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

erc Geometric Statistics

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, Moceri, 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 3 and 4. 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 3 and 4. 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 3 and 4 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 3 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 3 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 4 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 3.4.1.2), 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 3.3.2 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 3 and 4 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 3, 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 4 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,

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

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

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 4 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, in Chapter 4, can also be considered a generalization of the principal flow to higher dimensions (in Section 4.4.3.1 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

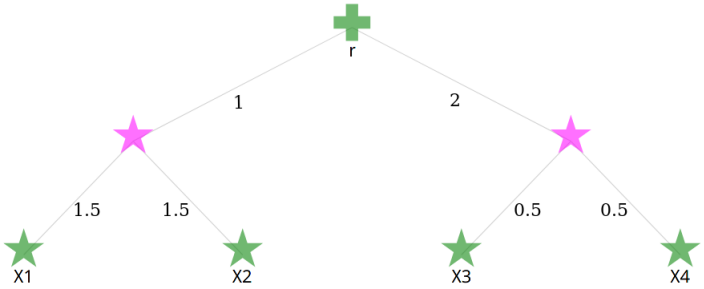
Phylogenetic PCA for manifold-valued observations

Contents

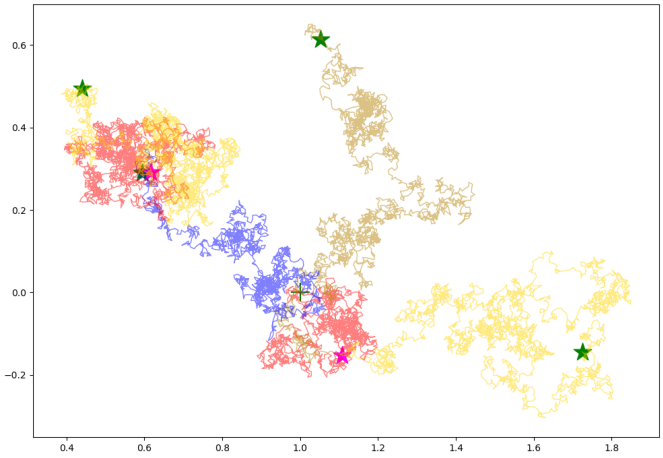
3.1 Introduction

From a mathematical point of view, a *phylogenetic tree* is a rooted, bifurcating tree graph, as exemplified in Figure *1a*. In evolutionary biology, such graphs represent evolutionary relationships between species. Each node represents a species, and its potential parent and child nodes represent its ancestor and descendant species, respectively. Each node (species) has an associated value, called a *trait*, for example a DNA sequence or a continuous measurement such as height. Assume, for the time being, that traits are values in $\mathbb{R}^d$. According to the Brownian motion model of evolution of such traits ([Cavalli-Sforza and Edwards 1967], [Felsenstein 1973]), the edges (also called branches) of a tree represent Brownian motions; the value (trait) of a node $n \in \mathbb{R}^d$ is the endpoint of a Brownian motion starting at the parent node $m \in \mathbb{R}^d$. The length of the edge connecting the nodes equals the duration of the Brownian motion - often in units of billions of years. In this work, the data we are given is, *1.*, a tree graph, representing a phylogenetic tree and, *2.*, the values of the leaf nodes of the tree. All other node values are unknown, in line with the fact that the leaf nodes represent recent, observable states while inner nodes represent hypothetical, unobserved states. From a statistical point of view, the important property of such data is that the observations (leaf node values) are not necessarily independent. The more recent their latest common ancestor is, the closer the values are likely to be. The field of *phylogenetic comparative methods* ([Felsenstein 1985], [Martins and Hansen 1997], [Harmon 2019]) comprises statistical methods that take into account the dependencies induced by the tree structure. One such method is *phylogenetic PCA* [Revell 2009], a version of PCA adapted to this specific type of non-independence.

