p$ by:

$$(\mathrm{d}x^i)_p \left(\left(\frac{\partial}{\partial x^j}\right)_p\right) = \delta_{ij}$$

for $j \in \{1, \dots, k\}$. The family $(\mathrm{d}x^i)_{i=1,\dots,k}$ is a basis of $T_p\mathcal{M}^*$. Any linear form η on $T_p\mathcal{M}$ has coordinates in this basis and we denote $\eta = \eta_i(\mathrm{d}x^i)_p$.

Covectors change of coordinates. As before, one can show:

$$(\mathrm{d}x^i)_p = \left(\frac{\partial x^i}{\partial x'^j}\right)_{\Phi'(p)} (\mathrm{d}x'^j)_p, \tag{8.3}$$

So that if $\eta = \eta_i (\mathrm{d}x^i)_p = \eta'_i (\mathrm{d}x'^i)_p$, then:

$$\eta'_i = \left(\frac{\partial x^j}{\partial x'^i}\right)_{\Phi'(p)} \eta_j$$

Definition 7. *Let $\mathcal{M}$ be smooth manifolds. Let f be a smooth function on M and X a smooth vector field on $\mathcal{M}$. Let $h \in \mathcal{C}(M, \mathbb{R})$. Then we define:*

$$(fX)(h)(p) = f(p)X(h)(p) \tag{8.4}$$

Definition 8. *A metric on $\mathcal{M}$ is a smooth section g of the tangent bundle of symmetric bilinear forms on $\mathcal{M}$. If g is everywhere positive definite, we say that $\mathcal{M}$ is a Riemannian manifold on which g is a Riemannian metric.*

8.2 Integral on the manifold

Integral of k-forms on the manifold. We can define the integral on k-forms (i.e. multi-linear alternated of $\mathcal{M}$. We suppose that $\mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^k$ to be positively oriented [1]. Let T be a k-form on U, we denote $X = f\mathrm{d}x^1 \wedge \ldots \mathrm{d}x^k$ and define:

$$\int_U X = \int_{\Phi(U)} f(x) \mathrm{d}x^1 \ldots \mathrm{d}x^k \tag{8.5}$$

This definition of the integral is chart invariant We write the k-form T in two different coordinate charts x' and x defined by Φ' and Φ:

$$T = f'(x_1, \ldots, x_k)\mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^k = f(x'_1, \ldots, x_k)\mathrm{d}x'^1 \wedge \ldots \wedge \mathrm{d}x'^k$$

We assume that Φ' and Φ have the same orientation. Then, since f and f' are related by:

$$f'(x'_1, \ldots, x'_k) = \det\left(\frac{\partial x'^i}{\partial x^j}\right)(x'_1, \ldots, x'_k) f(\Phi \circ \Phi^{-1\prime}(x_1, \ldots, x_k))$$

we have:

$$\begin{aligned}
\int_{\Phi(U)} f(x)\mathrm{d}x^1 \ldots \mathrm{d}x^k &= \int_{\Phi'(U)} f(\Phi \circ \Phi'^{-1}(x'_1, \ldots, x'_k)) \det\left(\frac{\partial x^{i\prime}}{\partial x^j}\right)(x'_1, \ldots, x'_k)\mathrm{d}x'^1 \ldots \mathrm{d}x'^k \\
&= \int_{\Phi'(U)} f(x)\mathrm{d}x'^1 \ldots \mathrm{d}x'^k
\end{aligned}$$

where the first equality is obtained by standard multivariate change of coordinates.

[1] A manifold is orientable if it has a non vanishing smooth k-form. A choice of such a k-form defines the orientation. A chart is positively oriented if $\mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^k$ is a positive function times the fixed non-vanishing k-form.

Integral of functions on the manifold. We now assume that $\mathcal{M}$ is equipped with a smooth metric g. The volume form on $\mathcal{M}$ is the k-form defined in any positively oriented system of coordinates Φ by:

$$\eta = \sqrt{|g|}\mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^k. \tag{8.6}$$

It is the volume of an infinitesimal element on the tangent space of $\mathcal{M}$ and hence a measure $\mathrm{d}\mathcal{M}(x) = \sqrt{|g(x)|}\mathrm{d}x$ on the manifold. t $f : U \to \mathcal{M}$, we define:

$$\int_U f = \int_U f\eta = \int_{\Phi(U)} f(\Phi^{-1}(x))\sqrt{|g(x)|}\mathrm{d}x^1 \ldots \mathrm{d}x^k \tag{8.7}$$

where $g(x)$ is the matrix of the metric in coordinates. This is chart invariant.

8.3 Affine connection

Definition 9. *An affine connection on $\mathcal{M}$ is a bilinear map $\nabla : \mathcal{C}^\infty(M, TM) \times \mathcal{C}^\infty(M, TM) \to C^\infty(M, TM)$ such that:*

- $\nabla_{fX} = f\nabla_X$ *for all* $f \in C^\infty(M, \mathbb{R})$.
- $\nabla_X(fY) = X(f)Y + f\nabla_X Y$

The affine connection can be applied to any tensor using Leibniz rule. For instance, if g is a a $(0, 2)$-smooth tensor like the metric:

$$(\nabla_X g)(Y, Z) = \nabla_X(g(Y, Z)) + g(\nabla_X(Y), Z) + g(Y, \nabla_X Z) \tag{8.8}$$

We define, for any connection ∇, in a coordinate chart, the Christoffel symbols[2]:

$$\nabla_i e_j := \nabla_{e_i} e_j = \Gamma^k_{ij} e_k. \tag{8.9}$$

Definition 10. *A connection ∇ is torsion-free if $\nabla_a\nabla_b f = \nabla_b\nabla_a f = 0$ for any function f. This is equivalent to $\Gamma^\mu_{\nu\lambda} = \Gamma^\mu_{\lambda\nu}$ for all λ, μ, ν.*

Proposition 12 (Coordinates of push-forward)**.** *Let $\mathcal{M}$ and $\mathcal{N}$ be two smooth manifolds of dimension m and n respectively. Let $\Phi : \mathcal{M} \mapsto \mathcal{N}$ be a smooth map. We denote x^μ (resp. y^α) coordinates on $\mathcal{M}$ and $\mathcal{N}$. We view y as functions of x by Φ. Let $X = X^\mu \left.\frac{\partial}{\partial x^\mu}\right|_p \in T_p\mathcal{M}$. Then, $\Phi_*(X) \in T_{\Phi(p)}\mathcal{M}$ has coordinates:*

$$\Phi_*(X) = X^\mu \left(\frac{\partial y^\alpha}{\partial x^\mu}\right)_p \left.\frac{\partial}{\partial y^\alpha}\right|_{\Phi(p)}. \tag{8.10}$$

[2]Strictly speaking, the Christoffel symbols are the coefficients of the Levi-Civita connection, but they can be defined for any affine connection.

