Riemannian and sub-Riemannian methods for dimension reduction

Morten Akhøj Pedersen

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Méthodes riemanniennes et sous-riemanniennes pour la réduction de dimension

Morten Akhøj Pedersen

Université Côte d'Azur, France & University of Copenhagen, Denmark

Préparée à l'Inria, équipe Epione, et UCPH, équipe Applied Geometry, Department of Computer Science

Présentée en vue de l'obtention du grade de docteur en Automatique, traitement du signal et des images d'Université Côte d'Azur, dirigée par Xavier Pennec et Stefan Sommer, soutenue le 23 novembre 2023

Devant le jury composé de :

Rapporteurs	Alain Trouvé	Professeur	Université Paris-Saclay
	Klas Modin	Professor	Chalmers University, SE
Examinateurs	Mads Nielsen	Professor	University of Copenhagen, DK
	Pierre Alliez	Directeur de recherches	Univ. Côte d'Azur & Inria, Sophia Antipolis, FR
Codirecteurs :			
	Xavier Pennec	Directeur de recherches	Univ. Côte d'Azur & Inria, Sophia Antipolis, FR
	Stefan Sommer	Professor	University of Copenhagen, DK

Inria Epione e-patient / e-medicine erc Geometric Statistics

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Riemannian and sub-Riemannian methods for dimension reduction

Morten Akhøj Pedersen

Submitted at Université Côte d'Azur, France & University of Copenhagen, Denmark

Prepared at Inria, team Epione, and the Department of Computer Science at UCPH, team Applied Geometry

Jury :

Reporters	Alain Trouvé	Professor	Université Paris-Saclay
	Klas Modin	Professor	Chalmers University, SE
Examiners	Mads Nielsen	Professor	University of Copenhagen, DK
	Pierre Alliez	Director of research	Univ. Côte d'Azur & Inria, Sophia Antipolis, FR
Supervisors :			
	Xavier Pennec	Director of research	Univ. Côte d'Azur & Inria, Sophia Antipolis, FR
	Stefan Sommer	Professor	University of Copenhagen, DK

Inria Epione e-patient / e-medicine

Riemannian and sub-Riemannian methods for dimension reduction

In this thesis, we propose new methods for dimension reduction based on differential geometry, that is, finding a representation of a set of observations in a space of lower dimension than the original data space. Methods for dimension reduction form a cornerstone of statistics, and thus have a very wide range of applications. For instance, a lower dimensional representation of a data set allows visualization and is often necessary for subsequent statistical analyses. In ordinary Euclidean statistics, the data belong to a vector space and the lower dimensional space might be a linear subspace or a non-linear submanifold approximating the observations. The study of such smooth manifolds, differential geometry, naturally plays an important role in this last case, or when the data space is itself a known manifold. Methods for analysing this type of data form the field of *geometric statistics*. In this setting, the approximating space found by dimension reduction is naturally a submanifold of the given manifold. The starting point of this thesis is geometric statistics for observations belonging to a known Riemannian manifold, but parts of our work form a contribution even in the case of data belonging to Euclidean space, $\mathbb{R}^d$.

An important example of manifold valued data is *shapes*, in our case discrete curves or surfaces. In evolutionary biology, researchers are interested in studying reasons for and implications of morphological differences between species. Shape is one way to formalize morphology. This application motivates the first main contribution of the thesis. We generalize a dimension reduction method used in evolutionary biology, *phylogenetic principal component analysis* (P-PCA), to work for data on a Riemannian manifold - so that it can be applied to shape data. P-PCA is a version of PCA for observations that are assumed to be leaf nodes of a phylogenetic tree. From a statistical point of view, the important property of such data is that the observations (leaf node values) are not necessarily independent. We define and estimate intrinsic weighted means and covariances on a manifold which takes the dependency of the observations into account. We then define phylogenetic PCA on a manifold to be the eigendecomposition of the weighted covariance in the tangent space of the weighted mean. We show that the mean estimator that is currently used in evolutionary biology for studying morphology corresponds to taking only a single step of our Riemannian gradient descent algorithm for the intrinsic mean, when the observations are represented in Kendall's shape space.

Our second main contribution is a non-parametric method for dimension reduction that can be used for approximating a set of observations based on a very flexible class of submanifolds. This method is novel even in the case of Euclidean data. The method works by constructing a subbundle of the tangent bundle on the data manifold M via local PCA. We call this subbundle the *principal subbundle*. We then observe that this subbundle induces a *sub-Riemannian* structure on M and we show that the resulting sub-Riemannian geodesics with respect to this structure stay close to the set of observations. Moreover, we show that sub-Riemannian geodesics starting from a given point locally generate a submanifold which is radially aligned with the estimated subbundle, even for non-integrable subbundles. Non-integrability is likely to occur when the subbundle is estimated from noisy data, and our method demonstrates that sub-Riemannian geometry is a natural framework for dealing which such problems. Numerical experiments illustrate the power of our framework by showing that we can achieve impressively large range reconstructions even in the presence of quite high levels of noise.

Keywords: *geometric statistics, differential geometry, Riemannian geometry, sub-Riemannian geometry, mathematical statistics, machine learning*

Méthodes riemanniennes et sous-riemanniennes pour la réduction de dimension

Nous proposons dans cette thèse de nouvelles méthodes de réduction de dimension fondées sur la géométrie différentielle. Il s'agit de trouver une représentation d'un ensemble d'observations dans un espace de dimension inférieure à l'espace d'origine des données. Les méthodes de réduction de dimension constituent la pierre angulaire des statistiques et ont donc un très large éventail d'applications. Dans les statistiques euclidiennes ordinaires, les données appartiennent à un espace vectoriel et l'espace de dimension inférieure peut être un sous-espace linéaire ou une sous-variété non linéaire approximant les observations. L'étude de telles variétés lisses, la géométrie différentielle, joue naturellement un rôle important dans ce dernier cas. Lorsque l'espace des données est lui-même une variété, l'espace approximant de dimension réduite est naturellement une sous-variété de la variété intiale. Les méthodes d'analyse de ce type de données relèvent du domaine des *statistiques géométriques*. Les statistiques géométriques pour des observations appartenant à une variété riemannienne sont le point de départ de cette thèse, mais une partie de notre travail apporte une contribution même dans le cas de données appartenant à l'espace euclidien, $\mathbb{R}^d$.

Les *formes*, dans notre cas des courbes ou des surfaces discrètes, sont un exemple important de données à valeurs dans les variétés. En biologie évolutive, les chercheurs s'intéressent aux raisons et aux implications des différences morphologiques entre les espèces. Cette application motive la première contribution principale de la thèse. Nous généralisons une méthode de réduction de dimension utilisée en biologie évolutive, l'*analyse en composantes principales phylogénétiques* (P-PCA), pour travailler sur des données à valeur dans une variété riemannienne - afin qu'elle puisse être appliquée à des données de forme. P-PCA est une version de PCA pour des observations qui sont les feuilles d'un arbre phylogénétique. D'un point de vue statistique, la propriété importante de ces données est que les observations ne sont pas indépendantes. Nous définissons et estimons des moyennes et des covariances intrinsèquement pondérées sur une variété qui prennent en compte cette dépendance des observations. Nous définissons ensuite l'ACP phylogénétique sur une variété comme la décomposition propre de la covariance pondérée dans l'espace tangent de la moyenne pondérée. Nous montrons que l'estimateur de moyenne actuellement utilisé en biologie évolutive pour étudier la morphologie correspond à ne prendre qu'une seule étape de notre algorithme de descente de gradient riemannien pour la moyenne intrinsèque, lorsque les observations sont représentées dans l'espace des formes de Kendall.

Notre deuxième contribution principale est une méthode non paramétrique de réduction de dimension fondée sur une classe très flexible de sous-variétés qui est novatrice même dans le cas de données euclidiennes. Grâce à une PCA locale, nous construisons tout d'abord un sous-fibré du fibré tangent sur la variété des données que nous appelons le *sous-fibré principal*. Cette distribution (au sens géométrique) induit une structure *sous riemannienne*. Nous montrons que les géodésiques sous-riemanniennes correspondantes restent proches de l'ensemble des observations et que l'ensemble des géodésiques partant d'un point donné génèrent localement une sous-variété qui est radialement alignée avec le sous-fibré principal, même lorsqu'il est non intégrables, ce qui apparait lorsque les données sont bruitées. Notre méthode démontre que la géométrie sous-riemannienne est le cadre naturel pour traiter de tels problèmes. Des expériences numériques illustrent la puissance de notre cadre en montrant que nous pouvons réaliser des reconstructions d'une extension importante, même en présence de niveaux de bruit assez élevés.

Mots-clés: *statistiques géométriques, géométrie différentielle, géométrie Riemannienne, géométrie sous-Riemannienne, statistique mathématique, apprentissage automatique*

Riemannske og sub-Riemannske metoder til dimensionsreduktion

I denne afhandling præsenteres nye metoder til dimensionsreduktion, baseret på differential geometri. Det vil sige metoder til at finde en repræsentation af et datasæt i et rum af lavere dimension end det opringelige rum. Sådanne metoder spiller en helt central rolle i statistik, og har et meget bredt anvendelsesområde. En lavere-dimensional repræsentation af et datasæt tillader visualisering og er ofte nødvendigt for efterfølgende statistisk analyse. I traditionel, Euklidisk statistik ligger observationerne i et vektor rum, og det lavere-dimensionale rum kan være et lineært underrum eller en ikke-lineær undermangfoldighed som approksimerer observationerne. Studiet af sådanne glatte mangfoldigheder, differential geometri, spiller en vigtig rolle i sidstnævnte tilfælde, eller hvis rummet hvori observationerne ligger i sig selv er en mangfoldighed. Metoder til at analysere observationer på en mangfoldighed udgør feltet *geometrisk statistik*. I denne kontekst er det approksimerende rum, fundet via dimensionsreduktion, naturligt en submangfoldighed af den givne mangfoldighed. Udgangspunktet for denne afhandling er geometrisk statistik for observationer på en a priori kendt Riemannsk mangfoldighed, men dele af vores arbejde udgør et bidrag selv i tilfældet med observationer i Euklidisk rum, $\mathbb{R}^d$.

Et vigtigt eksempel på data på en mangfoldighed er *former*, i vores tilfælde diskrete kurver eller overflader. I evolutionsbiologi er forskere interesseret i at studere grunde til og implikationer af morfologiske forskelle mellem arter. Former er én måde at formalisere morfologi på. Denne anvendelse motiverer det første hovedbidrag i denne afhandling. We generaliserer en metode til dimensionsreduktion brugt i evolutionsbiologi, *phylogenetisk principal component analysis* (P-PCA), til at virke for data på en Riemannsk mangfoldighed - så den kan anvendes til observationer af former. P-PCA er en version af PCA for observationer som antages at være de yderste knuder i et phylogenetisk træ. Fra et statistisk synspunkt er den vigtige egenskab ved sådanne observationer at de ikke nødvendigvis er uafhængige. We definerer og estimerer intrinsiske vægtede middelværdier og kovarianser på en mangfoldighed, som tager højde for sådanne observationers afhængighed. Vi definerer derefter phylogenetisk PCA på en mangfoldighed som egendekomposition af den vægtede kovarians i tanget-rummet til den vægtede middelværdi. Vi viser at estimatoren af middelværdien som pt. bruges i evolutionsbiologi til at studere morfologi svarer til at tage kun et enkelt skridt af vores Riemannske gradient descent algoritme for den intrinsiske middelværdi, når formerne repræsenteres i Kendall´s form-mangfoldighed.

Vores andet hovedbidrag er en ikke-parametrisk metode til dimensionsreduktion som kan bruges til at approksimere et data sæt baseret på en meget flexibel klasse af submangfoldigheder. Denne metode er ny også i tilfældet med Euklidisk data. Metoden virker ved at konstruere et under-bundt af tangentbundet på datamangfoldigheden M via lokale PCA´er. Vi kalder dette underbundt *principal underbundtet*. Vi observerer at dette underbundt inducerer en *sub-Riemannsk* struktur på M og vi viser at sub-Riemannske geodæter fra et givent punkt lokalt genererer en submangfoldighed som radialt flugter med det estimerede subbundt, selv for ikke-integrable subbundter. Ved støjfyldt data forekommer ikke-integrabilitet med stor sandsynlighed, og vores metode demonstrerer at sub-Riemannsk geometri er en naturlig tilgang til at håndtere dette. Numeriske eksperimenter illustrerer styrkerne ved metoden ved at vise at den opnår rekonstruktioner over store afstande, selv under høje niveauer af støj.

Nøgleord: *geometrisk statistik, differentialgeometry, Riemannsk geometri, sub-Riemannsk geometri, matematisk statistik, maskinlæring*

Acknowledgments

This PhD was carried out in France and in Denmark, with the time split equally between each place. The work was supported by the European Research Council (ERC) under the EU Horizon 2020 research and innovation program (grant agreement G-Statistics No. 786854).

I'm grateful to have met and worked with Xavier, Stefan, James and Erlend. Thank you for being generous, inspiring and knowledgeable.

Thanks to all members of the teams in Sophia-Antipolis and Copenhagen, meeting you has changed my life (for the better). An especially warm thanks goes to, in the order of when we first met, Lennard, Mathias, Jon, Nicholas, Yann, Santi, Josquin, Dimby, Anna, Elodie, Jairo, Zhijie, Tom, Luiś, Marcus, Libby, Thomas, Lily, Sofia, Christian, Michael, Gefan, Liwei, Olivier.

Thank you Alain, Klas, Pierre and Mads for spending time on reading this manuscript and/or participating in the defence, I'm honored.

Thank you Lydia for being you.

Contents

Chapter 1

Introduction

1.1 Statistics on manifolds

Let $x_1, \ldots, x_N$ be a set of observations. The field of statistics traditionally assumes that such observations belong to a (finite dimensional) vector space. That is, the methods rely on being able to add observations, $x_i + x_j$, and scale them, $x_i \cdot c$, $c \in \mathbb{R}_{\geq 0}$. However, this assumption of a vector space structure is prohibitive; important phenomena are naturally modelled as belonging to non-vector spaces. This thesis provides new methods for analyzing such *non-linear* data.

A simple example of non-linear data is observations on a surface in $\mathbb{R}^3$, such as the sphere. Such observations can be given to us in the form of vectors in $\mathbb{R}^3$, i.e. $x_i = [(x_i)_1, (x_i)_2, (x_i)_3]$, but prior knowledge might imply that each observation satisfies the equation

$$\|x_i\| = \sqrt{(x_i)_1^2 + (x_i)_2^2 + (x_i)_3^2} = 1, \quad i \in \{1, \ldots, N\}, \tag{1.1.1}$$

meaning that they are points on the unit sphere embedded in $\mathbb{R}^3$. Such data on the sphere, and its analogues in different dimensions, is studied in the field of *directional statistics* ([K. V. Mardia, Jupp, and K. Mardia 2000], [Pewsey and García-Portugués 2021]). The observations can be literal *directions*, e.g. wind directions, they can be observations of time points, or more generally any type of Euclidean observations that has been normalized, meaning that only directions matter, not magnitude (see [Banerjee et al. 2005] for applications to text classification, gene expression analysis and other domains). Applying traditional Euclidean statistical methods to such data, i.e. treating them as points in $\mathbb{R}^3$, disregards this prior knowledge and thus the output might not be sensible. For example, computing the Euclidean mean of the observations on the sphere yields a point that does not satisfy the constraint (1.1.1) - it is not a point on the sphere. There is a need for statistical methods which can incorporate prior knowledge about the space in which a set of observations might reside. The particular type of non-linear data spaces studied in this thesis are *manifolds* - of which the sphere is an example. In particular, we assume that the space containing the observations is a *Riemannian* manifold, implying

that e.g. distances between points can be computed. This type of statistics is called *Riemannian geometric statistics*. Given data on a Riemannian manifold, the field comprises methods for computing mean values, dimension reduction, regression, classification etc. (see [Pennec, Sommer, and T. Fletcher 2019] and [S. F. Huckemann and Eltzner 2021] for an overview).