Phylogenetic PCA (p-PCA) is defined for variables taking values in Euclidean space. In studies of morphological evolution of species, p-PCA has so far been applied to landmark shapes which, after procrustes alignment, are treated as elements of a Euclidean space [Polly et al. 2013]. This is common practice in the field of *geometric morphometrics* (see e.g. [Mitteroecker and Schaefer 2022], [Adams, Rohlf, and Slice 2013]), and corresponds to working in an approximation of a tangent space to Kendall's shape space (see Section 3.4.1.1). As opposed to this, we generalize p-PCA and the underlying model of Brownian motions to the setting where the nodes of the tree can take values in a general finite-dimensional Riemannian manifold. The methods we present are based on intrinsic (i.e. exact) computations, and when we do Euclidean computations they are performed in the exact tangent space. This implies that our method is appropriate for



(a)



(b)

Figure 3.1: *a*). An example of a phylogenetic tree. r is the root node and $x_1, \ldots, x_4$ are the leaf nodes. The edge lengths are written next to each edge. *b*). A realization of the tree in Figure a), generated from Brownian motions in $\mathbb{R}^2$. The green '+' at $(1, 0)$ is the root node. The two pink stars are the inner nodes. The 4 green stars are the leaf nodes. There are 6 Brownian motions, corresponding to the 6 edges in Figure a). If two edges in Figure a) share a node, the corresponding Brownian motions has different colours.

datasets with larger variance (cf. Section 2.3.1) and on any finite-dimensional shape manifold, including Kendall's. The method we present thus enables doing dimension reduction for shapes properly treated as points on a manifold, while accounting for the non-independence induced by the underlying tree structure.

3.1.1 Chapter organization and contributions

In Section 3.2, we describe phylogenetic PCA in the Euclidean setting, based on existing work. In Section 3.3, we describe the Brownian motion model on a Riemannian manifold and our extension of p-PCA to observations on such a manifold. We call this method *tangent phylogenetic PCA*. The basic ingredients are estimation of the root-node of the phylogenetic tree and the covariance of the Brownian motions. We define the root on a Riemannian manifold as a weighted Fréchet mean, which can be estimated by Riemannian gradient descent. We show that the root-estimation performed in [Polly et al. 2013] correponds approximately to taking a single gradient descent step in Kendall's shape space. In Section 4.6 we introduce the particular shape spaces that we illustrate the method on; the sphere (equivalent to Kendall's shape space for triangles), on which we simulate data, and the LDDMM landmark manifold, that we use to represent and analyse a data set of mammal jaws.

Relation of chapter to publications The material presented in this chapter is based on the paper *Tangent phylogenetic PCA* [Akhøj, Pennec, and Sommer 2023] published in the proceedings of the *Scandinavian Conference on Image Analysis*, 2023. Since the published version of the paper, the method has been formulated such that all estimators are now intrinsically defined. In particular, the root node has been rephrased as a weighted Fréchet mean, and the algorithm for estimating it has been formulated as a Riemannian gradient descent. Furthermore, the phylogenetic covariance estimator has been rephrased as a weighted tensor product of logarithms.

3.2 Phylogenetic PCA in Euclidean space

In this section, we describe Euclidean phylogenetic PCA [Revell 2009]. We define the tree structures with leaf node observations which constitutes the input to the algorithm, that is, it is given to us in advance. Furthermore, we define the underlying Brownian motion model on the tree and describe maximum likelihood estimators.

Phylogenetic PCA takes as input a rooted, bifurcating tree. We denote this by

$$(V, E, L),$$

where V is the node set, E is the set of directed edges, and $L : E \to \mathbb{R}_{>0}$ is a function giving the 'length' of an edge, $L(e) = L(n_i, n_j) \in \mathbb{R}_{>0}$, for

$e = (n_i, n_j) \in E$. A node $n \in V$ can have an associated value $x_n \in \mathbb{R}^d$. We will call such a tree a *p-tree*. P-PCA assumes that only values of leaf nodes are known. What makes the model Euclidean is the fact that the node values belong to a Euclidean space. In the subsequent sections, we will generalize this to allow values in a Riemannian manifold.

The length of a path between two nodes is the sum of the edge lengths between nodes in the path. There is a unique shortest path from the root r to any node $n_i \in V$, and we denote the length of this path by $L(r, n_i)$.

The *most recent common ancestor* of two nodes n_i and n_j, $\mathrm{MRCA}(n_i, n_j)$, is the unique node $n \in V$ from which there is a path to both n_i and n_j, and which minimizes the path-lengths $L(n, n_i)$ and $L(n, n_j)$.