Proof. For any smooth $f : \mathcal{N} \mapsto \mathbb{R}$, we have:

$$\begin{aligned}\Phi_*(X)(f) &= X^\mu \left.\frac{\partial}{\partial x^\mu}\right|_p (f \circ \Phi) \\ &= X^\mu \left(\frac{\partial y^\alpha}{\partial x^\mu}\right)_p \left.\frac{\partial}{\partial y^\alpha}\right|_{\Phi(p)} (f).\end{aligned}$$

□

For vector fields, this rewrites:

$$\Phi_*(X)^\alpha_{\Phi(p)} = X^\mu_p \left(\frac{\partial y^\alpha}{\partial x^\mu}\right)_p \tag{8.11}$$

Similarly, for any one-form $\eta \in T^*_{\Phi(p)}\mathcal{N}$, one gets:

$$(\Phi^*(\eta)_\mu)_p = \left(\frac{\partial y^\alpha}{\partial x^\mu}\right)_p (\eta_\alpha)_{\Phi(p)}. \tag{8.12}$$

Theorem 5. *Let $\mathcal{M}$ be a smooth manifold, and let g be a metric on $\mathcal{M}$. Then, there exists a unique affine connection ∇ such that:*

- $\nabla g = 0$.
- $\nabla_X Y - \nabla_Y X = [X, Y]$ *for all smooth vector fields X and Y.*

This is called the Levi-Civita connection. Its coefficients are given by:

$$\Gamma^i{}_{k\ell} = \frac{1}{2} g^{im} \left(\frac{\partial g_{mk}}{\partial x^\ell} + \frac{\partial g_{m\ell}}{\partial x^k} - \frac{\partial g_{k\ell}}{\partial x^m}\right) = \frac{1}{2} g^{im} (g_{mk,\ell} + g_{m\ell,k} - g_{k\ell,m}), \tag{8.13}$$

8.4 Immersions

Let $\mathcal{M}$, $\mathcal{N}$ be smooth manifolds. $\Psi : \mathcal{N} \to \mathcal{M}$ is a $\mathcal{C}^k$-immersion if it is of class $\mathcal{C}^k$ and if for all $p \in \mathcal{N}$, $\mathrm{d}\Psi(p)$ has full-rank. In that case, we say that $\Psi(\mathcal{N})$ is an immersed submanifold of $\mathcal{M}$.

Definition 11. *Let $\Psi : \mathcal{N} \to \mathcal{M}$ be a smooth function. Let X be a smooth vector field on N. We can define the push-forward of $\Psi * X$ on $\mathcal{M}$ by:*

$$\Psi * (X)(f) = X(f \circ \Psi) \tag{8.14}$$

Let now g be a metric on $\mathcal{M}$. We define the pull-back of g on $\mathcal{N}$ by:

$$\Psi^*(g)(X, Y) = g(\Psi_*(X), \Psi_*(Y)) \tag{8.15}$$

for any smooth vector fields X, Y on $\mathcal{N}$. Note that for any $p \in \mathcal{M}$, $\Psi_* : T_p\mathcal{M} \to T_{\Psi(p)}\mathcal{N}$ is a linear map.

Note that if Ψ is a local diffeomorphism, then Φ_* is a one-to-one linear map from corresponding tangent spaces. We can then build the push-forward of a vector field using the local inverse Ψ^{-1}. We will denote $\Psi^* = \Psi_*^{-1}$.

Proposition 13. *Let $\Psi : \mathcal{N} \to \mathcal{M}$ be a $\mathcal{C}^k$-immersion. Let $p \in \mathcal{N}$. Since $\mathrm{d}\Psi(p)$ has full-rank, Ψ is locally invertible i.e. there exists a neighborhood U of p in $\mathcal{N}$, a neighborhood V of $\Psi(p)$ in $\Psi(N)$ and $\Psi^{-1} : V \to U$ of class $\mathcal{C}^k$ such that $\Psi \circ \Psi_{-1} = Id_{\mathcal{N}}$ on V and $\Psi^{-1} \circ \Psi = Id_{\mathcal{M}}$ on U. This local inverse can be used to define the pull-back of a vector field or the push-forward of a metric, locally.*

We note here that if Ψ is not injective, then there might be several pull-back/push-forward of tensors on $\mathcal{M}$ or on $\mathbb{N}$. We discuss this issue when considering push-forward metrics.

Proposition 14. *Let $\Psi : \mathcal{N} \to \mathcal{M}$ be a smooth function. Let $X, Y \in \mathcal{C}(N, TN)$ and $f \in \mathcal{C}(\mathcal{N}, \mathbb{R})$. Let $p \in \mathcal{N}$. Then using Proposition 13, Ψ is locally invertible in a neighborhood of p. Then, in a neighborhood of $\Psi(p)$:*

$$\Psi_*(fX) = (f \circ \Psi^{-1})\Psi_*(X) \tag{8.16}$$

Proof. Let $h \in \mathcal{C}(\mathcal{M}, \mathbb{R})$ and q in a neighborhood of p in $\mathcal{N}$. We have:

$$\Psi_*(fX)(h)(\Psi(p)) = (fX)(h \circ \Psi)(p) = f(p)X(h \circ \Psi)(p) = ((f \circ \Psi^{-1})\Psi_*(X)(h))(\Psi(p)) \tag{8.17}$$

Hence the result. □

An immersed submanifold is in general not a submanifold, but we have the following result, which is a consequence of Proposition 13:

Theorem 6. *Let $\Psi : \mathcal{N} \to \mathcal{M}$ be a $\mathcal{C}^k$-immersion and let $p \in \mathbb{N}$. Then there exists an open subset U of p in $\mathcal{N}$ such that $\Psi(U)$ is a $\mathcal{C}^k$ submanifold of $\mathcal{M}$.*

This means that an immersed submanifold is locally a submanifold.

Theorem 7. *Let $\mathcal{M}$, $\mathcal{N}$ be smooth manifolds. Let $\Psi : \mathcal{N} \to \mathcal{M}$ be a $\mathcal{C}^k$-immersion. Let $p \in \mathcal{N}$ and U, V, Ψ^{-1} as in Proposition 13. Let g be a metric on $\mathcal{N}$. Let ∇ be the Levi-Civita connection of g on $\mathcal{N}$. We define, for smooth vector fields on V, $\nabla : \mathcal{C}^\infty(M, TM) \times \mathcal{C}^\infty(M, TM) \to C^\infty(M, TM)$ by*

$$\nabla'_X Y = \Psi_*(\nabla_{\Psi^*(X)}\Psi^*(Y)) \tag{8.18}$$

Then ∇' is the Levi-Civita connection on $\mathcal{M}$ for the metric $\Psi_*(g)$.