Determining a set of constraints satisfied by the observations is a common way to arrive at a manifold data space in applications. Other examples of this are matrix manifolds such as the manifold of symmetric positive definite matrices,

$$SPD(d) = \{A \in \mathbb{R}^{d\times d} \mid A^T = A, \\ x \neq 0 \implies x^T A x > 0\}.$$

Such matrices constitutes the observations in e.g. the study of diffusion tensor images, used in medical imaging, where they form covariance matrices of Brownian motions describing the motion (diffusion) of water molecules in tissue (see e.g. [Lazar 2007]). Another example of a matrix manifold is the rotation group $SO(3)$, which is abundant in engineering applications (see [Chirikjian and Kyatkin 2021]). In [Hauberg 2015] it is used to model motion capture data.

Another approach to deriving a manifold representing a data set is by introducing *equivalence relations*, that is, to consider certain subsets of points in the original space as identical. The result is a quotient space, which is often a manifold. An example of this is the notion of a *shape space*, which is a central motivation for geometric statistics in general as well as for the work in this thesis in particular. Consider the Euclidean space $\mathbb{R}^{d\cdot k} \cong \mathbb{R}^{d\times k}$, interpreted as a set of matrix-valued points, where each $x \in \mathbb{R}^{d\times k}$ is a collection of k points in $\mathbb{R}^d$ (the columns), called *landmarks*. These k points constituting x could for example be points along a discretized curve in $\mathbb{R}^2$, outlining some 'shape'. What do we mean by 'shape'? One answer is that a shape is that which is left after removing the effects of translation, scaling and rotation. This can be formalized as a set of equivalence relations, where two points $x, x' \in \mathbb{R}^{d\times k}$ are identified if there exists a translation, a scaling and a rotation such that applying these operations to each landmark in x makes them equal to the corresponding landmarks in x'. Identifying such points in $\mathbb{R}^{d\times k}$ leads to *Kendall's shape space*, which is a quotient manifold [Kendall 1984]. Kendall's shape space is part of the theoretical foundation of the field of geometric morphometrics - the study of morphology (shape) in biology [Klingenberg 2020]. Another example of a shape manifold is the *LDDMM* framework [Younes 2010], that enables analysis of full shapes, e.g. continuous curves and surfaces. Such spaces of full shapes has been used in the context of computational anatomy and medical imaging for studying organs like

the heart or brain (see [Guigui et al. 2021] for an analysis of heart movement).

Lastly, we note that defining a manifold in terms of a set of constraints or a set of equivalence relations can be complementary views on the same object. For example, it turns out that Kendall's shape space for $k = 3$ landmarks in $\mathbb{R}^2$, interpreted as a space of triangle shapes, can be identified with the sphere. Thus, performing statistics on such triangle shapes is equivalent to performing statistics on the sphere.

In the next chapter, we introduce the subject of geometric statistics in more detail. We put a special emphasis on methods for dimension reduction, which is the main subject of the thesis. Our contributions to this problem are described in the following subsection.

1.2 Thesis organization and contributions

Chapter 2 In this chapter, we present background theory concerning the parts of Riemannian geometric statistics most closely related to our contributions in Chapters **??** and **??**. We start out by elaborating on the notion of doing statistics for observations belonging to non-vector spaces. We then summarize the notions in Riemannian geometry that are fundamental to Riemannian geometric statistics, and subsequently explain broadly how the various operations are used, and computed, for this type of statistics. We devote some space to the notion of tangent space approximations, and statistics based on these, since they are used extensively both a tool and a benchmark in the methods we develop in chapters **??** and **??**. The exposition of tangent space statistics in this Chapter, and in Appendix A.2 where more technical details are given, includes some points which are of both theoretical and computational importance but which has, to the best of our knowledge, not been explicitly presented in the literature on geometric statistics. The section is concluded with presentations of the two foundational methods for dimension reduction on Riemannian manifolds; *tangent PCA* and *Principal geodesic analysis* (PGA). The methods for dimension reduction that we present in chapters **??** and **??** are novel extensions and combinations of these two methods. In Appendix A.3 we elaborate on the complications with computing PGA (which is an open problem), and rewrite the objective function to something more tractable. In Appendix A.4 we derive a Taylor expansion of the PGA objective function which reduces to tangent PCA when only the 1st order term is considered and includes curvature in the higher order terms.

Chapter 3 This chapter presents an extension of *Phylogenetic principal component analysis* (p-PCA) to manifold valued data. P-PCA is a version of PCA for observations that are assumed to be leaf nodes of a phylogenetic tree. Mathematically, a *phylogenetic tree* is a rooted, bifurcating tree graph, which in evolutionary biology represents evolutionary relationships between species. From a statistical point of view, the important property of such data is that the observations (leaf node values) are not necessarily independent. In particular, they are the endpoints of Brownian motions that are more or less coupled depending on how closely related the species are. P-PCA consists of eigendecomposition of the socalled phylogenetic covariance matrix, which is centered around the socalled phylogenetic root, a type of mean. These notions of mean and covariance takes the dependency of the observations into account. In evolutionary biology there is a need for applying this method to data on manifolds, in particular shapes. We therefore extend to a general Riemannian manifold the probabilistic model of Brownian motions structured according to a tree, and define intrinsic estimators of the phylogenetic root and covariance. We then define phylogenetic PCA on a Riemannian manifold to be eigendecomposition of this covariance matrix in the tangent space at the mean. We show simulation results on the sphere (Kendall's shape space for triangles), demonstrating well-behaved error distributions and fast convergence of estimators. Furthermore, we apply the method to a data set of mammal jaws, represented as points in the LDDMM landmark manifold. Lastly, we discuss the relationship between our approach and the approach taken to the same problem in *geometric morphometrics* - the field of biology analysing shapes (morphology). We show that the phylogenetic root estimator used in geometric morphometrics corresponds to taking a single step of our Riemannian gradient descent algorithm for the intrinsic root, when the observations are represented in Kendall's shape space.

Chapter 4 The method presented in Chapter **??** is an extension of tangent PCA. Tangent PCA is based on geodesics of the manifold M to which the observations belong, which is a generalization of straight lines to such a manifold. In this chapter, we describe a method for dimension reduction that is based on a more flexible class of curves. Furthermore, whereas the previous chapter presented an extension of a method that is already well defined in Euclidean space, the method presented in this chapter is novel also in the case of $M = \mathbb{R}^d$. The method provides a solution to a number of dimension reduction problems, such as; construction of a lower-dimensional submanifold approximating the observations, representation of the observations in a lower-dimensional Euclidean

space, and metric learning, in the sense of estimating a distance metric on M which reflects a lower dimensional geometry. The method works by constructing a subbundle of the tangent bundle on M via *local* tangent PCA's, which we call the *principal subbundle*. We then observe that this subbundle induces a *sub-Riemannian* (SR) structure on M, and we show that sub-Riemannian geodesics with respect to this structure follow the point cloud and can be used to provide solutions to the aforementioned problems. For example, we show that it is possible to generate submanifolds consisting of such SR geodesics, even if the subbundle is non-integrable, which loosely means that it doesn't 'naturally' determine submanifolds of M. Non-integrability is likely to occur when the subbundle is estimated from noisy data, and our method demonstrates that sub-Riemannian geometry is a natural framework for dealing which such noise. Last, but not least, we implemented a class with methods for sub-Riemannian geometry in *geomstats* [Miolane et al. 2020], making computations in sub-Riemannian geometry available for future applications in geometric and Euclidean statistics.

Chapter 5 We round off by giving a brief summary of the contributions of the thesis and describing current and future directions of work related to phylogenetic PCA and principal subbundles.

1.2.1 Publications and invited talks

The material presented in Chapter **??** is an extension of the paper, *Tangent phylogenetic PCA* [Akhøj, Pennec, and Sommer 2023], published in the proceedings of the *Scandinavian Conference on Image Analysis*, 2023.

The material presented in Chapter **??** is based on the preprint *Principal subbundles for dimension reduction* [Akhøj, Benn, et al. 2023] which has been submitted to the journal *Foundations of computational mathematics* in July 2023. In April 2023, this work was presented by the author at the workshop *Statistics of shapes and Geometry of Shape Spaces*, at the Max-Planck-Institute for Mathematics in the Sciences, Leipzig. In May 2023 it was presented by the author at the *Analysis and PDE Seminar* at the university of Bergen, Department of mathematics.

Chapter 2

Background: geometric statistics

Contents

2.1 Generalizing Euclidean statistical methods

Many Euclidean statistical methods can be reformulated in terms of operations that does not depend on the vector space structure, suggesting that they make sense on more general spaces. Often, there exist multiple formulations of the same method that are equivalent in Euclidean space, but which lead to different solutions in the more general space. One approach to generalizing a given method is to reformulate it in terms of distances, making it well-defined on a general *metric space*.

Example 2.1.1 (The empirical mean in $\mathbb{R}^d$)**.** Let $\{x_i\}_{i=1\dots N}$ be observations in $\mathbb{R}$ and let $d(x,y) = |x-y|$ be the Euclidean distance on $\mathbb{R}$. Consider the optimization problem

$$\hat{\mu} = \operatorname*{argmin}_{\mu\in\mathbb{R}} \sum_{i=1}^{N} d(\mu, x_i)^2. \tag{2.1.1}$$

Differentiating with respect to μ gives

$$\frac{d}{d\mu}\sum_{i=1}^{N} |\mu - x_i|^2 = 2\mu N - 2\sum_{i=1}^{N} x_i$$

and $\frac{d^2}{d^2\mu}\sum_{i=1}^{N} |\mu - x_i|^2 = 2N > 0$. Thus, the unique minimizer of (2.1.1) is

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{2.1.2}$$

the arithmetic mean, or empirical first moment. Likewise, for $x_1,\dots,x_N \in \mathbb{R}^d$ one can similarly show that the solution is given by $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

In Equation (2.1.2), vector space operations addition and scalar multiplication are applied to the observations, while the minimization problem (2.1.1) only applies a distance metric to the observations.

The above formulation of the arithmetic mean as a distance minimization problem can be used as a *definition* of a mean on a general metric space (M,d), where M is some set and $d : M \times M \to \mathbb{R}_{\geq 0}$ is a distance metric on M. The point here is that M is not necessarily a vector space.

Definition 2.1.2 (Fréchet mean)**.** [Fréchet 1948] Let (M,d) be a metric space, and let $x_1,\dots,x_N$ be points on M. The *Fréchet mean* is defined as a minimizer of the following minimization problem,

$$\hat{\mu} = \operatorname*{argmin}_{\mu\in M} \sum_{i=1}^{N} d(\mu, x_i)^2. \tag{2.1.3}$$

The Fréchet mean optimization problem is an instance of the following more general problem, for observations $x_1, \ldots, x_N \in M$,

$$\widehat{U} \in \operatorname*{argmin}_{U \in \mathcal{Q}} \sum_{i=1}^{N} d(x_i, \pi_U(x_i))^2, \tag{2.1.4}$$

where $\mathcal{Q}$ is some family of closed subsets of M and π is a projection of a point $x \in M$ to a subset $U \in \mathcal{Q}$ with respect to the distance metric d;

$$\pi_U(x) \in \operatorname*{argmin}_{p \in U} d(x, p), \tag{2.1.5}$$

assumed to be unique for the considered points. The optimization problem (2.1.4) searches for a subset U which approximates the observations with respect to the distance d. This reduces to the Fréchet mean problem for $\mathcal{Q} = M$, in which case $\pi_U(x_i) = U$ for any point $U \in \mathcal{Q}$. The problem (2.1.4) can be called an *unsupervised* approximation problem - unsupervised because there are no independent variables on which the observations depend, as there is in e.g. a regression problem. We think of U as a subset that *represents* the data given a chosen set of constraints, expressed by $\mathcal{Q}$. For example, elements of $\mathcal{Q}$ might be of *lower dimension* than the ambient space of observations, and it might be a *linear subspace* if the ambient space is a vector space. For a particular problem, the question is then which family of subsets $\mathcal{Q} \subset \mathcal{P}(M)$ yield solutions with desired properties, e.g. properties similar to the case of Euclidean space. Principal component analysis, the foundation of the methods presented in this thesis, is another important example of problem (2.1.4).

Example 2.1.3 (Principal component analysis)**.**

Let $\{x_i\}_{i=1\ldots N} \subset \mathbb{R}^d$ be observations in $\mathbb{R}^d$ and let μ be their arithmetic mean. The two core steps of principal component analysis (see e.g. [Jolliffe and Cadima 2016]) are the following,

1. Compute the empirical covariance matrix

$$K = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T. \tag{2.1.6}$$

2. Compute the eigendecomposition $K = P\Lambda P^{-1}$, where Λ is a diagonal matrix containing the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ and P is a matrix containing the corresponding eigenvectors $e_1, \ldots, e_d$ as columns.

Assuming that λ_k is strictly greater than λ_{k+1}, we call $V_k \coloneqq \operatorname{span}\{e_1, \ldots, e_k\}$ the k'th *eigenspace*. This algebraically defined object can be characterized in a

different way, namely as the k-dimensional subspace that best approximates the observations, in the sense of minimizing the following problem

$$V_k = \underset{U \in \mathrm{Gr}(k,\mathbb{R}^d)}{\operatorname{argmin}} \sum_{i=1}^{N} d(x_i, \pi_U(x_i))^2, \qquad i = 1 \dots d-1, \tag{2.1.7}$$

where $\mathrm{Gr}(k, \mathbb{R}^d)$ is the Grassmannian manifold of all k-dimensional subspaces of $\mathbb{R}^d$ [Bendokat, Zimmermann, and Absil 2020]. This problem is of the form (2.1.4) for $\mathcal{Q} = \mathrm{Gr}(k, \mathbb{R}^d)$, in which case $\pi_U(x_i)$ is orthogonal projection of x_i to linear subspace $U \subset \mathbb{R}^d$.

This formulation of eigenspaces suggests that we can generalize them to more general metric spaces (M, d). However, the parameter space $\mathcal{Q}$, the Grassmannian manifold of linear subspaces of $\mathbb{R}^d$, only makes sense when the space of observations is a vector space. A natural idea is then to look for a family of subsets $\mathcal{Q}$ of M analogous to linear subspaces of a vector space. In the case of Riemannian manifolds, such subsets, in fact submanifolds, can be built up from collections of geodesics - a generalization of straight lines to manifolds. This leads to *Principal geodesic analysis* and its approximative version *tangent PCA* which will be described in Sections 2.4.1 and 2.4.2.