We interpret a p-tree with leaf observations as a phylogenetic tree, where only the leaf nodes are recent enough in time to be known. Estimating phylogenetic trees (i.e. nodes, edges and edge-lengths) for various species is a developed and active field of research. See for example [Nyakatura and Bininda-Emonds 2012], describing the estimated tree which we use for analysing the data set of mammal jaws in Section 4.6.

3.2.1 The Brownian motion model on a tree

In this section we describe the model of a p-tree with branches representing Brownian motions on $\mathbb{R}^d$, and how this leads to a joint normal distribution of the leaf nodes with explicit maximum likelihood estimates.

Let $\{B_t^{x_0}\}_{t \geq 0}$ denote a Brownian motion (BM) in $\mathbb{R}^d$ with initial value $x_0 \in \mathbb{R}^d$. Given a p-tree (V, E, L), the associated Brownian motion model assumes that each edge of the tree corresponds to the path of an $\mathbb{R}^d$-valued Brownian motion. All such Brownian motions are assumed to evolve according to a common covariance matrix. Namely, the values $x_{n_i}, x_{n_j} \in \mathbb{R}^d$ of two nodes $n_i, n_j \in V$ s.t. $(n_i, n_j) \in E$ and $L(n_i, n_j) = t'$ are related by

$$x_{n_j} \sim \mathbb{P}(B_{t'}^{x_{n_i}}), \tag{3.2.1}$$

that is, x_{n_j} is an observation from the time-t' transition distribution of $B^{x_{n_i}}$. Thus all nodes except for the root are random variables. Figure 1b shows one realization of the p-tree in Figure 1a.

The assumption that the nodes are endpoints of Brownian motions implies that the joint distribution of leaf-node values $\{x_1, \ldots, x_N\}$ is a normal distribution on $\mathbb{R}^{N \cdot d}$. In [Felsenstein 1973], it is shown that

$$(x_1, \ldots, x_N) \sim N(Dr, R \otimes C), \tag{3.2.2}$$

a normal distribution with mean Dr and covariance matrix $R \otimes C$. Here, D is an $(N \cdot d) \times d$ design matrix where $D_{ij} = 1$, if i, j satisfies the inequality $(j-1) \cdot N < i \geq j \cdot N$, and 0 otherwise. That is, the j'th column of Dr contains only 0's as well as the j'th element of r repeated N times. $r \in \mathbb{R}^d$ is the value of the root node. $R \in \mathbb{R}^{d \times d}$ is the *phylogenetic covariance matrix* which is the covariance matrix for each Brownian motion with R_{ij} being the covariance between coordinate i and j. $C \in \mathbb{R}^{N \times N}$ is the *evolutionary covariance matrix* determined by the p-tree; $C_{ij} = L(r, \text{MRCA}(n_i, n_j))$. That is, C_{ij} is the total length of edges occuring both in the shortest path from the root node to n_i and to n_j. $\otimes$ is the Kronecker product of matrices. To sum up; the Brownian motion model structured according to a p-tree implies that the N leaf node observations can be considered as a single observation from a normal distribution on the product space $\mathbb{R}^{N \cdot d}$. Notice that if the p-tree contains only 2 leaf-nodes each connected to the root by an edge of length 1, then C is the identity matrix and the leaves are 2 independent samples from a normal distribution with mean r and covariance R.

From this model, one can derive maximum likelihood estimates (MLE's) for the lower dimensional parameters $r \in \mathbb{R}^d$ and $R \in \mathbb{R}^{d \times d}$. The MLE's of the phylogenetic root, $\hat{r}$, and covariance, $\hat{R}$, is

$$\hat{r} = \left(1^T C^{-1} 1\right)^{-1} \left(1^T C^{-1} X\right), \tag{3.2.3}$$

$$\hat{R} = \frac{1}{N-1} \left(X - \hat{r}^T\right)^T C^{-1} \left(X - \hat{r}^T\right), \tag{3.2.4}$$

where $X \in \mathbb{R}^{N \times d}$ is the matrix containing the leaf-node values in rows and $X - \hat{r}^T \in \mathbb{R}^{N \times d}$ is the row-wise difference (see [Harmon 2019], [Martins and Hansen 1996]). These estimators are identical to the ordinary empirical mean and covariance, (2.1.1) and (2.1.6), except for the weighting caused by multiplying by C^{-1}. This amounts to downweighing leaves with longer branch lengths to the root, and the weights of two leaves are closer the more branches they share.