Proof. We first prove that ∇' is a connection on V which is a submanifold of $\mathcal{M}$. Let $f : V \to \mathbb{R}$ be a smooth function and X and Y be smooth vector fields on V. We have:

$$\begin{aligned}\nabla'_{fX}Y &= \Psi_*\left(\nabla_{\Psi^*(fX)}\Psi^*(Y)\right) = \Psi_*\left(\nabla_{(f\circ\Psi)\Psi^*(X)}\Psi^*(Y)\right)\\ &= \Psi_*\left((f\circ\Psi)\nabla_{\Psi^*(X)}\Psi^*(Y)\right) = f\Psi_*\left(\nabla_{\Psi^*(X)}\Psi^*(Y)\right)\\ &= f\nabla'_X Y.\end{aligned}$$

We also have:

$$\begin{aligned}\nabla'_X(fY) &= \Psi_*\left(\nabla_{\Psi^*(X)}\Psi^*(fY)\right) = \Psi_*\left(\nabla_{\Psi^*(X)}(f\circ\Psi)\Psi^*(Y)\right)\\ &= \Psi_*\left(\Psi^*(X)(f\circ\Psi)Y\right) + \Psi_*\left((f\circ\Psi)\nabla_{\Psi^*(X)}\Psi^*(Y)\right)\\ &= [\Psi^*(X)(f\circ\Psi)\circ\Psi^{-1}]\Psi_*(Y) + f\Psi_*\left(\nabla_{\Psi^*(X)}\Psi^*(Y)\right)\\ &= \Psi^*(X)(f\circ\Psi)\Psi_*(Y) + f\Psi_*\left(\nabla_{\Psi^*(X)}\Psi^*(Y)\right)\\ &= X(f)Y + f\nabla'_X Y\end{aligned}$$

Hence, ∇' is an affine connection on V. We now show that this is the Levi-Civita connection for the push-forward metric $\Psi_*(g)$ on V. Let X, Y, Z be smooth vector fields on $\mathcal{N}$. Then:

$$\begin{aligned}\nabla'_X(\Psi_*(g)(Y,Z)) &= \Psi_*\left(\nabla_{\Psi^*(X)}\Psi^*(\Psi_*(g)(Y,Z))\right)\\ &= \Psi_*\left(\nabla_{\Psi^*(X)}g(\Psi^*(Y),\Psi^*(Z))\right)\\ &= \Psi_*\left(g(\nabla_{\Psi^*(X)}\Psi^*(Y),\Psi^*(Z)) + g(\Psi^*(Y),\nabla_{\Psi^*(X)}(\Psi^*(Z)))\right)\\ &= \Psi_*(g)(\nabla'_X Y, Z) + \Psi_*(g)(Y, \nabla'_X Z)\end{aligned}$$

which shows that $\nabla'(\Psi_*(g)) = 0$. Finally, ∇' is torsion free since:

$$\begin{aligned}\nabla'_X Y - \nabla'_Y X &= \Psi_*(\nabla_{\Psi^*(X)}\Psi^*(Y) + \nabla_{\Psi^*(Y)}\Psi^*(X))\\ &= \Psi_*([\Psi^*(X),\Psi^*(Y)])\\ &= [X,Y].\end{aligned}$$

So ∇' is the Levi-Civita connection for the induced metric. $\square$

This proves that the images of geodesics by Ψ are geodesics, and that the images of parallel curves by Ψ are parallel curves. Indeed, let γ be a geodesic on $\mathbb{N}$. We have:

$$\nabla'_{\Psi(\dot{\gamma(t)})}\Psi(\dot{\gamma(t)}) = \nabla_{\dot\gamma}\dot\gamma = 0$$

The reasonings are similar for the parallel transport: the parallel transport of a vector on V is the push-forward of the parallel transport of the pull-back of the vector along the pulled-back trajectory on $\mathcal{N}$.

8.5 Geodesically complete manifolds

Definition 12. *Let $(\mathcal{M}, g)$ be a Riemannian manifold. We say that $(\mathcal{M}, g)$ is geodesically complete if the domain of definition of all geodesic can be extended to $\mathbb{R}$.*

A geodesically complete manifold does not have boundary or singular point that can be reached in finite time. A fundamental result is Hopf-Rinow theorem:

Theorem 8. *Let $(\mathcal{M}, g)$ be a connected Riemannian manifold. Then the following statements are equivalent:*

I. *The closed and bounded subsets of M are compact.*

II. *$\mathcal{M}$ is a complete metric space for the distance:*

$$d(x, y) = inf_\gamma \int_0^1 g_{\gamma(t)}\left(\dot{\gamma}(t), \dot{\gamma}(t)\right) \mathrm{d}t$$

where the infimum is taken of all smooth curves $\gamma : [0, 1] \to \mathcal{M}$ such that $\gamma(0) = x$ and $\gamma(1) = y$.

III. *$\mathcal{M}$ is geodesically complete; that is, for every $p \in \mathcal{M}$, the exponential map is defined on the entire tangent space $T_p\mathcal{M}$.*

This theorem also implies that for any two points there exists a length minimizing geodesic between these two points. It connects the domain of definition of geodesics with the properties of the metric space itself.

8.6 Geodesics, Riemannian logarithms and geodesic completeness when $\mathcal{M} = \mathbb{R}$

We show some analytical results when $\mathcal{M} = \mathbb{R}$. We work in this subsection in coordinates in the canonical chart of $\mathbb{R}$. In this setting, the metric is a smooth positive function $g : \mathbb{R} \to \mathbb{R}$.

Lemma 6. *Let g be a Riemannian metric on $\mathbb{R}$. Then the geodesics of $(\mathbb{R}, g)$ are the functions $t \mapsto F^{-1}(at + b)$ for $a, b \in \mathbb{R}$ where $F : \mathcal{M} \to F(\mathcal{M})$ is defined by $x \mapsto \int_0^x \sqrt{g(t)}\mathrm{d}t$.*

Proof. There is a single Christoffel symbol which is $\Gamma(x) = \frac{1}{2}\frac{g'(x)}{g(x)}$ for $x \in \mathcal{M}$ (see equation (8.13) in Appendix 8). The geodesic equation is

$$\ddot{\gamma}(t) + \frac{1}{2}\frac{g'(\gamma(t))}{g(\gamma(t))}(\dot{\gamma}(t))^2 = 0.$$

Hence γ is a geodesic if and only if there exists $a \in \mathbb{R}$ such that

$$\dot{\gamma}(t)\sqrt{g \circ \gamma(t)} = a$$

for all t where γ is defined. We rewrite this $(F \circ \gamma)'(t) = a$ where F is defined on $\mathbb{R}$ by $F(x) = \int_0^x \sqrt{g(t)}\mathrm{d}t$. Hence, γ is of the form $\gamma(t) = F^{-1}(at + b)$. Conversely, the functions of the form $F^{-1}(at + b)$ are geodesics on $(\mathbb{R}, g)$. □

Lemma 7 (Riemannian logarithm)**.** *Let g be a Riemannian metric on $\mathbb{R}$, that we assume geodesically complete. Let $x, y \in \mathcal{M}$. By the Hopf-Rinow theorem (see Appendix 8.5), since $\mathcal{M}$ is geodesically complete, there exists at least one length minimizing geodesic connecting x to y. Let γ be such a geodesic with $\gamma(0) = x$, $\gamma(1) = y$ that we assume affinely parametrized. Then, expressed in the canonical coordinate chart on $\mathbb{R}$, we have:*

$$\log_x y = \frac{\int_x^y \sqrt{g(t)}\mathrm{d}t}{\sqrt{g(x)}}.$$

Proof. We set $F(p) = \int_x^p \sqrt{g(t)}\mathrm{d}t$. Using Lemma 6, there exists $a, b \in \mathbb{R}$ such that $\gamma(t) = F^{-1}(at + b)$. a and b satisfy $F^{-1}(b) = x$ and $F^{-1}(a + b) = y$. This yields $b = F(x) = 0$ and $a = \int_x^y \sqrt{g(t)}\mathrm{d}t$. We now have:

$$\gamma'(0) = \log_x y = a(F^{-1})'(b) = \frac{\int_x^y \sqrt{g(t)}\mathrm{d}t}{\sqrt{g(x)}}.$$ □

We now find a necessary and sufficient condition for $(\mathbb{R}, g)$ to be a geodesically complete manifold. In the case $\mathcal{M} = \mathbb{R}$, the interval of definition of any geodesic can be extended to $\mathbb{R}$ if and only if the geodesic does not reach $\pm\infty$ in finite time. This is materialized in the following Proposition.