Minimal prerequisites for approaching a problem of the form (2.1.4) on some metric space (M, d) seems to be the ability to compute the distance metric d, the existence of a suitable family of subsets $\mathcal{Q}$ for the problem at hand, and last, but not least, the ability to solve optimization problems. In this thesis we study the case where M is a complete Riemannian manifold and where the distance metric d is the geodesic distance. In the next section, we give an overview of how the tools available on Riemannian manifolds enables us to solve problems in non-Euclidean statistics, such as (2.1.4).

2.2 Riemannian geometry

In this section, we give a brief overview of the central objects in Riemannian geometry used for Riemannian geometric statistics. For more details, we refer the reader to, e.g., [Lee 2018] or [Do Carmo and Flaherty Francis 1992].

A Riemannian metric g on M is a smoothly varying family of inner products on the tangent bundle TM. Being an inner product on each tangent space $T_xM, x \in M$, g enables measurement of lengths and angles of tangent vectors. From this, it defines the length of a smooth path $\varphi : [0, 1] \to M$ as $L(\varphi) := \int_0^1 g(\dot{\gamma}(t), \dot{\gamma}(t))dt$, and a distance between two points as the infimum of

lengths of paths joining them;

$$d(x,y) = \inf \left\{ L(\varphi) \;\middle|\; \begin{array}{c} \gamma : [0,1] \to M \text{ is smooth} \\ \gamma(0) = x, \gamma(1) = y \end{array} \right\}, \tag{2.2.1}$$

for any $x, y \in M$.

A crucial fact about (finite dimensional) Riemannian geometry, as opposed to sub-Riemannian geometry, is that the metric determines a canonical isomorphism between T_pM and its dual space $T_p^\star M$, called the *flat* map,

$$\flat : T_pM \to T_p^\star M : v \mapsto v^\flat.$$

The inverse of the flat map is called the *sharp* map, denoted $\sharp$. Given a real-valued function $f : M \to \mathbb{R}$ we can compute its differential at a point $p \in M$, which is a covector $d_pf \in T_p^\star M$. Given a Riemannian metric, the induced sharp map thus provides a canonical way to identify this cotangent vector with a tangent vector, called the Riemannian gradient of f at p, $\nabla f|_p = \sharp(d_pf) \in T_pM$.

Recall that a *connection* on M allows to define the *covariant derivative* of e.g. a vector field or a tensor field along another vector field. A vector field is *parallel* along a curve if its covariant derivative along the curve is zero. For any points $x, y \in M$, any tangent vector $v \in T_xM$ and any smooth curve φ betweeen x and y, there exists a parallel vector field along the curve, called the *parallel transport* of v along φ. The parallel vector field solves the parallel transport equation, a 1st-order linear ODE. A Riemannian metric induces a canonical connection, the *Levi-Civita connection*, as the unique connection satisfying the two conditions of torsion being constantly zero, $\nabla_X Y - \nabla_Y X = [X, Y]$, and compatibility with the metric. The latter condition is equivalent to the parallel transport operator $\Pi_x^y : T_xM \to T_yM$ along any smooth curve between $x, y \in M$ being an isometry ([Lee 2018], lemma 5.2).

Geodesics are curves whose tangent vectors are parallel with respect to a chosen connection. This implies that they satisfy the *geodesic equation*, a 2nd-order ODE derived from the parallel transport equation. When the connection is Levi-Civita, determined by a Riemannian metric, geodesics have the property of being locally length minimizing curves. In particular, the path distance d on M is a *geodesic* distance, meaning that a shortest path between $x, y \in M$ is a geodesic w.r.t. the Levi-Civita connection. Furthermore, the Hopf-Rinow theorem states that (M, d) is a complete metric space if and only if M is geodesically complete, meaning that solutions to the geodesic equations exist for all time.

The *exponential map* is the data-to-solution map for the geodesic equation and carries an initial point $p \in M$ and initial direction $v \in T_pM$ to the value of

the corresponding geodesic γ_p^v at time 1,

$$\exp_p \quad : \quad T_pM \to M \quad : \quad v \mapsto \exp_p(v) = \gamma_p^v(1). \tag{2.2.2}$$

The derivative of the exponential map at $v = 0$ is the identity map, so the inverse function theorem implies that exp is locally a diffeomorphism. Let $\mathcal{C}_p \subset T_pM$ be the largest subset containing 0 on which $\exp_p$ is a diffeomorphism, called the *injectivity domain*. Let $C_p = \exp_p(\mathcal{C}_p) \subset M$ be its exponential image. The inverse of exp on C_p is called the *Riemannian logarithmic map*,

$$\log_p \quad : \quad C_p \to T_pM \quad : \quad x \mapsto \log_p(x) = \exp_p^{-1}(x). \tag{2.2.3}$$

Since $T_pM \cong \mathbb{R}^d$, $\log_p$ thus defines a chart on M, called a normal chart.

Lastly, a crucial property of the log is its relation to the geodesic distance metric,

$$\| \log_x(y) \|_x = d(x, y), \tag{2.2.4}$$

where $\| \cdot \|_x$ is the norm induced by the Riemannian metric g_x at x.

2.3 Basic tools for Riemannian geometric statistics

In this section, we give an overview of how the objects and operations described in the previous section are used in Riemannian geometric statistics.

First of all, we mention that there is not necessarily a canonical choice of metric for a certain application. In the case of e.g. the LDDMM shape space (Section **??**), there is a natural parameterized *family* of metrics, but not a canonical way to choose among these. Other examples are the manifolds of correlation matrices and of covariance matrices, where there is a plethora of different families of metrics (see the PhD thesis [Thanwerdas 2022] for an overview). This degree of freedom can be seen as modelling flexibility or a nuissance, depending on the application. In other cases, such as Kendall's shape space, there is a natural choice of metric.

After a choice of metric, the geodesic distance induced by the Riemannian metric turns the manifold M into a metric space (M, d) with respect to which distance based statistics can be performed. In practice, the distance $d(x, y)$, $x, y \in M$, is computed via the logarithm, implicitly assuming that y is in the exponential image of the injectivity area of x.

The gradient induced by the Riemannian metric allows for formulating optimization algorithms, e.g. gradient descent, for minimizing objective functions such as (2.1.3) for the Fréchet mean. See e.g. [Boumal 2023] or [Absil, Mahony,

and Sepulchre 2008] and Section **??** where we discuss gradient descent for a weighted Fréchet mean.

If closed-form expressions for geodesics are not available for a particular manifold, we can approximate them by numerically integrating the geodesic equation. This, and a wide range of other computations on Riemannian manifolds, is available in Python libraries such as *geomstats* [Miolane et al. 2020] and *gaxgeometry* (http://bitbucket.org/stefansommer/jaxgeometry).

Computing the log is a crucial problem for practical geometric statistics, e.g. for computing the geodesic distance and normal charts. On a general Riemannian manifold it is not possible to characterize the injectivity domain neither theoretically nor numerically. When we write $\log_p(x_i)$ it will thus be under a, usually implicit, assumption that $x_i \in C_p$. In the case of no additional theoretical knowledge, in geomstats and jaxgeometry the logarithm is computed via a 'shooting method', which amounts to searching for a geodesic whose endpoint is as close as possible to the desired endpoint, w.r.t. some easy-to-compute ambient distance. Such an optimal geodesic is found via a numerical optimization scheme on the initial tangent vector. This is a relatively expensive optimization problem. We discuss this further in Appendix A.3.

Tangent spaces are important for Riemannian geometric statistics, and normal charts are a particularly convenient representation of these. Since T_pM is a vector space approximating M around p, it provides us with the possibility to do Euclidean statistics on the transformed dataset $\{\log_p(x_i)\} \subset T_pM$ in a way that approximates exact statistics on the manifold. This comes with the benefit of being more computationally efficient, as well as removing the need for reformulating a method in an intrinsic way. Equation (2.2.4) expresses that the distance between the base point x and any point $y \in C_x$ is preserved in a normal chart. However, if neither x nor y is the base point then the corresponding distance in the tangent space is distorted - see subsection 2.3.1 for details. Theoretically, the basis induced by a normal chart on T_pM is orthonormal. This is convenient, since Euclidean statistical formulas usually assumes an orthonormal basis. However, if the log is computed by an optimization in the tangent space, as described above, the basis is arbitrary, which then has to be taken into account. This and other points are not fully adressed in the literature. We discuss it further in Section 2.4.1 and in Appendix A.2.

Lastly, we mention the role of the parallel transport map along a geodesic. It is often necessary to transport computations done in one tangent space to another, or collecting tangent vectors from different tangent spaces into a common one, in which a computation on the set can be performed (see an example of this in

Appendix A.4 on Taylor approximation of PGA). We therefore need suitable maps between tangent spaces. If the manifold is a Lie group, the differential of the group multiplication map (left or right) can be used. On a general Riemannian manifold, the parallel transport map provides solution.

2.3.1 Distortion of distances in the tangent space

As mentioned, there is a trade-off between the accuracy of the exact computation of geodesic distances and the convenient properties of the tangent space distances, such as computational efficiency. This trade-off is a recurrent theme in this thesis, so we elaborate on it in this section.

Equation (2.2.4) says that the distance between the origin in a tangent space T_pM and each $\log_p(x_i) \in T_pM$, as measured with respect to the norm $\|\cdot\|_p$, equals the geodesic distance between p and x_i. Thus, the tangent space does not distort these *radial* distances. However, the distance $\|v_1 - v_2\|_p$ between arbitrary points $v_1, v_2 \in T_pM$ does not equal the corresponding geodesic distance $d(\exp_p(v_1), \exp_p(v_2))$ on the manifold. In particular, the distortion of this geodesic distance depends on the curvature around p in the following way (see e.g. [Do Carmo and Flaherty Francis 1992], Section 5 proposition 2.7),

$$\begin{aligned} d(exp_p(t \cdot v_1), exp_p(t \cdot v_2))^2 = t^2 \, \|v_1 - v_2\|_p^2 \\ - t^4 \, \frac{1}{3} \langle R(v_1, v_2)v_1, v_2 \rangle \\ + O(t^4), \end{aligned} \tag{2.3.1}$$

where R is the curvature tensor and $\langle R(v_1, v_2)v_1, v_2 \rangle_p$ therefore the sectional curvature with respect to the plane spanned by v_1, v_2 in T_pM. Thus, in the case of positive sectional curvature (such as for any point p on the sphere), the Euclidean distances in the tangent space gets underestimated, and vice versa in the case of negative curvature. The formula also implies that the smaller t is, i.e. the closer the points being compared are, the smaller is the distortion. For Riemannian geometric statistics this means that the more concentrated, in the sense of low variability, the data set is, the smaller is the error incurred by using a distance based Euclidean formula in the tangent space. What a sufficient level of concentration is, relative to an accepted level of error, depends on the curvature of the manifold in the vicinity of the observations.

2.4 Dimension reduction

In this section, we focus on the particular problem of dimension reduction on Riemannian manifolds. We give details on the two foundational methods, tangent PCA and principal geodesic analysis. The methods presented in Chapters **??** and **??** are variations and combinations of these two methods. In the last subsection, we give an overview of more recent methods.

2.4.1 Tangent PCA

Tangent PCA can be described roughly as performing PCA on the log-transformed observations $\log_p(x_1), \dots, \log_p(x_N) \in T_pM$, where p might be the Fréchet mean. However, this description leaves out some important theoretical and practical details that are not described in the literature, so we do that below. Our presentation is structured according to the two steps of PCA: step one, computing the empirical covariance matrix, and step two, the eigendecomposition.

Intrinsic definition of the empirical covariance matrix In Euclidean statistics, there is a canonical choice of an orthonormal basis. Maybe for this reason, it is rarely mentioned that the empirical covariance matrix (2.1.6) depends on the choice of basis, in the following sense. Let v_A and v_B be representations of a vector v in $\mathbb{R}^d$ with respect to bases A and B, and let Q be the change of basis matrix from A to B, so that $v_B = Qv_A$. Changing the basis of v changes an outer product in the following way,

$$v_B {v_B}^T = (Qv_A)\,(Qv_A)^T = Qv_A {v_A}^T Q^T.$$

In contrast, the coordinate representation of $\varphi_A = v_A {v_A}^T$, considered as a linear map, is changed by

$$\varphi_B = Q\varphi_A Q^{-1} = Qv_A {v_A}^T Q^{-1}.$$

Thus, if and only if Q is an orthogonal change of basis matrix, i.e. $Q^T = Q^{-1}$, does changing the basis of v_A lead to the desired change of basis of $v_A {v_A}^T$, considered as a linear map. Eigendecomposition of a linear map, represented by a matrix, is independent of the choice of basis, in the sense that changing the basis of the linear map leads to a corresponding change of basis of the eigenvectors (the eigenvalues are not affected by basis change). Recall that the empirical covariance matrix is a sum of outer products. Thus, PCA depends on the choice of basis, up to orthogonal basis change, since the empirical covariance matrix does.

On a Riemannian manifold there is no canonical basis for a tangent space T_pM. However, from the Riemannian metric one can compute a, non-unique, orthonormal basis and represent the vectors $\{\log_p(x_i)\}_{i=1..N}$ w.r.t. to this basis. The empirical covariance matrix can then be computed w.r.t. this orthonormal basis, and the discussion above implies that the result of the subsequent eigendecomposition will be independent on the particular choice of orthonormal basis. However, we can do better than this; we can give a coordinate-free definition of empirical covariance, which has the additional advantage of allowing computations to be performed in any basis - removing the need for computing, and transforming the data to, an orthonormal basis.

Following [Pennec, Fillard, and Ayache 2006], we define the empirical covariance matrix at $p \in M$ as the tensor

$$\Sigma(p) = \frac{1}{N-1}\sum_{i=1}^{N} \log_p(x_i) \otimes \log_p(x_i) \in T_pM \otimes T_pM. \tag{2.4.1}$$

After choosing a local chart around p, which induces a basis A for T_pM, the coordinate representation of this tensor is

$$[\Sigma(p)]_A = \frac{1}{N-1}\sum_{i=1}^{N} \left[\log_p(x_i)\right]_A \left(\left[\log_p(x_i)\right]_A\right)^T [g_p]_A, \tag{2.4.2}$$

where $[\cdot]_A$ denotes representation of an object with respect to basis A. See Appendix A.2.1 for a derivation of this coordinate expression. Thus, the coordinate representation of the tensor $\Sigma(p)$ is a $d \times d$ matrix as we would expect. If the basis A is orthonormal with respect to metric g, then $[g_p]_A$ is the identitity matrix. Thus, the common description of tangent PCA as eigendecomposition of the matrix $\frac{1}{N-1}\sum_{i=1}^{N} \left[\log_p(x_i)\right]_A \left(\left[\log_p(x_i)\right]_A\right)^T$ is only correct if the basis A is orthonormal. In theory this is the case when the basis is induced by a normal chart. However, if the log-map is computed numerically via an optimization scheme it is not necessarily the case (see Appendix A.1). Lastly, note that if (M, g) is Euclidean space then $\log_p(x) = x - p$, so that if p is the Fréchet mean then the covariance (2.4.2) equals the ordinary Euclidean empirical covariance formula (2.1.6).