3.2.2 Phylogenetic PCA

As described in Example 2.1.3, PCA is based on eigendecomposition of the empirical covariance matrix. Under the model of Brownian motions along a tree, the covariance is the covariance R of the Brownian motions. *Phylogenetic PCA* is defined as PCA based on eigendecomposition of the empirical estimate $\hat{R}$.

For clarity, and for comparison with the manifold version presented in Section 3.3.4, we present Algorithm 1 as an example of how to use phylogenetic PCA. The output of the algorithm is the observations projected to the k'th eigenspace.

Other possible outputs, depending on the particular application, could for example be visualization of the first eigenvector, i.e. the direction of greatest variation. Different such uses of the extracted eigenvectors and eigenvalues are straight-forward reformulations of Algorithm 1. Another choice we made is to let the dimension k be an input to the algorithm. The choice of k depends on the application. If the goal is visualization, k is set to 2 or 3. The eigenvalues of $\hat{R}$ can also be inspected to find a dimension k representing a sufficient, relative to the aim, amount of variation in the data. Automatic selection methods also exists, see e.g. [Minka 2000] which is based on Bayesian model selection.

Algorithm 1 Phylogenetic PCA

Input: A p-tree (V, E, L) with leaf observations $\{x_i\}_{i=1..N} \subset \mathbb{R}^d$, and a choice $k \in \{1, \ldots, d\}$ of dimension of the lower dimensional representation

Output: The observations projected to the k'th eigenspace.

1: Compute ML estimates of the root $\hat{r}$ (eq. (3.2.3)) and center the data around $\hat{r}$,

$$(x_i)_{\hat{r}} := x_i - \hat{r}, \quad i = 1..N.$$

2: Compute the covariance estimate $\hat{R}$ (eq. (3.2.4)).

3: Eigen-decompose $\hat{R}$ and proceed as in ordinary PCA: project the $\hat{r}$-centered observations $\{(x_i)_{\hat{r}}\}_{i=1..N}$ to the subspace spanned by the first k eigenvectors, to get principal components $\{\tilde{x}_i\}_{i=1..N}$.

return $\{\tilde{x}_i\}_{i=1..N}$

3.3 Phylogenetic PCA on a Riemannian manifold

We now assume given a p-tree (V, E, L) where the leaf observations $x_1, \ldots, x_N$ are points on a Riemannian manifold (M, g). The model of Euclidean Brownian motions structured according to a p-tree is directly transferable to the manifold setting. The edges now represent Brownian motions on the Riemannian manifold M so that Equation 3.2.1 still holds, with the transition distribution being that of a Brownian motion on M. Such Brownian motions are introduced in the next subsection. Following that, we generalize the phylogenetic root and covariance estimators to manifold valued data. *Tangent phylogenetic PCA* will then be defined as a tangent PCA centered at the root estimate and based

on an eigendecomposition of the estimated covariance matrix of a Riemannian Brownian motion.

3.3.1 Brownian motion on a Riemannian manifold

A Riemannian Brownian motion is a stochastic process on a Riemannian manifold, (M, g), whose infinitesimal generator is the Laplace-Beltrami operator [Hsu 2002]. Similarly to the Euclidean case, we will denote such a Brownian motion with initial point $p_0 \in M$ by $\{B_t^{p_0}\}_{t \geq 0}$. The process can be simulated via a scheme similar to Euler-Maruyama [Said and Manton 2012] as follows. Choose a stepsize $\tau > 0$ and an orthonormal basis of $T_{p_0}M$ w.r.t. the Riemannian metric. Sample a vector $v \in T_{p_0}M$ from a standard normal distribution w.r.t. this basis. Starting from p_0, the subsequent point is generated as $B_\tau^{p_0} = \exp_{p_0}\left(\frac{v}{\|v\|_g}\tau\right) \in M$. At any time-point t of the trajectory, the next point $B_{t+\tau}^{p_0}$ is generated in the same way but letting the base point be $B_t^{p_0}$. The resulting sequence of points will be approximate samples from a Riemannian Brownian motion.