Proposition 15. *Let g be a Riemannian metric on $\mathbb{R}$. We define $F : \mathbb{R} \to \mathbb{R}$ by $F(x) = \int_0^x \sqrt{g(t)}\mathrm{d}t$. $(\mathbb{R}, g)$ is geodesically complete if and only if:*

$$\lim_{x \to +\infty} F(x) = +\infty$$
$$\lim_{x \to +\infty} F(x) = -\infty.$$

Proof. First, suppose that $(\mathbb{R}, g)$ is geodesically complete. Since F is monotonic, let us assume for instance, towards a contradiction, that $\lim_{x \to +\infty} F(x) = M > 0$. Then F is a diffeomorphism from $\mathbb{R}$ onto $]a, M[$ where $a \in [-\infty, M[$. Then F^{-1} is defined on $]a, M[$. Let $x \in]a, M[$, and consider the geodesic $\gamma(t) = F^{-1}(t)$. γ is not defined for all time since it reaches $+\infty$ in finite time. This is a contradiction.

Conversely, let us assume that $\lim_{x \to +\infty} F(x) = +\infty$ and $\lim_{x \to +\infty} F(x) = -\infty$. Let γ be a geodesic on $(\mathbb{R}, g)$. Then there exists $a, b \in \mathbb{R}$ such that $\gamma(t) = F^{-1}(at + b)$. Since $\lim_{x \to +\infty} F(x) = +\infty$ and $\lim_{x \to +\infty} F(x) = -\infty$ and F is smooth and monotonic, it is a diffeomorphism of $\mathbb{R}$ and its inverse is defined on $\mathbb{R}$. Hence γ is defined for all time and $\mathcal{M}$ is geodesically complete. $\square$

Finally, we prove that every metric g on $\mathbb{R}$ such that $(\mathbb{R}, g)$ is geodesically complete can be obtained by pulling-back the Euclidean metric via a diffeomorphism of $\mathbb{R}$:

Proposition 16. *Let g be a smooth Riemannian metric on $\mathbb{R}$ such that $\mathbb{R}$ is geodesically complete. Then, there exists $\Phi : \mathbb{R} \mapsto \mathbb{R}$ diffeomorphism such that g is the pull-back of the Euclidean metric by Φ.*

Proof. We define, for $t \in \mathbb{R}$:

$$\Phi(t) = \int_0^t \sqrt{g(u)} \mathrm{d}u$$

Φ is well-defined for all $t \in \mathbb{R}$ using Proposition 15, it is smooth since g is smooth. Now, using (4.4), we notice that $\Phi^*(\eta)(p) = (\Phi')^2(p) = g(p)$. Finally, by the global inversion theorem Φ is a diffeomorphism of $\mathbb{R}$. $\square$

Chapter 9

Large Deformation Diffeomorphic Metric Mapping (LDDMM)

The Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework was built upon the idea that the comparison of shapes (images, meshes etc.) is best done by the analysis of transformations between these shapes. Such transformations should not create holes, and should not be allowed to fold the space, so they must be smooth bijections of the space in which the shapes are embedded. The LDDMM framework offers a principled way to generate such deformations.

We give a short introduction of this framework here. For mathematical details, please refer to [106].

9.1 Flow of diffeomorphisms

Let $d \in \mathbb{R}$ – which is to be thought 2 or 3– the dimension of the ambient space. The construction of diffeomorphisms is based on the integration of the flow of a time-varying velocity field. One of the main result:

Theorem 9. *Let v be a time-varying velocity field on an open bounded subset Ω of $\mathbb{R}^d$ such that:*

- *$v(t)$ is continuously differentiable for all $t \in [0, 1]$.*
- *$v(t)$ and $Dv(t)$ vanish on $\partial\Omega$ and at infinity for all $t \in [0, 1]$.*
- *v is absolutely integrable for the norm $\| \cdot \|_{1,\infty}$ (the sum of the supremum norms of the partial derivatives of v of order 1 or less).*

Then its flow Φ^v_{st} –its integration from s to t for $s, t \in [0, 1]$ – is a diffeomorphism of Ω.

9.2 Admissible vector space

Following the previous result, to obtain a large collection of diffeomorphisms, one needs a large collection of velocity fields satisfying the assumption of the previous Theorem. Hence the definition:

Definition 13. *A Banach space $V \subset \mathcal{C}_0^1(\Omega, \mathbb{R}^d)$ is admissible if it is canonically embedded in $\mathcal{C}_0^1(\Omega, \mathbb{R}^d)$ that is if there exists $C \geq 0$ such that for all $v \in V$:*

$$\|v\|_V \geq V\|v\|_{1,\infty}$$

If V is an admissible vector space, we denote $\chi_V^1(\Omega)$ the set of absolutely integrable time-dependent vector fields $(v(t, t \in [0,1])$.

We are now ready for the second main result:

Theorem 10. *Let V be admissible. Let $G_V = \{\Phi_{01}^v | v \in \chi_V^1(\Omega)\}$. Then G_V is a group for the composition of functions.*

Hence, the definition of an admissible vector space is enough to generate a group of diffeomorphisms. Now the algorithmic aspects of this construction come into play. A way to proceed now is to look at Reproducing Kernel Hilbert Spaces (RKHS).

9.3 The landmark manifold

The definition of the family of diffeomorphisms that we use in practice is a bit different from this generic LDDMM framework. We fix a manifold of landmark (points) and map this manifold onto a diffeomorphism space, using the integration of time-varying velocity fields as an intermediate step. Details about this construction can be found in [23]. We start with a definition

Definition 14. *Let $d \in \mathbb{N}$, let $n \in \mathbb{N}$. We define:*

$$\mathcal{M} = \{(x_1, \dots, x_n) | x_i \in \mathbb{R}^d \ \forall i \in \{1, \dots, n\}, x_i \neq x_j \ \forall i, j \in \{1, \dots, n\}\} \tag{9.1}$$

$\mathcal{M}$ is a smooth manifold. Now it is possible to equip $\mathcal{M}$ with a Riemannian metric:

Definition 15. *$K : \mathbb{R}^{dn} \times \mathbb{R}^{dn} \to \mathbb{R}$ is a positive definite symmetric kernel if for all $c \in \mathcal{M}$ and all $\alpha \in T_c\mathcal{M}^*$ we have:*

$$\sum_i \sum_j K(c_i, c_j)\alpha_i^\top \alpha \geq 0$$

Proposition 17. *Let $\alpha, \beta \in \mathbb{R}^{dn}$ be co-tangent vectors to $c \in \mathcal{M}$. Let K be a smooth positive definite symmetric kernel. Then:*

$$K_c(\alpha, \beta) = \sum_i \sum_j \alpha_i^\top K(c_i, c_j)\beta_j \tag{9.2}$$

is a Riemannian co-metric on $\mathcal{M}$. We denote g the associated metric.