Eigendecomposition in the tangent space Given this intrinsic definition of empirical covariance on a Riemannian manifold, we can move to step 2 of PCA, the eigendecomposition. Tangent PCA consists simply of performing eigendecomposition of $\Sigma(p)$. The eigendecomposition yields eigenvectors $e_1, \ldots, e_d \in T_pM$ and thus eigenspaces $V_k := \operatorname{span}(e_1, \ldots, e_k) \subset T_pM$

in the tangent space. The log transformed observations can be projected to eigenspaces, $\pi_{V_k}\left(\log_p(x_i)\right)$, and mapped back to M via the exponential map, $\exp_p\left(\pi_{V_k}\left(\log_p(x_i)\right)\right)$. Since the exponential map is a diffeomorphism on the tangent cut locus $\mathcal{C}_p$ it holds that $\exp_p(\mathcal{C}_p \cap V_k)$ is a k-dimensional embedded submanifold of M. This submanifold can be used as a k-dimensional approximation of the data. Since the image of a linear subspace under the exponential map is formed by a collection of geodesics, we interpret the resulting submanifold of M as an analogue of a linear subspace.

Tangent PCA approximates distances We now discuss the sense in which tangent PCA is approximative. As discussed in Example 2.1.3 eigenspaces of a linear map can be formulated as solutions to a distance minimization problem. In the setting of tangent PCA, eigenspace V_k, $k = 1 \dots d-1$, solves the optimization problem

$$V_k = \underset{U \in \mathrm{Gr}(k, T_pM)}{\operatorname{argmin}} \sum_{i=1}^{N} d(\log_p(x_i), \pi_U(\log_p(x_i)))^2, \tag{2.4.3}$$

where $d(x, y) = \|x - y\|_p^2$ is the distance between $x, y \in T_pM$ w.r.t. the norm induced by g_p. Thus, V_k is a linear subspace approximating the log-transformed observations in T_pM. However, as discussed in the previous section, the fact that V_k is close to the observations $\{\log_p(x_i)\}_{i=1\dots N}$ in the tangent space does not imply that the submanifold $\exp_p(V_k \cap \mathcal{C}_p)$ is close to the observations $\{x_i\}_{i=1\dots k}$ in M. More specifically, that $d(\log_p(x_i), \pi_{V_k}(\log_p(x_i)))$ is small does not imply that $d(x_i, \exp_p(\pi_{V_k}(\log_p(x_i))))$ is small. To ensure this, one has to minimize the geodesic distances instead of the tangent space distances. This is done in principal geodesic analysis, described in the next section.

In Chapter **??**, we derive a tensor representation of a particular type of covariance matrix used in evolutionary biology, and define 'tangent phylogenetic PCA' as eigendecomposition of the corresponding matrix. In Chapter **??** we define socalled *local* tangent PCA, which are combined in a certain way to construct approximating submanifolds that does not consist of (Riemannian) geodesics.

2.4.2 Principal geodesic analysis

In this section we assume that the base point is the Fréchet mean, μ, as it is presented in the original paper on PGA [T. P. Fletcher et al. 2004]. As we saw, tangent PCA approximates the observations with the submanifold $\exp_\mu(V_k \cap \mathcal{C}_\mu)$ where V_k is found by minimizing certain distances in the tangent space $T_\mu M$. In

principal geodesic analysis, $V_k \subset T_pM$ is found by minimizing the corresponding geodesic distances on M. The V_k's are obtained in the following sequential way. Let $V_1 := \text{span}\{v_1\}$, where v_1 minimizes the following objective function,

$$v_1 \in \underset{v \in T_\mu M}{\text{argmin}} \frac{1}{N} \sum_{j=1}^{N} \text{d}(x_j, \pi_{\exp_\mu(span(v) \cap \mathcal{C}_\mu)}(x_j))^2 \tag{2.4.4}$$

$$= \underset{v \in T_\mu M}{\text{argmin}} \frac{1}{N} \sum_{j=1}^{N} \| \log_{x_j}(\pi_{\exp_\mu(span(v) \cap \mathcal{C}_\mu)}(x_j))\|^2, \tag{2.4.5}$$

where d is the geodesic distance on (M, g), and π is projection with respect to this distance. The following subspaces $V_2 := \text{span}(v_1, v_2), \ldots, V_{d-1} := \text{span}(v_1, \ldots, v_{d-1})$ are found by optimizing for one basis vector v_i at a time,

$$v_i \in \underset{v \in T_\mu M}{\text{argmin}} \frac{1}{N} \sum_{j=1}^{N} \| \log_{x_j}(\pi_{\exp_\mu(V_{i-1} \cap \mathcal{C}_\mu)}(x_j))\|^2, \tag{2.4.6}$$

for $i = 2 \ldots d-1$. For practical purposes, to reduce the dimension of the parameter space, it can be helpful to add constraints enforcing unit length and orthogonality to the optimization problem, i.e. $\|v\|_g = 1$ and $v_i \perp \text{span}\{V_{i-1}\}$. Regardless, the PGA optimization problem is high-dimensional and computationally intensive. The exact dimension and complexity depends on how many operations are available closed-form, e.g. log or exp. In Appendix A.3 we elaborate on the complications with computing PGA, and rewrite the objective function to something more tractable. No publicly available implementation of PGA exists to this date (for work in this direction, see [Sommer, Lauze, et al. 2010]).

On a general manifold, as opposed to Euclidean space, the described sequential procedure of optimizing for one basis vector at a time is not equivalent to directly solving for the optimal k-dimensional subspace of T_pM. Likewise, in Euclidean space, the eigenspace can equivalently be defined as the solution to a variance maximization problem, whereas this is not equivalent to error (distance) minimization on a general manifold. In practice, the sequential approach leads to a simpler optimization problem, and the formulation in terms of distance minimization seems to be more stable than variance maximization [Sommer, Lauze, et al. 2010].

We end our presentation of PGA by deriving tangent PCA as an approximation of PGA. The tangent PCA objective function (2.4.3) can be derived from the PGA objective function by making the approximation

$$\log_\mu(\pi_{\exp_\mu(span(v) \cap \mathcal{C}_\mu)}(x_j)) \approx v \cdot \langle v, \log_\mu(x_j)\rangle$$

which leads to

$$\begin{aligned} \mathrm{d}(x_j, \pi_{\exp_\mu(span(v)\cap C_\mu)}(x_j)) &\approx \| \log_\mu(x_j) - \log_\mu(\pi_{\exp_\mu(span(v))}(x_j))\|_p && (2.4.7)\\ &\approx \| \log_\mu(x_j) - v \cdot \langle v, \log_\mu(x_j)\rangle\|_p && (2.4.8)\\ &\approx \| \log_\mu(x_j) - \pi_{\mathrm{span}(v)}(\log_\mu(x_j)))\rangle\|_p && (2.4.9) \end{aligned}$$

where the first approximation is based on Equation (2.3.1), and $\pi_{\mathrm{span}(v)}$ is orthogonal projection in T_pM w.r.t. the norm induced by g_μ. The latter expression is exactly what tangent PCA minimizes, cf. Equation (2.4.3). In Appendix A.4 we derive a Taylor expansion of the objective function of PGA whose first order term corresponds to the tangent PCA objective function - this is another way of deriving tangent PCA as an approximation to PGA. The higher order terms in the Taylor expansion takes the curvature into account.

Since PGA is based on computing exact distances in M, as opposed to approximated distances in $T_\mu M$, it is more sensitive to the curvature of M than PGA. However, it is not sensitive to potential non-geodesic variation in the dataset. This is in line with it being a manifold generalization of PCA, which assumes that the data is well approximated by linear subspaces. The method we present in Chapter **??** picks up on non-geodesic variation in the data, by combining PGA and *local* tangent PCA's.

2.4.3 Other methods for dimension reduction on Riemannian manifolds

We round off by a short survey of other, more advanced, methods for dimension reduction on Riemannian manifolds. For another, short review, see Section 5 of [S. F. Huckemann and Eltzner 2021].

The most direct extension of PGA is *Geodesic principal component analysis* (G-PCA) [S. Huckemann, Hotz, and Munk 2010], which also approximates the observations by geodesics, but which furthermore optimizes for the base point μ. The method of *Principal nested spheres* (PNS) [Jung et al. 2010] works for data belonging to spheres in any dimension (see also *principal nested symmetric spaces* [Curry, Marsland, and McLachlan 2019] for a generalization). The observations are approximated by subspheres of the original sphere, minimizing a sum-of-squared-errors (SSE) criterion where the error is measured by geodesic distance. By iteratively fitting a subsphere of one dimension less, the output is a nested sequence of subspheres (submanifolds). In comparison, the result of PGA is a nested sequence of submanifolds that are constructed by *adding* a dimension in each iteration. Geodesics on the sphere are great arcs, which can be identified

with circles of radius 1, while the arbitrary subspheres of PNS can have any radius. Thus, the class of approximating submanifolds of the sphere considered by PNS is larger than that of PGA and G-PCA.

Barycentric subspace analysis (BSA) [Pennec 2018] is a method which works for data on an arbitrary Riemannian manifold. This method approximates the data by the generalization to Riemannian manifolds of barycentric subspaces. Like PNS, BSA does not depend on a base point. It depends on a chosen set of points, e.g. observations, generating the barycentric subspace. By adding or removing points, the result is a nested sequence of barycentric subspaces of the original manifold.

All the methods mentioned so far consider a certain parameterized class of curves or submanifolds, and choose among these by minimizing a sum-of-squared-errors objective function, based on geodesic distance (although for BSA, the SSE approach is just one option). This corresponds to our earlier general formulation of the dimension reduction problem (2.1.4) based on a family of subsets $\mathcal{Q}$ of the original space. We now mention some methods which are not based on a parameterized family of subsets or on minimizing SSE. A method which is completely non-parametric, in the sense of the family of approximating curves being infinite dimensional, is the method of *principal flows* [Panaretos, Pham, and Yao 2014]. The method fits local tangent PCA's and generates a curve whose tangent vectors are aligned with the first eigenvector - i.e. it moves in the direction that locally describes the most variation in the data. This method only works for approximating the observations by a curve, but was generalized to higher dimensional approximations in [Yao, Eltzner, and Pham 2023]. Our method of *principal subbundles*, described in Chapter **??**, can also be considered a generalization of the principal flow to higher dimensions (in Section **??** we elaborate on the relationship between these methods). An approach which is not based on deterministic curves, let alone geodesics, is *infinitesmal probabilistic principal component analysis* ([Sommer 2019] and [Pennec, Sommer, and T. Fletcher 2019], Chapter 10), which is based on anisotropic Brownian motions on the manifold. This method uses maximum likelihood to fit a covariance matrix of lower rank than the ambient manifold, say k, describing the motion of the Brownian motions. This enables, among other things, a representation of the observations in a k-dimensional Euclidean space.

Chapter 3

Conclusion and future work

Contents

3.1 Summary of contributions

In Chapter 2 and its associated appendix A, we gave background on geometric statistics on Riemannian manifolds, with a special focus on computational aspects and on the two foundational methods for dimension reduction; Tangent PCA and Principal geodesic analysis. We gave details, not contained in the literature, on how to do statistics in the tangent space in a way that does not depend on the chosen coordinates. A special attention was given to the empirical covariance tensor, central to tangent PCA. Furthermore, we discussed and rewrote the objective function of PGA from a computational point of view, and derived a Taylor expansion of it.

In Chapter **??**, we formulated a version of Tangent PCA adapted to applications in evolutionary biology; *Tangent phylogenetic PCA*. This method assumes that the data are leaf-nodes of a phylogenetic tree, and takes the implied correlation structure of such data into account. We defined, and derived estimators for, the associated notions of phylogenetic mean and covariance. This work is a generalization to Riemannian manifolds of a method, *phylogenetic PCA*, for Euclidean data. In studies of morphology in evolutionary biology the method is being applied to Procrustes aligned point configurations representing landmark shapes. In our version, we give a Riemannian gradient descent algorithm for estimating the intrinsic phylogenetic mean, and show that the way the mean is currently estimated in morphometrics corresponds to taking only a single step of this algorithm, when the Riemannian manifold is Kendall's shape space.

In Chapter **??**, we developed a method for dimension reduction on a Riemannian manifold M which approximates the data based on a very flexible class of curves, and submanifolds composed of such curves. The method works by constructing a subbundle of the tangent bundle on the manifold M via local PCA's. We call this subbundle the *principal subbundle*. This subbundle induces a *sub-Riemannian* (SR) structure on M, and we show that sub-Riemannian geodesics with respect to this structure stay close to the set of observations. We show that it is possible to approximate the data set by submanifolds consisting of such SR geodesics, even if the subbundle is non-integrable. In particular, we show that the image under the SR exponential map of the dual space to the subbundle at a base point μ is an embedding into M, yielding a submanifold of M whose dimension equals the rank of the subbundle. Non-integrability is likely to occur when the subbundle is estimated from noisy data, and our method demonstrates that sub-Riemannian geometry is a natural framework for dealing which such noise. We also contributed to the software library *geomstats* [Miolane

et al. 2020], with a class representing a sub-Riemannian metric and associated computations of geodesic etc. This makes computations in sub-Riemannian geometry available for future applications in geometric, and Euclidean, statistics.

3.2 Future work related to phylogenetic PCA

Using anisotopic normal distributions Sommer et al. defines an *anisotropic normal distribution* ([Sommer and Svane 2015], [Pennec, Sommer, and T. Fletcher 2019]) on a Manifold M, equipped with a connection or a Riemannian metric, as the time-t transition distribution of an anisotropic Brownian motion on M. This distribution has mean μ, called the *diffusion mean*, and covariance Σ if the initial point of the Brownian motion is μ and its covariance is Σ (more precisely Σ is a frame, so (μ, Σ) is a point in the frame bundle on M). The mean and covariance can be estimated by maximum likelihood methods (see [Sommer 2015] and [Grong and Sommer 2022]). The diffusion mean, defined through a stochastic process, is an alternative to the Fréchet mean, defined via the geodesic distance. The framework of anisotropic normal distributions (ANS) seems natural to use in relation to phylogenetic comparative methods (PCM) since the underlying model for the latter is exactly based on such Brownian motions. To adapt the ANS framework to the PCM setting, however, requires to modify the framework to deal with Brownian motions structured according to a tree-graph. This is ongoing work. An off-the-shelf solution is to exchange the weighted Fréchet mean with the weighted diffusion mean (see [Sommer and Bronstein 2020] and [Jensen and Sommer 2022]).

Geometric morphometrics in relation to recent developments in computational geometric statistics We see potential in revisiting the geometric morphometrics (GM) literature in view of the recent developments in computational differential geometry and geometric statistics. The basis of the methods in GM was developed before software libraries such as geomstats and jaxgeometry made available intrinsic computations on manifolds, in particular for geometric statistics. As discussed in Chapter **??**, the field of geometric morphometrics builds on a wide range of approximations of intrinsic computations, e.g. different ways to compute local linear approximations to Kendall's shape space. It is possible that, in light of the mentioned recent computational developments, some of these are superfluous, or that some of them can be exchanged for more suitable, intrinsic methods. We hope that our work on phylogenetic PCA is one step in this direction.