A Riemannian Brownian motion, the type of process described above, is isotropic, meaning that it spreads out equally in all directions w.r.t. the Riemannian metric. This can be generalized to non-isotropic Brownian motions, parameterized by an initial point p_0 and an initial covariance matrix R. The pair (p_0, R) is modelled as a point in the frame bundle of M. The construction of this stochastic process is more involved than in the isotropic case, since the frame (covariance matrix) needs to transported between tangent spaces by means of parallel transport. We refer to [Pennec, Sommer, and T. Fletcher 2019], chapter 10, for details on the construction.

3.3.2 Phylogenetic root estimation on a Riemannian manifold

In order to formulate the root estimator intrinsically, we rewrite the Euclidean expression (Equation (3.2.3)) as follows,

$$\begin{aligned}\hat{r} &= \left(1^T C^{-1} 1\right)^{-1} \left(1^T C^{-1} X\right) \\ &= \sum_{i=1}^{N} x_i w_i \\ &= \operatorname*{argmin}_{r \in \mathbb{R}^d} \sum_{i=1}^{N} w_i \cdot d_{\mathbb{R}^d}(r, x_i)^2. \end{aligned} \tag{3.3.1}$$

Here $w_i = \frac{1}{\sum_{j=1}^{N} \tilde{w}_j} \tilde{w}_i$ and the unnormalized weights are the column sums $\tilde{w}_i = \sum_{j=1}^{N} (C^{-1})_{j,i}$. The distance, $d_{\mathbb{R}^d}(p, q) = \sqrt{\sum_{i=1}^{d} (p_i - q_i)^2}$, is the ordinary

Euclidean one. If this distance is exchanged for the geodesic distance w.r.t. g, denoted by d, then Equation (3.3.1) is well defined on a Riemannian manifold. We therefore define the Riemannian root estimate by

$$\hat{r} = \operatorname*{argmin}_{r \in M} \sum_{i=1}^{N} w_i \cdot d_g(r, x_i)^2,$$

which is a weighted Fréchet mean. For conditions regarding existence and uniqueness of weighted Fréchet means, see [Karcher 1977], [Afsari 2011], [M. Arnaudon and Miclo 2014]. Minimizers of the corresponding objective function,

$$f(p) = \sum_{i=1}^{N} w_i \cdot d_g(r, x_i)^2, \tag{3.3.2}$$

can be found by means of Riemannian gradient descent. For details on optimization on Riemannian manifolds, see [Udriste 1994] and [Absil, Mahony, and Sepulchre 2008]. Karcher [Karcher 1977] showed that the gradient of f is

$$\nabla f \quad : \quad p \mapsto \nabla f|_p = -2 \sum_{i=1}^{N} w_i \log_p(x_i) \ \in \ T_p M,$$

which is the Euclidean root estimate of the observations represented in the tangent space, scaled by a factor of 2 (which could of course be removed by rescaling the objective function). The gradient descent algorithm for estimating the phylogenetic root on a Riemannian manifold is given in Algorithm 2. See Appendix **??** for references discussing convergence of gradient descent algorithms on Riemannian manifolds.

Algorithm 2 Root node estimation

Input: A p-tree (V, E, L) with leaf observations $\{x_i\}_{i=1\ldots N} \subset M$, an initial guess $r_0 \in M$, a convergence threshold $\epsilon > 0$ and a stepsize $\tau > 0$.

Output: An estimate of the root $\hat{r} \in M$.

1: Set $\hat{r} = r_0$ and pick any $v_{\hat{r}} \in T_{\hat{r}} M$ s.t. $\|v_{\hat{r}}\|_g > \epsilon$.

2: **while** $\|v_{\hat{r}}\|_g > \epsilon$ **do**

3: Map the data to $T_{\hat{r}} M$, $(x_i)_{\hat{r}} := \log_{\hat{r}}(x_i)$, $i = 1 \ldots N$.

4: Compute the gradient $v_{\hat{r}} := \nabla f|_{\hat{r}} = -\sum_{i=1}^{N} w_i (x_i)_{\hat{r}}$.