Proof. g is smoothly varying and defines an inner product on $T_c\mathcal{M}^*$ for all $c \in \mathcal{M}$. $\square$

We now connect this construction to the notion of admissible spaces for diffeomorphic flow.

Proposition 18. *Let Ω be an open bounded subset of $\mathbb{R}$. Let us consider the space of all possible vector fields that can be generated when the control points c lie in Ω:*

$$V = \{x \to v(x) = \sum_{i=1}^{n} K(c_i, x)\alpha_i | x_1, \dots, x_n \in \Omega, \alpha_1, \dots \alpha_n \in \mathbb{R}^d\}$$

We equip V with the scalar product: $< (c, \alpha), (d, \beta) >_V = \sum_i \sum_j \alpha_i^\top K(c_i, d_j)\beta_j$ and denote $\| \cdot \|_V$ the corresponding norm. Then V is an admissible space of vector fields.

As mentioned above, this space is a RKHS and in the algorithmic aspects underlying this construction, we benefit from the use of the kernel.

Definition 16. *Let $c \in \mathcal{M}$ and $\alpha \in T_c\mathcal{M}$. We denote $\Phi_t^{c,\alpha}$ the diffeomorphism obtained by integration of the time-varying velocity field:*

$$v(t) = K(c(t), c(t))\alpha(t) \tag{9.3}$$

here $c(t), \alpha(t)$ is the geodesic γ and its momenta at time t with initial position c and momenta α.

The operation $\Pi : (c, \alpha) \mapsto \Phi_1^{c,\alpha}$ maps control points and vectors to diffeomorphisms. When we work with shapes and Riemannian geometry, we actually perform all computations on the landmark manifold (geodesics, parallel transport etc) and map back the results to the diffeomorphism space after hand. Strictly speaking, we do not perform the transport on a manifold of diffeomorphisms, but on the manifold of landmarks.

For details regarding the implementation of this LDDMM instance, we refer the reader to [9]. A major improvement in the implementations came from the use of the PyTorch library which enables both automatic differentiation and efficient GPU usage.

List of publications

[1] Alexandre Bône, Maxime Louis, Olivier Colliot, Stanley Durrleman, Alzheimer's Disease Neuroimaging Initiative, et al. Learning low-dimensional representations of shape data sets with diffeomorphic autoencoders. In *International Conference on Information Processing in Medical Imaging*, pages 195–207. Springer, Cham, 2019.

[2] Alexandre Bône, Maxime Louis, Benoît Martin, and Stanley Durrleman. Deformetrica 4: an open-source software for statistical shape analysis. In *International Workshop on Shape in Medical Imaging*, pages 3–13. Springer, Cham, 2018.

[3] Alexandre Bône, Maxime Louis, Alexandre Routier, Jorge Samper, Michael Bacci, Benjamin Charlier, Olivier Colliot, and Stanley Durrleman. Prediction of the progression of subcortical brain structures in alzheimer's disease from baseline. In *6th MICCAI Workshop on Mathematical Foundations of Computational Anatomy*, 2017.

[4] Alexandre Bône, Benoît Martin, Maxime Louis, Olivier Colliot, and Stanley Durrleman. Hierarchical modeling of alzheimer's disease progression from a large longitudinal mri data set. 2019.

[5] Raphael Couronne, Maxime Louis, and Stanley Durrleman. Longitudinal autoencoder for multi-modal disease progression modelling. 2019.

[6] Igor Koval, Alexandre Bône, Maxime Louis, Simona Bottani, Arnaud Marcoux, Jorge Samper-Gonzalez, Ninon Burgos, Benjamin Charlier, Anne Bertrand, Stéphane Epelbaum, et al. Simulating alzheimer's disease progression with person-alised digital brain models. 2018.

[7] Maxime Louis, Alexandre Bône, Benjamin Charlier, Stanley Durrleman, Alzheimer's Disease Neuroimaging Initiative, et al. Parallel transport in shape analysis: a scalable numerical scheme. In *International Conference on Geometric Science of Information*, pages 29–37. Springer, Cham, 2017.

[8] Maxime Louis, Benjamin Charlier, and Stanley Durrleman. Geodesic discriminant analysis for manifold-valued data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 332–340, 2018.

[9] Maxime Louis, Benjamin Charlier, and Stanley Durrleman. Learning riemannian geometry for mixed-effect models using deep generative networks. 2018.

[10] Maxime Louis, Benjamin Charlier, Paul Jusselin, Pal Susovan, and Stanley Durrleman. A fanning scheme for the parallel transport along geodesics on riemannian manifolds. 2017.

[11] Maxime Louis, Raphaël Couronné, Igor Koval, Benjamin Charlier, and Stanley Durrleman. Riemannian geometry learning for disease progression modelling. In *International Conference on Information Processing in Medical Imaging*, pages 542–553. Springer, Cham, 2019.

[12] Alexander D Rider, David C Moore, Charles P Blakemore, Maxime Louis, Marie Lu, and Giorgio Gratta. Search for screened interactions associated with dark energy below the 100 μ m length scale. *Physical review letters*, 117(10):101101, 2016.

Bibliography

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] G. Arvanitidis, L. K. Hansen, and S. Hauberg. A locally adaptive normal distribution. In *Advances in Neural Information Processing Systems*, pages 4251–4259, 2016.

[3] G. Arvanitidis, L. K. Hansen, and S. Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

[4] G. Arvanitidis, L. K. Hansen, and S. Hauberg. Maximum likelihood estimation of riemannian metrics from euclidean data. In *International Conference on Geometric Science of Information*, pages 38–46. Springer, 2017.

[5] M. Beg, M. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *IJCV*, 61(2):139–157, 2005.

[6] E. Bigio, L. Hynan, E. Sontag, S. Satumtira, and C. White. Synapse loss is greater in presenile than senile onset alzheimer disease: implications for the cognitive reserve hypothesis. *Neuropathology and applied neurobiology*, 28(3):218–227, 2002.

[7] A. Bône, O. Colliot, and S. Durrleman. Learning distributions of shape trajectories from longitudinal datasets: a hierarchical model on a manifold of diffeomorphisms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9271–9280, 2018.