A broader picture In continuation of the previous subsection on GM, we state the following long term goal: to formulate on Riemannian manifolds those phylogenetic comparative methods that are relevant for morphometrics and where intrinsic differential geometry can make a contribution. Such methods could be other basic statistical operations than PCA, e.g. *phylogenetic generalized least squares regression* [Symonds and Blomberg 2014]. Or it could be in the direction of allowing more flexible probabilistic models, e.g. based on processes more general than Brownian motions, and where the parameters are not necessarily the same throughout the tree (see [Harmon 2019] and [Mitov, Bartoszek, and Stadler 2019]). Or it could be in the direction of uncertainty quantification, e.g. how does uncertainty in tree estimation propagate to uncertainty in the estimated parameters.

Choosing landmark positions that are comparable (homologuous) between species can be difficult, and the annotation of such datasets involves costly manual labour. Therefore it seems desirable to also develop phylogenetic comparative methods for 'full shapes', i.e. shapes defined by continuous curves and surfaces instead of landmarks, e.g. using the LDDMM framework for continuous curves or surfaces (see [Younes 2010]) or other shape representations (see [Salili-James et al. 2022] for a comparison of different representations in the context of shape classification). See [Mitteroecker and Schaefer 2022] for a discussion of potential pros and cons of 'landmark-free' shape representations in the context of biology and geometric morphometrics. However, geometric statistics is less developed for such infinite-dimensional manifolds compared to the finite dimensional setting treated in this thesis.

3.3 Future work related to principal subbundles

Evaluation We illustrated the framework of principal subbundles on a number of experiments in Chapter 4, but it was not fully benchmarked with respect to competing methods for manifold reconstruction, dimension reduction and metric learning. This is our most immediate direction of future work.

Improving computations As mentioned in Appendix B.2.1, we have so far experimented with two ways of integrating the Hamiltonian equations; standard Euler integration and symplectic Euler. We plan to test other, e.g. higher-order, symplectic integrators, with the aim of decreasing computational cost (by allowing for increased step size) while keeping the error tolerable. One candidate could be the second-order leapfrog scheme (see e.g. [Hairer 2001]), which is

the most commonly used integrator in the field of Hamiltonian Monte-Carlo according to [Betancourt 2017].

A second strategy for speeding up computations is to limit the number of points in which the local PCA is computed, in the following way. We suggest to precompute local PCA's only at observations, and interpolate between these by local averages. This strategy is used in surface reconstruction methods based on normals (in [Kazhdan, Bolitho, and Hoppe 2006] Poisson surface reconstruction is computed in this way). In that setting, the interpolation is simply a local average of normal vectors. In the higher dimensional case we need to compute an average of k-dimensional subspaces (see [Marrinan et al. 2014] for an overview of different such averages).

A third strategy for speeding up computations is to lower the number of observations that are considered in each local PCA when integrating the Hamiltonian equations. Let p be a point on the ambient manifold $\mathcal{N}$ and choose a radius $r > 0$. For a moment, we will consider only observations that are inside the geodesic ball $B_r(p) \subset \mathcal{N}$. Choose a weight threshold $\epsilon > 0$ such that the distance corresponding to this weight, $r_0 = K_\alpha^{-1}(\epsilon)$, is below r, where K_α is the kernel function in of the weighted second moment, Definition **??**. Then it holds that $B_{r_0}(q) \subseteq B_r(p)$ for any q such that $d(p, q) \leq r - r_0$. This holds in particular for any point along a geodesic $q_t = \exp_p(tv)$ up to time $t \leq r - r_0$ if $\|v\| = 1$. This means that, relative to the threshold ϵ, the observations contained in $B_r(p)$ are sufficient for computing the principal subbundle (local PCA) at points along this geodesic until time $t' = r - r_0$. At $q_{t'}$ a new ball $B_r(q_{t'})$ can be computed, and the process can be repeated. We think of this process as changing between 'charts' $B_r(\cdot)$ containing a relevant subset of observations for a given time interval. We have presented the strategy for computing a single geodesic, but we expect a similar procedure to be useful when computing neighbouring geodesics generating a principal submanifold (Algorithm **??**).

Extensions In this section, we mention two alternative ways to use a principal subbundle. The common theme is to consider other classes of curves than sub-Riemannian geodesics. This is the subject of ongoing work and is therefore only described cursorily.

One extension is to consider horizontal stochastic processes instead of sub-Riemannian geodesics. E.g. a sub-Riemannian equivalent of Brownian motions. The time-t transition distribution of such processes can act as an analogue to anisotropic normal distributions; they are probability distributions concentrated around the data. To perform maximum likelihood estimation for

such distributions we need the sub-Riemannian equivalent of bridge-processes (a stochastic process conditioned on hitting a specified point), which is work-in-progress. Such bridge processes are also expected to be of independent interest, e.g. for generarative models, as a process that moves between two specified observations while following the point cloud.

Another extension aims at generating curves that are attracted to the point cloud in a different way than the sub-Riemannian geodesics described in Chapter **??**. Sub-Riemannian geodesics determined by the principal subbundle are constrained to move horizontally w.r.t. the subbundle. I.e. they are subjected to constraints on their velocity vectors (as well as the constraint of being locally length-minimizing). When these curves are initialized within the point cloud they remain within it - at least in principle, and up to numerical error. Apart from this, they are not attracted to the data. It might be useful to introduce, in some way, an 'attraction term' that forces a curve initialized outside of the data to move towards it. This could also act as a correction to the deviation from the point cloud that is sometimes incurred by the numerical integration of the geodesic equations. We are currently considering multiple ways of adding such a data attraction term. One strategy is to add a term to the metric matrix which goes to infinity when evaluated outside of the point cloud, similar to what is done in [Hauberg 2018]. This changes the geometry of the space, in the sense of changing the metric. Another approach is to consider more general Hamiltonians than the sub-Riemannian Hamiltonian (Section **??**)). This does not change the geometry, but only the solutions to the Hamiltonian equations, which are then no longer geodesics.

Appendices

Appendix A

Appendix for Chapter 2

A.1 Computational aspects of geometric statistics

For doing statistics on a Riemannian manifold we need to be able to compute operations such as those described in Section 2.3. For a given manifold, closed form expressions may or may not be available - most often numerical approximations are needed at some stage of a computation. Fortunately, a large effort has been put into implementing numerical differential geometry and geometric statistics in recent years, with software libraries such as *geomstats* [Miolane et al. 2020] and *jaxgeometry* (http://bitbucket.org/stefansommer/jaxgeometry).

Two main computational bottlenecks for geometric statistics is computation of the exponential and logarithmic maps. The former is computed by numerical integration of the geodesic equation in a chart, either as a 2nd order ODE in the tangent bundle or as two coupled first order ODE's in the cotangent bundle (i.e. the Hamiltonian geodesic equations, see **??**). The logarithm can be computed by a shooting method, i.e. by solving an optimization problem in the tangent space that searches for the initial tangent vector minimizing some distance between the endpoint of the candidate geodesic and the desired endpoint. This distance should be efficient to compute, e.g. an ambient Euclidean distance if we have available an embedding of the manifold in a Euclidean space. In more detail, the optimization problem is the following,

$$\log_p(y) \in \operatorname*{argmin}_{v \in T_pM} d_0(\exp_p(v), y) + \lambda \|v\|_g. \tag{A.1.1}$$

The term $\lambda\|v\|_g$ facilitates finding a *length-minimizing* geodesic, with the hyperparameter $\lambda > 0$ determining how much emphasis is placed on this length-minimizing property compared to the 'matching term' $d(\exp_p(v), y)$. In Geomstats and in Jax geometry, this optimization problem is solved using the BFGS algorithm [Wright, Nocedal, et al. 1999] with automatic differentiation.

Thus, the objective function in the optimization problem for the log map involves computing the exponential map, i.e. solving the geodesic equations. This is computed at least once at each step of the optimization algorithm. This has led us to develop a discrete approximation of the log map, presented in the following section. We use this approximation in Chapter **??**, e.g. when

computing projections to a principal submanifold in Section **??**, and we expect it to be of wider interest for computations on Riemann manifolds.

Lastly, we note that when the log is computed by solving (A.1.1), the basis for the tangent space is arbitrary, in particular it is not necessarily orthonormal as one might expect if considering the computed log to be a point in a normal chart.

A.1.1 Discretization of the log map

We propose a discrete approximation of the Riemannian logarithmic map. The idea is to pick a finite set of relatively evenly distributed tangent vectors on the unit sphere in T_pM. For each such tangent vector we integrate the geodesic equations up to a chosen time point $r > 0$. For each point along a geodesic initialized by one of these tangent vectors, we store its time index and its position in M. The collection of all such positions forms a discretized geodesic ball of radius r in M. For a given point $x \in M$, $\log_p(x)$ is then approximated by the tangent vector corresponding to the nearest point on the generated grid. 'Nearest' is here defined w.r.t. a distance metric d_0 that should be efficient to compute, e.g. an ambient Euclidean metric, as for the objective function (A.1.1). The procedure is described in 3 steps below.

1. Generate an even grid $\overline{\mathbb{S}^d} \subset \mathbb{S}^d \subset T_pM$ on the unit sphere (w.r.t. the Riemannian meric g) in the tangent space at p, and a grid of timepoints $\overline{[0,r]} \subset [0,r]$, for some final time $r > 0$.

2. Generate a discrete ball $\overline{B_p}$ of radius $r > 0$ around p on M by

$$\overline{B_p} \coloneqq \{\exp_p(tv) \mid t \in \overline{[0,r]}, v \in \overline{\mathbb{S}^d}\}.$$

   For each point $y_{t_0}^{v_0} \coloneqq \exp_p(t_0 v_0) \in \overline{B_p}$, store the parameters of the corresponding initial tangent vector, i.e. t_0, v_0.

3. Define the discrete log, $\overline{log}_p(y)$, by

$$\overline{\log}_p(y) \in \operatorname*{argmin}_{y_t^v \in \overline{B_p}} d_0(y_t^v - y) + \lambda|t| \tag{A.1.2}$$

where λ is a trade-off parameter, as described in relation to Equation (A.1.1). The optimization problem (A.1.2) is discrete. It can be vectorized, such that the only computationally heavy part is the initialization steps 1 and 2. We expect the discrete log to be the most useful in situations where 'many' logs, $\log_p(\cdot)$,

need to be computed for the same base point $p \in M$, such as in the gradient descent for a Fréchet mean (Algorithm **??**, Chapter **??**).

A.2 Statistics in the tangent space

A general strategy for performing statistics on data belonging to a Riemannan manifold, $x_1, \ldots, x_N \in M$, is to map the data to a tangent space $T_\mu M$ at a well chosen base point $\mu \in M$, and apply Euclidean statistical methods in this vector space. To minimize distortion of distances due to tangent space linearization (see Section 2.3.1), it is preferable that the base point is relatively close to the observations. A Fréchet mean is a good candidate, since it is a minimizer of the sum of squared distances to the observations - note however that due to curvature it can be well outside the support of the data (this is the case for data distributed along a great arc on the sphere, for example). Having chosen μ, the next step is to map the observations to $T_\mu M$ via the Riemannian logarithm, i.e. the data is represented as

$$\log_\mu(x_1), \ldots, \log_\mu(x_N) \in T_\mu M.$$

The Euclidean statistical method can then be applied to the transformed dataset. For some methods, it is a reasonable final step to map the result back to M via the Riemannian exponential. In this way, e.g. a regression line in $T_\mu M$ is mapped to a curve on M, or a subspace of $T_\mu M$ found by e.g. PCA is mapped to a submanifold of M.

When following this strategy, one needs to take into account the fact that there is no canonical basis for a tangent space, and that Euclidean formulas usually assume the data to be represented in an orthonormal coordinate system. In various presentations of tangent space statistics (e.g. the tangent PCA approximation of PGA in [T. P. Fletcher et al. 2004]) is it implicitly assumed that the observations are represented in a basis induced by a normal chart. This basis is orthonormal - the local representation of the Riemannian metric is the identity matrix. However, if the chart does not induce an orthonormal basis on the tangent space, one needs to change to one, or take the non-orthonormal basis into account. As described in Appendix A.1, when the log is computed numerically by the optimization problem (A.1.1), the basis chosen for the tangent space is arbitrary, not necessarily orthonormal.

To find an orthonormal basis for $T_\mu M$, one can do the following. Let $p \in M$ and let g be the Riemannian metric on M, represented in local coordinates

around p. Let

$$LL^T = g^{-1}(\mu)$$

be the cholesky decomposition of the cometric. Then the columns of L form an orthonormal basis for T_pM.

In the next section, we focus on the particular case of tangent PCA, i.e. Euclidean PCA performed in a tangent space.

A.2.1 Expressing the second moment in coordinates

The geometric statistics literature is unclear regarding the meaning of the covariance matrix defined as the tensor product (2.4.1), and what its coordinate expression is. Therefore, in this section, we show how to derive the coordinate expression (2.4.2) from (2.4.1).

For some $v, u \in T_p\mathcal{N}$, the expression $v \otimes u$ can be identified with an endomorphism on $T_p\mathcal{N}$. Its coordinate representation is thus a $d \times d$ matrix. There seems to be some confusion about this in the geometric statistics literature, so we give details below. For the remainder of this section we denote a Riemannian metric by h, to align the notation with Chapter **??**, where multiple metrics are at play.

Lemma A.2.1. *Let $(\mathcal{N}, h)$ be a Riemannian manifold, and $u, v \in T_p\mathcal{N}$. Given a choice of basis for $T_p\mathcal{N}$, the tensor $v \otimes u \in T_p\mathcal{N} \otimes T_p\mathcal{N}$ can be expressed in coordinates as*

$$vu^T h_p \in \mathbb{R}^{d\times d}, \tag{A.2.1}$$

where $u, v \in \mathbb{R}^{d\times 1}$ are the vectors and $h_p \in \mathbb{R}^{d\times d}$ is the Riemannian metric represented w.r.t. the chosen basis.

Proof. The tensor $v \otimes u$ is an element of the tensor product space $T_p\mathcal{N} \otimes T_p\mathcal{N}$. After choosing a Riemannian metric, there is a canonical isomorphism between $T_p\mathcal{N}$ and its dual space, $T_p^\star\mathcal{N}$, given by the Riemannian flat map,

$$\flat : T_p\mathcal{N} \to T_p^\star\mathcal{N} : u \mapsto h_p(u, \cdot) := u^\flat.$$

Thus

$$T_p\mathcal{N} \otimes T_p\mathcal{N} \cong T_p\mathcal{N} \otimes T_p^\star\mathcal{N},$$

where elements of the latter space are denoted $(1, 1)$ tensors. Furthermore, there is a canonical isomorphism, independent of a Riemannian metric,

$$T_p\mathcal{N} \otimes T_p^\star\mathcal{N} \cong \mathrm{End}(T_p\mathcal{N}),$$

where $\mathrm{End}(T_p\mathcal{N})$ is the space of endomorphisms on $T_p\mathcal{N}$. This isomorphism is given by the map Φ which takes an endomorphism A to the $(1,1)$ tensor $\Phi(A)$ that acts on $w \in T_p\mathcal{N}$ and $\eta \in T_p^\star\mathcal{N}$ by $\Phi(A)(w,\eta) = \eta(Aw)$. The linear map corresponding to a (1,1) tensor of the form $v \otimes u^\star, v \in T_p\mathcal{N}, u^\star \in T_p^\star\mathcal{N}$, is $w \mapsto \Phi^{-1}(v \otimes u^\star)(w) = v \cdot u^\star(w)$, i.e. a scaling of v by $u^\star(w) \in \mathbb{R}$.

After choosing a basis for $T_p\mathcal{N}$, the tangent vectors v, w can be represented as column vectors $v, w \in \mathbb{R}^{d\times 1}$. The flat map can be represented by the matrix h_p, which is the matrix representation of the Riemannian metric at p. After identifying covectors with row vectors (i.e. coordinate representations of linear maps from $T_p\mathcal{N}$ to $\mathbb{R}$), $u^\flat$ can be represented as the row vector $u^\flat = (h_p u)^T \in \mathbb{R}^{1\times d}$. This acts on w by $u^\flat(w) = (h_p\, u)^T\, w$. Thus, w.r.t. some chosen basis, the matrix representation of our desired endomorphism is given by

$$\Phi^{-1}(v \otimes u^\flat) = v(h_p u)^T = v u^T h_p.$$

□

A.2.1.1 Verifying independence of the coordinate system

Let Q be the change-of-basis matrix from basis a of $T_p\mathcal{N}$ to basis b. Then $Q^\star = (Q^T)^{-1}$ is the corresponding change-of-basis matrix from basis $a^\star$ to $b^\star$ for $T_p^\star\mathcal{N}$, where these bases are dual to a, b. Thus, the change of basis of tangent vector v from a to b is computed as $v_b = Q_{ab} v_a$. The flat map $\flat$ is a linear map from $T_p\mathcal{N}$ to $T_p^\star\mathcal{N}$, so if $(h_p)_a$ is its representation w.r.t. bases a and $a^\star$, then its representation w.r.t. bases b and $b^\star$ is computed as

$$(h_p)_b = Q^\star (h_p)_a Q^{-1} = (Q^T)^{-1}(h_p)_a Q^{-1}.$$

We verify that the change of basis of the individual elements u, v, h_p matches the change of basis of the matrix (as a linear map) (A.2.1):

$$v_b u_b^T (h_p)_b = Q v_a (Q u_a)^T (Q^T)^{-1} (h_p)_a Q^{-1} \tag{A.2.2}$$
$$= Q v_a u_a^T (h_p)_a Q^{-1}. \tag{A.2.3}$$

As opposed to this, the expression $v_b u_b^T$ does *not* transform properly under basis change: $v_b u_b^T = Q v_a (Q u_a)^T = Q v_a u_a^T Q^T$ is only equal to $Q v_a u_a^T Q^{-1}$ if $Q^T = Q^{-1}$, i.e. if the basis change matrix is orthogonal, meaning that it only rotates the basis.

A.3 Computing principal geodesic analysis

In this section, we elaborate on the optimization problem (A.3.1) that defines principal geodesic analysis (PGA). We will see that, if no closed form expressions for the Riemannian log and projection w.r.t. to geodesic distance are available, then it is a triply nested optimization problem. As of this date, to the best of our knowledge, there exists no publicly available implementation of PGA. The first principal direction solves

$$v_1 \in \operatorname*{argmin}_{v \in T_\mu M} \frac{1}{N} \sum_{j=1}^{N} \mathrm{d}(x_j, \pi_{\exp_\mu(span(v) \cap \mathcal{C}_\mu)}(x_j))^2 \tag{A.3.1}$$

$$= \operatorname*{argmin}_{v \in T_\mu M} \frac{1}{N} \sum_{j=1}^{N} \| \log_{x_j}(\pi_{\exp_\mu(span(v) \cap \mathcal{C}_\mu)}(x_j)) \|_g^2, \tag{A.3.2}$$

The projection problem π can be rephrased as

$$\pi_{\exp_\mu(span(v))}(q) \in \operatorname*{argmin}_{p \in \exp_\mu(span(v))} d(p, q) \tag{A.3.3}$$

$$= \operatorname*{argmin}_{\alpha \in \mathbb{R}} \quad d\left(\exp_\mu(\alpha \cdot v), q\right). \tag{A.3.4}$$

The log is given by the optimization problem (A.1.1). The outer optimization for $v \in T_\mu M$ thus makes this a triply nested optimization problem.

Automatic differentiation [Margossian 2019] provides an efficient way to compute gradients for solving optimization problems, and packages like geomstats and jaxgeometry relies heavily on it. At the time of writing, automatic differentiation of such 'implicitly' defined functions, where the output is the solution to an optimization problem, is still at an early stage of development. An alternative approach is to rewrite the problem into a large global optimization instead of a nested one.

We first move out the inner optimization for the projection,

$$\operatorname*{argmin}_{v \in T_\mu M} \frac{1}{N} \sum_{j=1}^{N} \| \log_{x_j}(\pi_{\exp_\mu(span(v))}(x_j)) \|_g^2 \tag{A.3.5}$$

$$= \operatorname*{argmin}_{\substack{\alpha \in \mathbb{R}^N \\ v \in T_\mu M}} \frac{1}{N} \sum_{j=1}^{N} \| \log_{x_j}(\exp_\mu(\alpha_j v)) \|_g^2 \tag{A.3.6}$$

We then move out the inner optimization for the log,

$$v_1 \in \operatorname*{argmin}_{\substack{\alpha \in \mathbb{R}^N \\ v \in T_\mu M}} \frac{1}{N} \sum_{j=1}^{N} \| \log_{x_j}(\exp_\mu(\alpha_j v)) \|_g^2 \tag{A.3.7}$$

$$\approx \underset{\substack{\alpha\in\mathbb{R}^N \\ v\in T_\mu M \\ W\in\mathbb{R}^{d\times N}}}{\operatorname{argmin}} \frac{1}{N}\sum_{j=1}^{N} \|W_{:j}\|_g^2 + \lambda d_0(\exp_{\exp_\mu(\alpha_j v)}(W_{:j}), x_j)^2, \tag{A.3.8}$$

where W is a matrix whose columns are the log-vectors, one for each observation. λ is the same trade-off parameter as in the log-optimization problem (A.1.1), controlling the emphasis on minimizing approximation error in the *log* relative to minimizing the length of the corresponding geodesic. d_0 is a distance that can be computed easily - henceforth we will assume that $d_0(x, y) = \|x - y\|_{Eucl}$ is the Euclidean distance and that the manifold is embedded in a Euclidean space.

The interpretation of the optimization problem (A.3.8) can be problematic, in the following sense. An optimization scheme will produce a sequence of parameter candidates in the parameter space $\mathbb{R}^{(d+1)N+d}$, but for each update of parameters $(\alpha_j)_0 \in \mathbb{R}$, $j = 1 \dots N$, and $v_0 \in T_\mu M$ the vector $W_{:j} \in \mathbb{R}^d$ is used as a tangent vector belonging to a different tangent space, $T_{\exp_\mu((\alpha_j)_0 v_0)}M$. On a general Riemannian manifold (that is not e.g. a Lie group), there is no canonical way to identify tangent spaces. One way to map between different tangent spaces is the parallel transport map, $\Pi_x^y : T_x M \to T_y M$. This map is an isometry, meaning that it preserves angles and lengths of vectors. Parallel transport depends on a chosen curve between points x, y, which we will assume to be the length-minimizing geodesic (assumed unique) between x, y. Therefore, we can use parallel transport to ensure that the vectors $W_{:j}$, $j = 1, \dots, N$, represent tangent vectors in the same tangent space, namely $T_\mu M$. Let $\widetilde{\alpha_j v} := \exp_\mu(\alpha_j v) \in M$. Then this amounts to the following version of the optimization problem,

$$\underset{\substack{\alpha\in\mathbb{R}^N \\ v\in T_\mu M \\ W\in\mathbb{R}^{d\times N}}}{\operatorname{argmin}} \frac{1}{N}\sum_{j=1}^{N} \|W_{:j}\|_g^2 + \lambda \|\exp_{\widetilde{\alpha_j v}}(\Pi_\mu^{\widetilde{\alpha_j v}} W_{:j}) - x_j\|_{Eucl}^2, \tag{A.3.9}$$

The parameter space of the final optimization problem (A.3.9) for the first principal direction v_1 is of dimension $(d + 1)N + d$. For the i'th principal component, solving (2.4.6), the projection problem causes the projection parameter α to grow in dimension; $\alpha \in \mathbb{R}^{i \cdot N}$, since a point in an i dimensional subspace is parameterized by i coordinates. Thus the parameter space dimension for the i'th component is $(d+i)N+d$. Evaluating the objective function involves solving the parallel transport equation once and computing two exponential maps, i.e. integrating the geodesic equations twice. The authors have implemented this scheme with promising results for lower dimensional parameter spaces, but more work needs to be done in testing it for higher dimensions.

A.4 Taylor-approximation of PGA

In this section we derive a Taylor approximation of the PGA objective function. The expansion contains 3 terms, the first of which corresponds to the objective function of tangent PCA. The subsequent terms takes curvature into account. The expansion is derived by expressing the PGA objective function in terms of the socalled neighbouring log map, which is a map with a known Taylor expansion. The neighbouring log and its Taylor expansion was first introduced by Pennec in [Pennec 2019] building on work by Gavrilov ([Gavrilov 2006], [Gavrilov 2007]).

Definition A.4.1 (The neighbouring log map)**.** Let p be a point on M, and let v, w be tangent vectors in T_pM. Let $\Pi_x^y(v)$ be the parallel transport of v from T_xM to T_yM along the (assumed unique) length-minimizing geodesic from $x \in M$ to $y \in M$. Let $p_u \coloneqq \exp_p(u)$. Then the neighbouring log map is given by

$$l_p(u,v) = \Pi^p_{\exp_p(u)} \left(\log_{\exp_p(u)}(\exp_p(v)) \right) \tag{A.4.1}$$

$$= \Pi^p_{p_u} \left(\log_{p_u}(p_v) \right). \tag{A.4.2}$$

The Taylor expansion of the neighbouring log is given by

$$l_p(v,w) = w - v + \frac{1}{6}R(w,v)(v-2w) + \frac{1}{24}(\nabla_v R)(w,v)(2v-3w) + \frac{1}{24}(\nabla_w R)(w,v)(v-2w) + O(5),$$

where R is the curvature tensor and $\nabla.R$ its covariant derivative. See [Pennec 2019] for details.

We now formulate the PGA objective function, for the first principal component, in terms of the neighbouring log map. The objective function for the subsequent components can be found similarly. Below, we will use the notation $\overline{x_j} \coloneqq \log_\mu(x_j)$, and $\|\cdot\|_p$ for the norm on T_pM induced by that Riemannian metric, and $\langle \cdot, \cdot \rangle_p$ for the Riemannian metric at $p \in M$.

$$v_1 \in \underset{v \in T_\mu M}{\operatorname{argmin}} \frac{1}{N} \sum_{j=1}^{N} \mathrm{d}(x_j, \pi_{\exp_\mu(span(v))}(x_j))^2 \tag{A.4.3}$$

$$= \underset{\substack{\alpha \in \mathbb{R}^N \\ v \in T_\mu M}}{\operatorname{argmin}} \frac{1}{N} \sum_{j=1}^{N} \mathrm{d}(x_j, \exp_\mu(\alpha_j v))^2 \tag{A.4.4}$$

$$= \underset{\substack{\alpha\in\mathbb{R}^N \\ v\in T_\mu M}}{\operatorname{argmin}} \frac{1}{N}\sum_{j=1}^{N} \|\log_{x_j}(\exp_\mu(\alpha_j v))\|_{x_j}^2 \tag{A.4.5}$$

$$= \underset{\substack{\alpha\in\mathbb{R}^N \\ v\in T_\mu M}}{\operatorname{argmin}} \frac{1}{N}\sum_{j=1}^{N} \|\Pi_{x_j}^{\mu}\log_{x_j}(\exp_\mu(\alpha_j v))\|_{\mu}^2 \tag{A.4.6}$$

$$= \underset{\substack{\alpha\in\mathbb{R}^N \\ v\in T_\mu M}}{\operatorname{argmin}} \frac{1}{N}\sum_{j=1}^{N} \|l_\mu\left(\overline{x_j}, \alpha_j\, v\right)\|_{\mu}^2 \tag{A.4.7}$$

$$\approx \underset{\substack{\alpha\in\mathbb{R}^N \\ v\in T_\mu M}}{\operatorname{argmin}} \frac{1}{N}\sum_{j=1}^{N} \Bigg(\|\alpha_j v - \overline{x_j}\|_{\mu}^2 \tag{A.4.8}$$

$$+ \frac{1}{3}\langle R(\alpha_j v, \overline{x_j})(\alpha_j v), \overline{x_j}\rangle_\mu \tag{A.4.9}$$

$$+ \frac{1}{12}\langle (\nabla_{\overline{x_j}+\alpha_j v} R)(\alpha_j v, \overline{x_j})(\alpha_j v), \overline{x_j}\rangle_\mu \Bigg) \tag{A.4.10}$$

Equation (A.4.4) uses the fact that

$$\pi_{\exp_\mu(span(v))}(q) \in \underset{p\in\exp_\mu(span(v))}{\operatorname{argmin}} d(p,q) = \underset{\alpha\in\mathbb{R}}{\operatorname{argmin}} \quad d\left(\exp_\mu(\alpha\cdot v), q\right),$$

and Equation (A.4.6) uses the fact that parallel transport is an isometry. Thus, the expansion consists of 3 terms. Including only the first order term yields the objective function of tangent PCA. The higher order terms take curvature into account.

The complexity of the approximated optimization problem is comparable to 'tangent PCA', thus relatively inexpensive. In particular, the log-map needs to be solved only once for every observation. In contrast to this, the exact PGA problem is a nested optimization problem, as we discussed in Section A.3.

Appendix B

Appendix for Chapter ??

B.1 Proofs

B.1.1 Smoothness of the principal subbundle

We show smoothness first on $\mathbb{R}^d$ and then on a Riemannian manifold $(\mathcal{N}, h)$. The proof of the latter utilizes the former result in a chart, as well as smoothness results for the involved maps, which are only non-trivial in the manifold case.

*

Proof. Let $p \in \mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}$ be arbitrary. We will show that there exists a local frame of smooth vector fields spanning the subspace $\mathcal{E}_{p'}^{\alpha,k}$ at every point p' on an open set $\mathscr{U}$ around p. By Lemma 10.32 in [Lee 2013], this is equivalent to the subbundle being smooth on $\mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}$.

The eigenvalues of $\Sigma_\alpha(p)$ at p are

$$\lambda_1(p) \geq \cdots \geq \lambda_k(p) > \lambda_{k+1}(p) \geq \cdots \geq \lambda_d(p),$$

where only λ_k and λ_{k+1} are assumed to be different. Since $\Sigma_\alpha : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a smooth map, Theorem 3.1 of [Sun 1990] implies that there exists an open set $\mathscr{B}(p) \subset \mathbb{R}^d$ around p and d continuous functions $\bar{\lambda}_i(\cdot) : \mathscr{B}(p) \to \mathbb{R}$ satisfying that $\bar{\lambda}_i(p')$ is an eigenvalue of $\Sigma_\alpha(p')$ for all $p' \in \mathscr{B}$ and $\bar{\lambda}_i(p) = \lambda_i(p), i = 1 \dots d$.