[8] A. Bône, M. Louis, O. Colliot, S. Durrleman, A. D. N. Initiative, et al. Learning low-dimensional representations of shape data sets with diffeomorphic autoencoders. In *International Conference on Information Processing in Medical Imaging*, pages 195–207. Springer, 2019.

[9] A. Bône, M. Louis, B. Martin, and S. Durrleman. Deformetrica 4: an open-source software for statistical shape analysis. In *International Workshop on Shape in Medical Imaging*. Springer, 2018.

[10] A. Bône, M. Louis, A. Routier, J. Samper, M. Bacci, B. Charlier, O. Colliot, S. Durrleman, A. D. N. Initiative, et al. Prediction of the progression of subcortical brain structures in alzheimer's disease from baseline. In *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, pages 101–113. Springer, 2017.

[11] A. Bône, M. Louis, A. Routier, J. Samper, M. Bacci, B. Charlier, O. Colliot, S. Durrleman, A. D. N. Initiative, et al. Prediction of the progression of subcortical brain structures in alzheimer's disease from baseline. In *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, pages 101–113. Springer, 2017.

[12] A. Chartsias, T. Joyce, et al. Multimodal mr synthesis via modality-invariant latent representation. *IEEE transactions on medical imaging*, 2018.

[13] M. Chupin, E. Gérardin, R. Cuingnet, C. Boutet, L. Lemieux, S. Lehéricy, H. Benali, L. Garnero, and O. Colliot. Fully automatic hippocampus segmentation and classification in alzheimer's disease and mild cognitive impairment applied on data from adni. *Hippocampus*, 19(6):579–587, 2009.

[14] R. Couronne, M. Louis, and S. Durrleman. Longitudinal autoencoder for multi-modal disease progression modelling. working paper or preprint, Apr. 2019.

[15] R. Cui, M. Liu, and G. Li. Longitudinal analysis for alzheimer's disease diagnosis using rnn. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 1398–1401. IEEE, 2018.

[16] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, A. D. N. Initiative, et al. Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. *NeuroImage*, 56(2):766–781, 2011.

[17] A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999.

[18] J. Damon and J. Marron. Backwards principal component analysis and principal nested relations. *Journal of Mathematical Imaging and Vision*, 50(1-2):107–114, 2014.

[19] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128, 1999.

[20] M. P. Do Carmo. *Riemannian geometry*. Birkhauser, 1992.

[21] S. Durrleman, X. Pennec, A. Trouvé, J. Braga, G. Gerig, and N. Ayache. Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *International journal of computer vision*, 103(1):22–59, 2013.

[22] S. Durrleman, M. Prastawa, N. Charon, J. R. Korenberg, S. Joshi, G. Gerig, and A. Trouvé. Morphometry of anatomical shape complexes with dense deformations and sparse parameters. *NeuroImage*, 2014.

[23] S. Durrleman, M. Prastawa, N. Charon, J. R. Korenberg, S. Joshi, G. Gerig, and A. Trouvé. Morphometry of anatomical shape complexes with dense deformations and sparse parameters. *NeuroImage*, 2014.

[24] L. Falissard, G. Fagherazzi, N. Howard, and B. Falissard. Deep clustering of longitudinal data. *arXiv preprint arXiv:1802.03212*, 2018.

[25] J. Fishbaugh, M. Prastawa, G. Gerig, and S. Durrleman. Geodesic regression of image and shape data for improved modeling of 4D trajectories. In *ISBI 2014 - 11th International Symposium on Biomedical Imaging*, pages 385 – 388, Apr. 2014.

[26] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen.*, 7:179–188, 1936.

[27] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004.

[28] T. Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *Int J Comput Vis*, 105(2):171–185, 2013.

[29] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[30] L. Gao, H. Pan, F. Liu, X. Xie, Z. Zhang, J. Han, A. D. N. Initiative, et al. Brain disease diagnosis using deep learning features from longitudinal mr images. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 327–339. Springer, 2018.

[31] S. Garbarino, M. Lorenzi, A. D. N. Initiative, et al. Modeling and inference of spatio-temporal protein dynamics across brain networks. In *International Conference on Information Processing in Medical Imaging*, pages 57–69. Springer, 2019.

[32] G. Gerig, B. Davis, P. Lorenzen, S. Xu, M. Jomier, J. Piven, and S. Joshi. Computational anatomy to assess longitudinal trajectory of brain growth. In *Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, pages 1041–1047. IEEE, 2006.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[34] P. Gori, O. Colliot, Y. Worbe, L. Marrakchi-Kacem, S. Lecomte, C. Poupon, A. Hartmann, N. Ayache, and S. Durrleman. *Bayesian Atlas Estimation for the Variability*

Analysis of Shape Complexes, pages 267–274. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[35] M. Hadj-Hamou, M. Lorenzi, N. Ayache, and X. Pennec. Longitudinal analysis of image time series with diffeomorphic deformations: A computational framework based on stationary velocity fields. *Frontiers in Neuroscience*, 10:236, 2016.

[36] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–176, 1996.

[37] S. Hauberg. Only bayes should learn a manifold (on the estimation of differential geometric structure from data). *arXiv preprint arXiv:1806.04994*, 2018.

[38] S. Hauberg, O. Freifeld, and M. J. Black. A geometric take on metric learning. In *Advances in Neural Information Processing Systems*, pages 2024–2032, 2012.

[39] T. Hotz, S. Huckemann, A. Munk, D. Gaffrey, and B. Sloboda. Shape spaces for prealigned star–haped objects-studying the growth of plants by principal components analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(1):127–143, 2010.

[40] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[41] J.-H. Jacobsen, A. Smeulders, and E. Oyallon. i-revnet: Deep invertible networks. *arXiv preprint arXiv:1802.07088*, 2018.

[42] S. H. Joshi, Q. Xie, S. Kurtek, A. Srivastava, and H. Laga. Surface shape morphometry for hippocampal modeling in alzheimer's disease. In *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*, pages 1–8. IEEE, 2016.

[43] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

[44] D. G. Kendall. A survey of the statistical theory of shape. *Statist. Sci.*, 4(2):87–99, 05 1989.

[45] A. Kheyfets, W. A. Miller, and G. A. Newton. Schild's ladder parallel transport procedure for an arbitrary connection. *International Journal of Theoretical Physics*, 39(12):2891–2898, 2000.

[46] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[47] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[48] I. Koval, J.-B. Schiratti, A. Routier, M. Bacci, O. Colliot, S. Allassonnière, S. Durrleman, A. D. N. Initiative, et al. Statistical learning of spatiotemporal patterns from longitudinal manifold-valued networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 451–459. Springer, 2017.

[49] L. Kuhnel, T. Fletcher, S. Joshi, and S. Sommer. Latent space non-linear statistics. *arXiv preprint arXiv:1805.07632*, 2018.

[50] L. Kühnel and S. Sommer. Computational anatomy in theano. In *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, pages 164–176. Springer, 2017.

[51] N. M. Laird and J. H. Ware. Random-Effects Models for Longitudinal Data. *Biometrics*, dec 1982.