Since each $\bar{\lambda}_i$ is continuous, there exists an open subset $\mathscr{U} \subset \mathscr{B}$ on which the ordering $\bar{\lambda}_1(p') \geq \cdots \geq \bar{\lambda}_d(p')$ holds for all $p' \in \mathscr{U}$, and where $\bar{\lambda}_i(p') = \bar{\lambda}_j(p')$ is only possible for i, j s.t. $\bar{\lambda}_i(p) = \bar{\lambda}_j(p)$. In particular $\bar{\lambda}_i(p') < \bar{\lambda}_{k+1}(p')$ for all $i < k+1$ and $p' \in \mathscr{U}$.

Theorem 3.2 of [Sun 1990] now says that there exists a frame of analytic vector fields $p \mapsto \{X_1(p), \dots, X_k(p)\}$ such that, for all $p' \in \mathscr{U}$,

$$\operatorname{span}\{X_1(p'), \dots, X_k(p')\} = V_{\bar{\lambda}_1(p'),\dots,\bar{\lambda}_k(p')}(\Sigma_\alpha(p'))$$

where $V_{\bar{\lambda}_1(p'),\dots,\bar{\lambda}_k(p')}(\Sigma_\alpha(p'))$ denotes the eigenspace of $\Sigma_\alpha(p')$ corresponding to eigenvalues $\bar{\lambda}_1(p'), \dots, \bar{\lambda}_k(p')$, which is exactly the principal subbundle subspace $\mathcal{E}_{p'}^{\alpha,k}$. □

To show that the principal subbundle on a Riemannian manifold is smooth, we need a result on smoothness of a certain map involving parallel transport.

Lemma B.1.1. *Let the map $f : \mathcal{N} \to \mathcal{N}$ and the vector field O on $\mathcal{N}$ be smooth. Let $\Pi_x^y : T_x\mathcal{N} \to T_y\mathcal{N}$ denote parallel transport along the (assumed unique) length-minimizing geodesic from x to y. Then the vector field*

$$p \mapsto \Pi_{f(p)}^{p} O(p) \in T_p\mathcal{N} \tag{B.1.1}$$

is smooth for every $p \notin Cut(f(p))$.

Proof. For $x, y \in \mathcal{N}$, the parallel transported vector $\Pi_x^y W \in T_y\mathcal{N}$ of $W \in T_x\mathcal{N}$ along a curve $\gamma : (0,1) \to \mathcal{N}$ is the value at time 1 of a vector field V along γ satisfying the linear initial value problem (an ODE)

$$\dot{V}^k(t) = -V^j(t)\dot{\gamma}^i(t)\Gamma_{ij}^k(\gamma(t)) \tag{B.1.2}$$
$$V(0) = W, \tag{B.1.3}$$

where Γ_{ij}^k, $i, j, k \in \{1, \ldots, d\}$ are the Christoffel symbols determined by the metric h. See [Lee 2018], Section 4, for details.

If γ is a geodesic with initial velocity $Q \in T_x\mathcal{N}$ then it is a solution to the geodesic equations (equations (B.1.5) and (B.1.6), below). In this case, we can write the parallel transport equation and the geodesic equations as a single, coupled, ODE:

$$\dot{V}^k(t) = -V^j(t)\dot{\gamma}^i(t)\Gamma_{ij}^k(\gamma(t)) \tag{B.1.4}$$
$$\dot{\gamma}^k(t) = U^k(t) \tag{B.1.5}$$
$$\dot{U}^k(t) = -U^i(t)U^j(t)\Gamma_{ij}^k(\gamma(t)) \tag{B.1.6}$$
$$U(0) = Q \tag{B.1.7}$$
$$V(0) = W \tag{B.1.8}$$
$$\gamma(0) = x. \tag{B.1.9}$$

Note that the equation for V is coupled with the equations for γ and U, but not vice versa, so that, in practice, the whole path γ can be computed first, and then subsequently V.

This is again a linear initial value problem, and the fundamental theorem for ODE's states that solutions exist, and depend smoothly on the initial conditions Q, W, x. This shows smoothness of the parallel transport operator in the case where $\gamma((0,1))$ is contained in a single chart. For the more general case, we refer to the technique used in the proof of Proposition 4.32 in [Lee 2018] for showing that solutions found on individual charts overlap smoothly.

The map (B.1.1) takes a point $p \in \mathcal{N}$ to a vector field at time 1 satisfying equations (B.1.4)-(B.1.9). For each p, the initial conditions are

$$\begin{aligned} x &= f(p) \\ Q &= log^h_{f(p)}(p) \\ W &= O(p) \end{aligned}$$

all of which depend smoothly on p, if $p \notin \mathrm{Cut}(f(p))$. Since the solution to the ODE depends smoothly on the initial conditions, and since the initial conditions depends smoothly on p, the vector field (B.1.1) is smooth. □

*

Proof. As in the Euclidean case, we want to prove the existence of a smooth frame around every point $p \in \mathcal{S}'_{\alpha,k}$ spanning the subbundle locally around p. We will make use of the corresponding result for $\mathcal{N} = \mathbb{R}^d$, in a chart. In order to do this, we need to make sure that all of the involved maps are smooth as a function of p.

The tangent mean map $m : \mathcal{N} \to \mathcal{N}$ and the tensor field $p \mapsto \Sigma_\alpha(p) \in T_p\mathcal{N} \otimes T_p\mathcal{N}$ is smooth if each logarithm $\log^h_p(x_i)$, $i = 1 \dots N$, is smooth as a function of the base point $p \in \mathcal{N}$. This is ensured by the cut locus conditions in $\mathcal{S}'_{\alpha,k}$.

Assuming smoothness of Σ_α, we now consider charts (U, φ) on $\mathcal{N}$ and (O, ϕ) on $T\mathcal{N} \otimes T\mathcal{N}$, $U \subset \mathbb{R}^d, \varphi : U \to \varphi(U) \subset \mathcal{N}$, respectively $O \subset \mathbb{R}^{d\times d}, \phi : O \to \varphi(O) \subset T\mathcal{N} \otimes T\mathcal{N}$ (identifying each $T_p\mathcal{N} \otimes T_p\mathcal{N}$ with the space of endomorphisms on $T_p\mathcal{N}$, cf. Section A.2.1), around a point $p \in \mathcal{N}$ and $\varphi(p) \in T\mathcal{N} \otimes T\mathcal{N}$. In this chart,

$$f := \phi^{-1} \circ \Sigma_{\alpha,k} \circ m \circ \varphi$$

is a smooth map from $\mathbb{R}^d$ to $\mathbb{R}^{d\times d}$. Eigendecomposition of the matrix $f(p')$, $p' \in U$, is independent of the basis and thus of the choice of charts. As shown in the proof of Proposition **??**, there exists a smooth frame $p' \mapsto \{X_1(p'), \dots, X_k(p')\}$, $X_i(p') \in \mathbb{R}^d$, defined on some open subset $\mathscr{U} \subset \mathbb{R}^d$ around $\varphi^{-1}(p)$ s.t.

$$\mathrm{span}\{X_1(p'), \dots, X_k(p')\} = V_k(f(p')), \quad \forall p' \in \mathscr{U},$$

where the right hand side is the eigenspace of $f(p')$ corresponding to the largest k eigenvalues. We have thus shown the existence of a smooth frame on $\varphi(U) \subset \mathcal{N}$ spanning the corresponding eigenspaces of $\Sigma_\alpha \circ m$ at every point of $\varphi(U)$.

The last thing we need to take account of is the parallel transport map. Since parallel transport is an isometry, it holds that

$$\operatorname{span}\{\Pi^y_{p'} X_1(p'), \dots, \Pi^y_{p'} X_k(p')\} = \operatorname{span}\{\Pi^y_{p'} F_1(p'), \dots, \Pi^y_{p'} F_k(p')\} \subset T_y\mathcal{N},$$

where $\{F\}_{i=1..k}$ is any other frame spanning the same subspace as $\{X\}_{i=1..k}$ at p'. Thus, the parallel transported frame X spans the same subspace as the parallel transported eigenvectors $\{e_i\}_{i=1...k}$ at p' (the X_i's are not necessarily eigenvectors, as explained in [Sun 1990]). By Lemma B.1.1, the map $p \mapsto \Pi^p_{m(p)} V(p)$ is smooth, for a smooth vector field V. We have thus shown that the principal subbundle at p is spanned by a smooth frame around p. □

B.2 Notes on implementation

At each step of the integration of a geodesic, eigenvectors needs to be computed at the current position p. This involves evaluating the kernel $K_\alpha(|x_i - p|)$ for all $i = 1..N$. For large datasets, we suggest to do this using libraries specialized at such kernel-operations, such as KEOPS, as well as automatically filtering out points far away from p whose weight will be close 0 anyway. We have not had the need to implement these optimizations in order to run the examples of Section **??**.

The integration of the L geodesics in the algorithm for the principal submanifold can be parallelized; the computation of each one is independent from the rest.

To speed up computations further, we suggest to compute the sub-Riemannian metric at $p \in \mathbb{R}^d$ as a weighted mean of the metric computed at a finite number of points around p, possibly at every observation. This is similar to the approach used to compute Poisson surface reconstruction [Kazhdan, Bolitho, and Hoppe 2006] (where the surface normal at $p \in \mathbb{R}^3$ is evaluated as the mean of normals computed nearby p) and the learned Riemannian metrics in [Hauberg, Freifeld, and Black 2012]. In this way, the derivatives of the metric, and therefore the Hamiltonian, can be computed closed form, removing the need for automatic differentiation.

B.2.1 Choice of integration scheme

The integration of Hamilton's equations can be done using a symplectic integration scheme which aims at keeping the Hamiltonian constant. A constant hamiltonian is equivalent to constant speed, cf. eq. (**??**). This is desired because the computation of curve length and distance via eq. (**??**) assumes constant

speed. We compared ordinary Euler integration to semi-implicit Euler (see e.g. [Hairer et al. 2006]), a first-order symplectic integrator, and found the Hamiltonian to be better preserved using ordinary Euler integration in our experiments.

B.3 Choosing the kernel range α and bundle rank k

Firstly, note that these parameters can be considered to be a modelling choice, expressing the scale at which we want to analyze the data - what scale of variation to take into account. However, one can aim to find the 'lowest level of variation that is not due to random noise'. Secondly, note that the 'optimal' value of one hyperparameter depends on the value of the other. Since the rank k takes a finite number of values $k \in \{1, \dots, d-1\}$, we suggest to start by estimating this. See [Bac et al. 2021] for a survey and benchmarking of different methods. Given an estimated k, we suggest to select a range parameter for which the separation between eigenvalues λ_k and λ_{k+1} is the most clear on average. The optimal kernel range depends on the level of noise and the rate of change of the affine subspace $\mathcal{E}_p$ as a function of p, which, in the case of the manifold hypothesis, is an expression of the curvature of the underlying manifold. A fast varying $\mathcal{E}$ calls for a smaller α, while high levels of noise as well as a lower number of observations calls for a larger α.

B.4 Algorithm for combining principal submanifolds for 2D surface reconstruction

In this section, we present an algorithm for combining principal submanifolds $\{M^k_{\mu_j}(r_j)\}_{j=1..l}$ based at different base points $\mu_j, j = 1 \dots l$. In this case, $k = 2$ and we'll write M_{μ_j} instead of $M^2_{\mu_j}$. Given a point $x \in \mathbb{R}^3$, the algorithm first projects x to a set of nearest principal submanifolds, and then represents x as a weighted average of these projections, weighted by the SR distance between a projection and its corresponding base point. The point x can e.g. be an observation, $x \in \{x_i\}_{i=1..N}$, or a point in a principal submanifold, $x \in M_{\mu_j}$. The algorithm can then be run for each point x in $\{x_i\}_{i=1..N}$ or in $M_{\mu_j}, j = 1..l$.

The point sets representing principal submanifolds $M_{\mu_j}(r_j), j = 1 \dots l$, are generated by Algorithm 1. For each point $p \in M_{\mu_j}(r_j)$, we assume that the corresponding initial cotangent $\eta(p) \in \mathcal{E}^\star_{\mu_j}$ has been stored.

Apart from the hyperparameters of the principal subbundle and submanifolds, the algorithm needs a 'threshold parameter' $\epsilon > 0$. x will not be projected to

principal submanifold M_{μ_j} if the distance between x and its projection $\hat{x}_j$ to M_{μ_j} is greater than ϵ. Thus, the size of ϵ should be comparable to an estimate of the noise-level in the point cloud.

The algorithm is the following.

1. *Project to each submanifold:* project x to each $M_{\mu_j}(r_j), j = 1..l$, wrt. Euclidean distance, i.e. find the closest point in $M_{\mu_j}(r_j)$ w.r.t. Euclidean distance. Denote this projection of x to $M_{\mu_j}(r_j)$ by $\hat{x}_j$. Denote the corresponding initial cotangent by $\eta(\hat{x}_j)$ and the distance by $d_j := d(\mu_j, \eta(\hat{x}_j)) = \|\eta(\hat{x}_j)\|$.

2. *Filter out projections:* let $B := \{j \in \{1, \dots, l\} \,\big|\, |x - \hat{x}_j| < \epsilon\}$ consist of indices of the basepoints satisfying that the projection of x to M_{μ_j} is sufficiently close to x.

3. *Rescale distances:* set $\tilde{d}_j := d_j \cdot 1/s_j(d_j)$, where s_j is a continuous, decaying bijection with domain and image given by $s_j : [0, r_j] \to [0, 1]$. We suggest to use the affine function satisfying these constraints.

4. *Compute weighted average:* the weighted representation of x is now computed as

$$\hat{x} = \frac{1}{\sum_{j \in B} w_j} \sum_{j \in B} w_j \hat{x}_j,$$

where (unnormalized) weights w_j are given by

$$w_j(x) = e^{-(\tilde{d}_j - \tilde{d}_{j^\star})^2/(2\sigma)}, j = 1 \dots |B|,$$

and $j^\star := \operatorname{argmin}_{j \in B} d_j$ is the index of the principal submanifold that is closest w.r.t. SR distance. The standard deviation σ in w_j controls how fast the weights should go to zero. A general-purpose choice is $\sigma = \max_{j \in \{1,..,l\}} r_j\}$.

B.5 Supplementary figures

B.5.1 Illustration of overlapping submanifolds

Figure B.1 is a supplement to figure **??**, zooming in on the region of overlap between the two principal submanifolds.

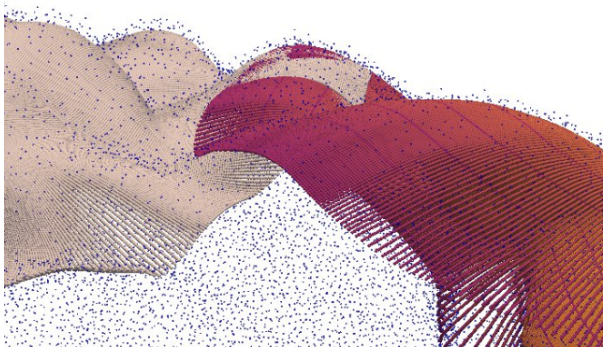


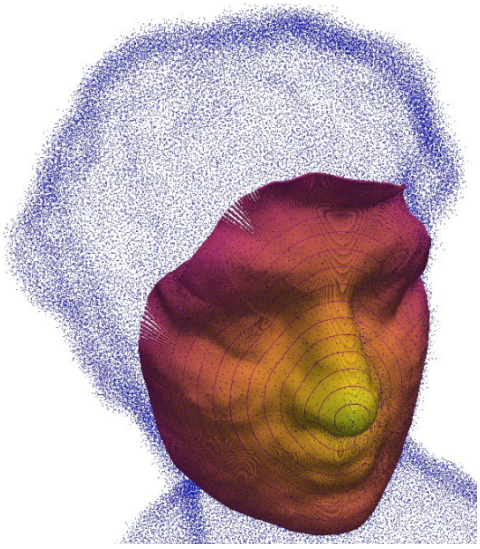
Figure B.1: Supplementary figure to Figure **??**, zooming in on the region where the two submanifolds overlap. The left, beige submanifold in this figure is the purple one in Figure **??**, recolored to be able to distinguish more clearly the two submanifolds.

B.5.2 Reconstruction of head sculpture surface under noise level 3 out of 3

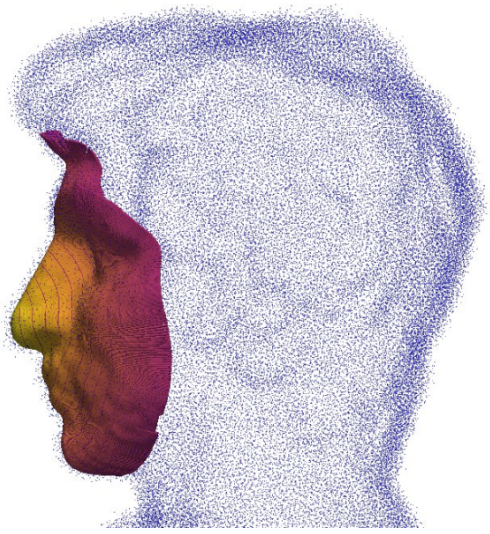
Figure B.2 illustrates the reconstruction of the face of the 'head sculpture' (from the benchmark dataset described in [Huang et al. 2022]), with noise level 3 out of 3. The parameters are the same as for the experiment described in section **??** except for a slightly larger kernel range.

B.5.3 Illustration of the log map on a 4-dimensional sphere in $\mathbb{R}^{50}$

Figure B.3 shows a single computed geodesic, found by solving the log problem $\log_p(q)$, for p, q and observations as described in section**??**. The distance $d^{\mathcal{E}}(p, q)$ is estimated as the length of the computed geodesic. The blue points are observations on the 4-dimensional sphere embedded in $\mathbb{R}^5 0$ projected to $\mathbb{R}^3$.



(a) Frontal view.



(b) Side view.

Figure B.2: Figures *(a)* and *(b)* show a principal submanifold recontructing the 'head sculpture' surface from a point cloud (purple points) with noise level 3 out of 3. The submanifold is centered around the tip of the nose. The figure shows the raw points generated by Algorithm **??** - no subsequent processing, apart from coloring, has been applied. The skewed circles on the face are geodesic balls, i.e. points on the same circle has the same SR distance to the center point. The colors of the face depends on the SR distance to the base point at the tip, a lighter color signifies shorter distance.

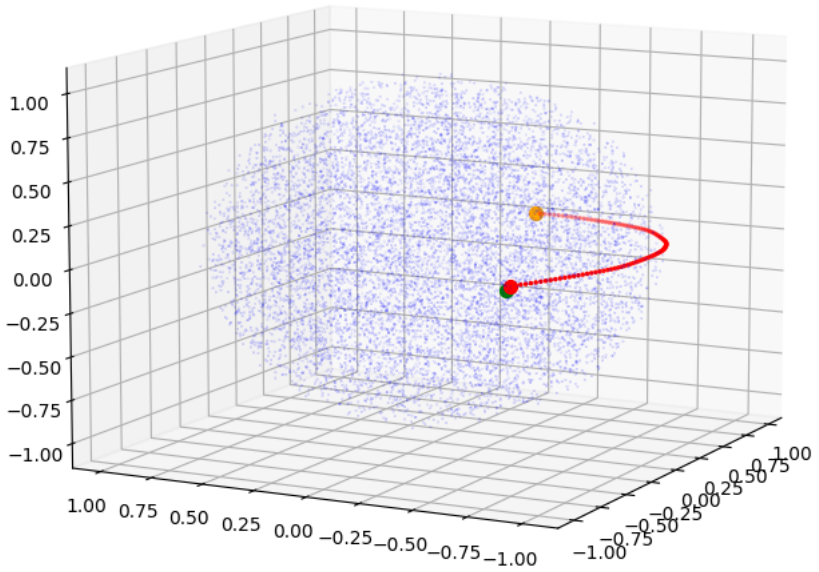


Figure B.3: Illustration of a single computed geodesic found by solving the log problem $\log_p(q)$, for p, q and observations as described in section **??**. The blue points are observations on the 4-dimensional sphere embedded in $\mathbb{R}^{50}$ projected to $\mathbb{R}^3$. The orange point is the initial point p, the red points are points on the geodesic, the green point is the target point q.

Bibliography

Absil, P.-A., Mahony, R., and Sepulchre, R. (2008). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Akhøj, M., Benn, J., Grong, E., Sommer, S., and Pennec, X. (2023). "Principal subbundles for dimension reduction". In: *arXiv preprint arXiv:2307.03128*.

Akhøj, M., Pennec, X., and Sommer, S. (2023). "Tangent phylogenetic PCA". In: *Scandinavian Conference on Image Analysis*. Springer, pp. 77–90.

Bac, J., Mirkes, E. M., Gorban, A. N., Tyukin, I., and Zinovyev, A. (2021). "Scikit-dimension: a python package for intrinsic dimension estimation". In: *Entropy* vol. 23, no. 10, p. 1368.

Banerjee, A., Dhillon, I. S., Ghosh, J., Sra, S., and Ridgeway, G. (2005). "Clustering on the Unit Hypersphere using von Mises-Fisher Distributions." In: *Journal of Machine Learning Research* vol. 6, no. 9.

Bendokat, T., Zimmermann, R., and Absil, P.-A. (2020). "A Grassmann manifold handbook: Basic geometry and computational aspects". In: *arXiv preprint arXiv:2011.13699*.

Betancourt, M. (2017). "A conceptual introduction to Hamiltonian Monte Carlo". In: *arXiv preprint arXiv:1701.02434*.

Boumal, N. (2023). *An introduction to optimization on smooth manifolds*. Cambridge University Press.

Chirikjian, G. S. and Kyatkin, A. B. (2021). *Engineering applications of noncommutative harmonic analysis: with emphasis on rotation and motion groups*. CRC Press.

Curry, C., Marsland, S., and McLachlan, R. I. (2019). "Principal symmetric space analysis". In: *Journal of Computational Dynamics* vol. 6, no. 2, pp. 251–276.

Do Carmo, M. P. and Flaherty Francis, J. (1992). *Riemannian geometry*. Vol. 6. Springer.

Fletcher, T. P., Lu, C., Pizer, S. M., and Joshi, S. (2004). "Principal geodesic analysis for the study of nonlinear statistics of shape". In: *IEEE transactions on medical imaging* vol. 23, no. 8, pp. 995–1005.

Fréchet, M. (1948). "Les éléments aléatoires de nature quelconque dans un espace distancié". In: *Annales de l'institut Henri Poincaré*. Vol. 10. 4, pp. 215–310.

Gavrilov, A. V. (2006). “Algebraic properties of covariant derivative and composition of exponential maps”. In: *Matematicheskie Trudy* vol. 9, no. 1, pp. 3–20.

— (2007). “The double exponential map and covariant derivation”. In: *Siberian Mathematical Journal* vol. 48, pp. 56–61.

Grong, E. and Sommer, S. (2022). “Most probable paths for anisotropic Brownian motions on manifolds”. In: *Foundations of Computational Mathematics*, pp. 1–33.

Guigui, N., Moceri, P., Sermesant, M., and Pennec, X. (2021). “Cardiac motion modeling with parallel transport and shape splines”. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1394–1397.

Hairer, E. (2001). “Geometric integration of ordinary differential equations on manifolds”. In: *BIT Numerical Mathematics* vol. 41, pp. 996–1007.

Hairer, E., Hochbruck, M., Iserles, A., and Lubich, C. (2006). “Geometric numerical integration”. In: *Oberwolfach Reports* vol. 3, no. 1, pp. 805–882.

Harmon, L. J. (2019). *Phylogenetic comparative methods*. Independent.

Hauberg, S. (2015). “Principal curves on Riemannian manifolds”. In: *IEEE transactions on pattern analysis and machine intelligence* vol. 38, no. 9, pp. 1915–1921.

— (2018). “Only bayes should learn a manifold (on the estimation of differential geometric structure from data)”. In: *arXiv preprint arXiv:1806.04994*.

Hauberg, S., Freifeld, O., and Black, M. (2012). “A geometric take on metric learning”. In: *Advances in Neural Information Processing Systems* vol. 25.

Huang, Z., Wen, Y., Wang, Z., Ren, J., and Jia, K. (2022). *Surface Reconstruction from Point Clouds: A Survey and a Benchmark*. eprint: arXiv:2205.02413.

Huckemann, S., Hotz, T., and Munk, A. (2010). “Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions”. In: *Statistica Sinica*, pp. 1–58.

Huckemann, S. F. and Eltzner, B. (2021). “Data analysis on nonstandard spaces”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* vol. 13, no. 3, e1526.

Jensen, M. H. and Sommer, S. (2022). “Mean Estimation on the Diagonal of Product Manifolds”. In: *Algorithms* vol. 15, no. 3, p. 92.

Jolliffe, I. T. and Cadima, J. (2016). “Principal component analysis: a review and recent developments”. In: *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* vol. 374, no. 2065, p. 20150202.

Jung, S., Liu, X., Marron, J., and Pizer, S. M. (2010). "Generalized PCA via the backward stepwise approach in image analysis". In: *Brain, Body and Machine*. Springer, pp. 111–123.

Kazhdan, M., Bolitho, M., and Hoppe, H. (2006). "Poisson surface reconstruction". In: *Proceedings of the fourth Eurographics symposium on Geometry processing*. Vol. 7.

Kendall, D. G. (1984). "Shape manifolds, procrustean metrics, and complex projective spaces". In: *Bulletin of the London mathematical society* vol. 16, no. 2, pp. 81–121.

Klingenberg, C. P. (2020). "Walking on Kendall's shape space: understanding shape spaces and their coordinate systems". In: *Evolutionary Biology* vol. 47, no. 4, pp. 334–352.

Lazar, N. A. (2007). "Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance imaging data: Commentary". In: *Journal of the American Statistical Association* vol. 102, no. 480, pp. 1105–1110.

Lee, J. M. (2013). *Introduction to smooth manifolds*. Springer.

— (2018). *Introduction to Riemannian manifolds*. Vol. 176. Springer.

Mardia, K. V., Jupp, P. E., and Mardia, K. (2000). *Directional statistics*. Vol. 2. Wiley Online Library.

Margossian, C. C. (2019). "A review of automatic differentiation and its efficient implementation". In: *Wiley interdisciplinary reviews: data mining and knowledge discovery* vol. 9, no. 4, e1305.

Marrinan, T., Ross Beveridge, J., Draper, B., Kirby, M., and Peterson, C. (2014). "Finding the subspace mean or median to fit your need". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1082–1089.

Miolane, N., Guigui, N., Brigant, A. L., Mathe, J., Hou, B., Thanwerdas, Y., Heyder, S., Peltre, O., Koep, N., Zaatiti, H., Hajri, H., Cabanes, Y., Gerald, T., Chauchat, P., Shewmake, C., Brooks, D., Kainz, B., Donnat, C., Holmes, S., and Pennec, X. (2020). "Geomstats: A Python Package for Riemannian Geometry in Machine Learning". In: *Journal of Machine Learning Research* vol. 21, no. 223, pp. 1–9.

Mitov, V., Bartoszek, K., and Stadler, T. (2019). "Automatic generation of evolutionary hypotheses using mixed Gaussian phylogenetic models". In: *Proceedings of the National Academy of Sciences* vol. 116, no. 34, pp. 16921–16926.

Mitteroecker, P. and Schaefer, K. (2022). "Thirty years of geometric morphometrics: Achievements, challenges, and the ongoing quest for biological meaning-

fulness". In: *American Journal of Biological Anthropology* vol. 178, no. S74, pp. 181–210.

Panaretos, V. M., Pham, T., and Yao, Z. (2014). "Principal flows". In: *Journal of the American Statistical Association* vol. 109, no. 505, pp. 424–436.

Pennec, X. (2018). "Barycentric subspace analysis on manifolds". In: *The Annals of Statistics* vol. 46, no. 6A, pp. 2711–2746.

— (2019). "Curvature effects on the empirical mean in Riemannian and affine Manifolds: a non-asymptotic high concentration expansion in the small-sample regime". In: *arXiv preprint arXiv:1906.07418*.

Pennec, X., Fillard, P., and Ayache, N. (2006). "A Riemannian framework for tensor computing". In: *International Journal of computer vision* vol. 66, no. 1, pp. 41–66.

Pennec, X., Sommer, S., and Fletcher, T. (2019). *Riemannian geometric statistics in medical image analysis*. Academic Press.

Pewsey, A. and García-Portugués, E. (2021). "Recent advances in directional statistics". In: *Test* vol. 30, no. 1, pp. 1–58.

Salili-James, A., Mackay, A., Rodriguez-Alvarez, E., Rodriguez-Perez, D., Mannack, T., Rawlings, T. A., Palmer, A. R., Todd, J., Riutta, T. E., Macinnis-Ng, C., et al. (2022). "Classifying organisms and artefacts by their outline shapes". In: *Journal of the Royal Society Interface* vol. 19, no. 195, p. 20220493.

Sommer, S. (2015). "Anisotropic distributions on manifolds: Template estimation and most probable paths". In: *Information Processing in Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28-July 3, 2015, Proceedings 24*. Springer, pp. 193–204.

— (2019). "An infinitesimal probabilistic model for principal component analysis of manifold valued data". In: *Sankhya A* vol. 81, no. 1, pp. 37–62.

Sommer, S. and Bronstein, A. (2020). "Horizontal flows and manifold stochastics in geometric deep learning". In: *IEEE transactions on pattern analysis and machine intelligence* vol. 44, no. 2, pp. 811–822.

Sommer, S., Lauze, F., Hauberg, S., and Nielsen, M. (2010). "Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations". In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11*. Springer, pp. 43–56.

Sommer, S. and Svane, A. M. (2015). "Modelling anisotropic covariance using stochastic development and sub-Riemannian frame bundle geometry". In: *arXiv preprint arXiv:1512.08544*.

Sun, J.-g. (1990). “Multiple eigenvalue sensitivity analysis”. In: *Linear algebra and its applications* vol. 137, pp. 183–211.

Symonds, M. R. and Blomberg, S. P. (2014). “A primer on phylogenetic generalised least squares”. In: *Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice*, pp. 105–130.

Thanwerdas, Y. (May 2022). “Riemannian and stratified geometries on covariance and correlation matrices”. Theses. Université Côte d’Azur.

Wright, S., Nocedal, J., et al. (1999). “Numerical optimization”. In: *Springer Science* vol. 35, no. 67-68, p. 7.

Yao, Z., Eltzner, B., and Pham, T. (2023). *Principal Sub-manifolds*. arXiv: 1604.04318 [stat.ME].

Younes, L. (2010). *Shapes and diffeomorphisms*. Vol. 171. Springer.