[52] B. Lam, M. Masellis, M. Freedman, D. T. Stuss, and S. E. Black. Clinical, imaging, and pathological heterogeneity of the alzheimer's disease syndrome. *Alzheimer's research & therapy*, 5(1):1, 2013.

[53] G. Lebanon. Learning riemannian metrics. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2002.

[54] S. Lee, N. Charon, B. Charlier, K. Popuri, E. Lebed, M. V. Sarunic, A. Trouvé, and M. F. Beg. Atlas-based shape analysis and classification of retinal optical coherence tomography images using the functional shape (fshape) framework. *Medical image analysis*, 35:570–581, 2017.

[55] C. Lenglet, M. Rousson, R. Deriche, and O. Faugeras. 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, 2006.

[56] M. J. Lindstrom and D. M. Bates. Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics*, 46(3):673, sep 1990.

[57] M. Lorenzi, N. Ayache, G. Frisoni, and X. Pennec. 4D registration of serial brain's MR images: a robust measure of changes applied to Alzheimer's disease. *Spatio Temporal Image Analysis Workshop (STIA), MICCAI*, 2010.

[58] M. Lorenzi, N. Ayache, and X. Pennec. Schild's ladder for the parallel transport of deformations in time series of images. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 463–474. Springer, 2011.

[59] M. Lorenzi and X. Pennec. Geodesics, parallel transport & one-parameter subgroups for diffeomorphic image registration. *International journal of computer vision*, 105(2):111–127, 2013.

[60] M. Lorenzi and X. Pennec. Geodesics, parallel transport & one-parameter subgroups for diffeomorphic image registration. *IJCV*, 105(2):111–127, Nov. 2013.

[61] M. Lorenzi and X. Pennec. Parallel transport with pole ladder: Application to deformations of time series of images. In *Geometric Science of Information*, volume 8085, pages 68–75, 2013.

[62] M. Louis, A. Bône, B. Charlier, S. Durrleman, A. D. N. Initiative, et al. Parallel transport in shape analysis: a scalable numerical scheme. In *International Conference on Geometric Science of Information*, pages 29–37. Springer, 2017.

[63] M. Louis, A. Bône, B. Charlier, S. Durrleman, A. D. N. Initiative, et al. Parallel transport in shape analysis: a scalable numerical scheme. In *International Conference on Geometric Science of Information*, pages 29–37. Springer, 2017.

[64] M. Louis, B. Charlier, and S. Durrleman. Geodesic discriminant analysis for manifold-valued data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 332–340, 2018.

[65] M. Louis, B. Charlier, P. Jusselin, S. Pal, and S. Durrleman. A fanning scheme for the parallel transport along geodesics on riemannian manifolds. *SIAM Journal on Numerical Analysis*, 56(4):2563–2584, 2018.

[66] M. Louis et al. Riemannian geometry learning for disease progression modelling. In *International Conference on Information Processing in Medical Imaging*, 2019.

[67] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[68] D. J. C. MacKay. Probable networks and plausible predictions - - a review of practical bayesian methods for supervised neural networks, 1995.

[69] S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

[70] P. Manfredo. Riemannian geometry. 1992.

[71] V. S. Matveev. Geodesically equivalent metrics in general relativity. *Journal of Geometry and Physics*, 62(3):675–691, 2012.

[72] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[73] C. Metz, S. Klein, M. Schaap, T. van Walsum, and W. Niessen. Nonrigid registration of dynamic medical imaging data using nd + t b-splines and a groupwise optimization approach. *Medical Image Analysis*, 15(2):238 – 249, 2011.

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

[75] M. I. Miller, A. Trouvé, and L. Younes. Geodesic shooting for computational anatomy. *Journal of Mathematical Imaging and Vision*, 24(2):209–228, Mar 2006.

[76] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6:26094 EP –, 05 2016.

[77] J. Ngiam, A. Khosla, M. Kim, et al. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011.

[78] X. Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127, 2006.

[79] X. Pennec et al. Barycentric subspace analysis on manifolds. *The Annals of Statistics*, 46(6A):2711–2746, 2018.

[80] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[81] J. Peyrat, H. Delingette, M. Sermesant, X. Pennec, C. Xu, and N. Ayache. Registration of 4D time-series of cardiac images with multichannel diffeomorphic Demons. *Med Image Comput Comput Assist Interv*, 2008.

[82] S. J. D. Prince. Probabilistic Linear Discriminant Analysis for Infermences About Identity. In *Proc. International Conference on Computer Vision*, 2007.

[83] A. Qiu, L. Younes, M. I. Miller, and J. G. Csernansky. Parallel transport in diffeomorphisms distinguishes the time-dependent pattern of hippocampal surface deformation due to healthy aging and the dementia of the alzheimer's type. *NeuroImage*, 40(1):68–76, 2008.

[84] D. E. Rumelhart, G. E. Hinton, R. J. Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[85] M. Rumpf and B. Wirth. Variational time discretization of geodesic calculus. *IMA Journal of Numerical Analysis*, 35(3):1011–1046, 2014.

[86] J.-B. Schiratti, S. Allassonniere, O. Colliot, and S. Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *Advances in Neural Information Processing Systems*, pages 2404–2412, 2015.

[87] J.-B. Schiratti, S. Allassonnière, O. Colliot, and S. Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *NIPS 28*, pages 2404–2412. Curran Associates, Inc., 2015.

[88] J.-B. Schiratti, S. Allassonniere, O. Colliot, and S. Durrleman. A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. Accepted at Journal of Machine Learning Research., Sept. 2016.

[89] H. Shao, A. Kumar, and P. T. Fletcher. The riemannian geometry of deep generative models. *arXiv preprint arXiv:1711.08014*, 2017.

[90] N. Singh, J. Hinkle, S. Joshi, and P. T. Fletcher. Hierarchical geodesic models in diffeomorphisms. *International Journal of Computer Vision*, 117(1):70–92, 2016.

[91] S. Sommer. Horizontal dimensionality reduction and iterated frame bundle development. In *Geometric Science of Information*, pages 76–83. Springer, 2013.

[92] S. Sommer, A. Arnaudon, L. Kuhnel, and S. Joshi. Bridge simulation and metric estimation on landmark manifolds. In *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, pages 79–91. Springer, 2017.

[93] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen. *Manifold Valued Statistics, Exact Principal Geodesic Analysis and the Effect of Linear Approximations*, pages 43–56. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[94] N. Srivastava, E. Mansimov, et al. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, 2015.

[95] T. I. Standish, D. W. Molloy, M. Bédard, E. C. Layne, E. A. Murray, and D. Strang. Improved reliability of the standardized alzheimer's disease assessment scale (sadas) compared with the alzheimer's disease assessment scale (adas). *Journal of the American Geriatrics Society*, 44(6):712–716, 1996.

[96] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[97] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Non-parametric diffeomorphic image registration with the demons algorithm. In *Med Image Comput Comput Assist Interv*, pages 319–326. Springer Berlin Heidelberg, 2007.

[98] K. Vogtmann, A. Weinstein, and V. Arnol'd. *Mathematical Methods of Classical Mechanics*. Graduate Texts in Mathematics. Springer New York, 1997.

[99] L. Wang, F. Beg, T. Ratnanather, C. Ceritoglu, L. Younes, J. C. Morris, J. G. Csernansky, and M. I. Miller. Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the alzheimer type. *IEEE Transactions on Medical Imaging*, 26(4):462–470, 2007.

[100] G. Wu, Q. Wang, J. Lian, and D. Shen. Estimating the 4d respiratory lung motion by spatiotemporal registration and building super-resolution image. In *MICCAI*, pages 532–539, 2011.

[101] L. Wu, C. Shen, and A. van den Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 65:238–250, 2017.

[102] Q. Xie, S. Kurtek, H. Le, and A. Srivastava. Parallel transport of deformations in shape space of elastic surfaces. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 865–872, 2013.

[103] T. Yang, G. Arvanitidis, D. Fu, X. Li, and S. Hauberg. Geodesic clustering in deep generative models. *arXiv preprint arXiv:1809.04747*, 2018.

[104] X. Yang et al. Deep multimodal representation learning from temporal data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[105] L. Younes. Jacobi fields in groups of diffeomorphisms and applications. *Quarterly of applied mathematics*, pages 113–134, 2007.

[106] L. Younes. *Shapes and diffeomorphisms*, volume 171. Springer Science & Business Media, 2010.

[107] L. Younes. *Shapes and diffeomorphisms*. Heidelberg: Springer, 2010. "A direct application of what is presented in the book is a branch of the computerized analysis of medical images called computational anatomy"–Back cover.

[108] L. Younes, A. Qiu, R. L. Winslow, and M. I. Miller. Transport of relational structures in groups of diffeomorphisms. *Journal of mathematical imaging and vision*, 32(1):41–56, 2008.

[109] M. Zhang and P. T. Fletcher. Probabilistic principal geodesic analysis. In *Advances in Neural Information Processing Systems*, pages 1178–1186, 2013.

[110] M. Zhang and P. T. Fletcher. Bayesian principal geodesic analysis in diffeomorphic image registration. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. SpringerLink, 2014.

[111] M. Zhang and T. Fletcher. Probabilistic principal geodesic analysis. In *Advances in Neural Information Processing Systems*, pages 1178–1186, 2013.

Contents

Introduction **i**

I Parallel transport: an efficient numerical scheme and its applications **1**

1 A numerical scheme and a proof of convergence **5**

1.1 Introduction 5
1.2 Rationale 6
1.2.1 Notations and assumptions 6
1.2.2 The key identity 7
1.2.3 Convergence rate on $\mathbb{S}^2$ 8
1.3 The numerical scheme 9
1.3.1 The algorithm 9
1.3.2 Possible variations 10
1.3.3 The convergence Theorem 12
1.4 Proof of the convergence Theorem 1 13
1.5 Numerical experiments 16
1.5.1 Setup 16
1.5.2 Results 16
1.5.3 Comparison with Schild's ladder 19
1.6 Conclusion 19
1.7 Pseudo-code and proofs 20
1.7.1 Pseudo-code 20
1.8 Proofs 21

2 Application to shape analysis [62, 10] **31**

2.1 Introduction 31
2.2 Parallel transport in the context of shape analysis 32
2.2.1 The chosen family of diffeomorphisms 32
2.2.2 Parallel transport on $\mathcal{G}_c$ 33
2.3 Method for future shape prediction 35
2.4 Results 37
2.4.1 Data, pre-processing, parameters and performance metric 37

2.4.2 Estimating the error associated to a single parallel transport 38
2.4.3 Geodesic regression extrapolation 39
2.4.4 Non reparametrized transport 42
2.4.5 Refining with cognitive dynamical parameters 45
2.5 Conclusion 45

II Geodesic Discriminant Analysis for manifold-valued data 49

3 Geodesic Discriminant Analysis for manifold-valued data 51
3.1 Introduction 51
3.2 Geometric Geodesic Discriminant Analysis 53
3.2.1 Inference 54
3.2.2 Dimension reduction and classification 55
3.3 Probabilistic Geodesic Discriminant Analysis 55
3.3.1 Inference 57
3.3.2 Dimension reduction and classification 57
3.4 Probabilistic GDA for shape analysis. 58
3.4.1 Embedding shapes and images on a manifold 58
3.4.2 A generative model 59
3.5 Applications and Results 61
3.5.1 Geometric GDA on $\mathbb{S}^2$ 61
3.5.2 Kimia-216 62
3.5.3 Brain structures in the course of Alzheimer's disease 63
3.6 Conclusion 66

III Riemannnian geometry learning 67

4 Riemannian metrics so as to be normally distributed 75
4.1 The toy example $\mathcal{M} = \mathbb{R}$ 75
4.2 Experiments 79
4.3 Shedding light on Generative Adversarial Networks 82
4.3.1 Push-forward versus pull-back metric 85

5 Riemannian metrics for geodesicity 87
5.1 The toy example $\mathcal{M} = \mathbb{R}$ 88
5.2 Geodesically rigid metrics 89
5.3 A more general existence result 91
5.4 A parametric family of Riemannian metrics on $\mathbb{R}^d$ 92
5.4.1 Cholesky decomposition 92

5.4.2 Building a Riemannian metric 92
5.5 Experiments 97
5.6 Parametric Riemannian metric learning for longitudinal disease modelling . 99
5.6.1 Original modelling 99
5.6.2 Proposed extension 100
5.6.3 Results 101

6 Disease modelling using deep neural networks [66] 103
6.1 Introduction 103
6.2 Propagation model and deep generative models 105
6.2.1 Push-forward of a Riemannian metric 105
6.2.2 Model for longitudinal progression 105
6.2.3 Encoding the latent variables 106
6.2.4 Regularization 107
6.2.5 Cost function and inference 107
6.3 Experimental results 108
6.3.1 On a synthetic set of images 108
6.3.2 On cognitive scores 109
6.3.3 On anatomical MRIs 112
6.4 Conclusion 113

7 Longitudinal auto-encoder for multimodal disease progression modeling [14] 115
7.1 Introduction 115
7.2 Methods 116
7.2.1 Decoding : Non linear mixed effect model 116
7.2.2 Encoding 119
7.2.3 Regularization, cost function and optimization 119
7.3 Experimental results 120
7.3.1 Cognitive scores: proof of concept 120
7.3.2 A synthetic data set 123
7.3.3 Application to Alzheimer's disease future image prediction 124
7.4 Conclusion and perspectives 125

IV Conclusion and perspectives 127

V Appendix 131

8 Riemannian geometry 133

8.1 Manifold, tangent vector, metric. 133
8.2 Integral on the manifold 135
8.3 Affine connection 136
8.4 Immersions 137
8.5 Geodesically complete manifolds 140
8.6 Geodesics, Riemannian logarithms and geodesic completeness when $\mathcal{M} = \mathbb{R}$ 140

9 Large Deformation Diffeomorphic Metric Mapping (LDDMM) 143
9.1 Flow of diffeomorphisms 143
9.2 Admissible vector space 143
9.3 The landmark manifold 144

List of publications 147

Bibliography 148