5: Update $\hat{r}$ by $\hat{r} \leftarrow \exp_{\hat{r}}(-\tau v_{\hat{r}}) \in M$

6: **end while**

return $\hat{r}$

3.3.3 Phylogenetic covariance estimation on a Riemannian manifold

As we did for the root estimator, we reformulate the Euclidean phylogenetic covariance estimator (3.2.4) as a weighted sum - this time a weighted sum of outer products;

$$\hat{R} = \frac{1}{N-1}\left(X - \hat{r}^T\right)^T C^{-1}\left(X - \hat{r}^T\right) \tag{3.3.3}$$

$$= \frac{1}{N-1}\sum_{i,j=1}^{N}(C^{-1})_{i,j}\cdot(x_i - \hat{r})(x_j - \hat{r})^T. \tag{3.3.4}$$

As explained in Section 2.4.1 and appendix A.2.1, we can make intrinsic sense of an outer product by considering it to be a coordinate expression of a certain tensor product. Thus, we define the intrinsic phylogenetic covariance estimator on a Riemannian manifold to be

$$\hat{R} = \frac{1}{N-1}\sum_{i,j=1}^{N}(C^{-1})_{i,j}\cdot log_{\hat{r}}(x_i)\otimes log_{\hat{r}}(x_i). \tag{3.3.5}$$

Cf. Lemma A.2.1, after choosing a chart around $\hat{r}$ and thus a basis for $T_{\hat{r}}M$, the coordinate expression of the estimator is

$$\hat{R} = \frac{1}{N-1}\sum_{i,j=1}^{N}(C^{-1})_{i,j}\cdot log_{\hat{r}}(x_i) log_{\hat{r}}(x_i)^T g_{\hat{r}}. \tag{3.3.6}$$

where $g_p \in \mathbb{R}^{d\times d}$ denotes the matrix representation of the Riemannian metric at a point $p \in M$, in the chosen chart.

3.3.4 Phylogenetic PCA on Riemannian manifolds

Given an estimate of the root node, $\hat{r}$, and of the phylogenetic covariance, $\hat{R}$, we define phylogenetic PCA on a Riemannian manifold as tangent PCA computed in the tangent space at $\hat{r}$, using covariance $\hat{R}$ (see Section 3.3.4).

Algorithm 3 is the manifold version of Algorithm 1. It extracts the k-dimensional eigenspace $V \subset T_{\hat{r}}M$ of the covariance estimate $\hat{R}$, projects the log-transformed observations $\{\log_{\hat{r}}(x_i)\}_{i=1..N}$ to V and maps them out to M via the exponential map. As discussed in Section 2.4.1 on tangent PCA, since $V \subset T_pM$ is a k-dimensional subspace, then $\exp_p(V \cap \mathcal{C}_p) \subset M$ is a k-dimensional submanifold, where $\mathcal{C}_p$ is the injectivity domain at p. Therefore, if $\{\log_{\hat{r}}(x_i)\}_{i=1..N} \subset \mathcal{C}_{\hat{r}}$, then the exponential image of their projection to V are points on a k-dimensional submanifold of M. In this sense they constitute a k-dimensional representation of the original data.

Algorithm 3 Tangent phylogenetic PCA

Input: A p-tree (V, E, L) with leaf observations $\{x_i\}_{i=1..N} \subset M$, an estimate of the root node $\hat{r}$ and a dimension $k \in \{1, \dots, d\}$ of the reduced representation.
Output: The observations mapped to a k dimensional submanifold of M.

1: Map the data to $T_{\hat{r}}M$, $\bar{x}_i := \log_{\hat{r}}(x_i)$.

2: Compute the phylogenetic covariance matrix $\hat{R} = \sum_{i,j=1}^{N} (C^{-1})_{i,j} \cdot \bar{x}_i \bar{x}_i^T g(\hat{r})$.

3: Eigen-decompose $\hat{R}$ and proceed as in ordinary PCA in $T_{\hat{r}}M$: project the observations $\{\bar{x}_i\}_{i=1...N}$ to the subspace spanned by the first k eigenvectors to get principal components $\{\bar{x}_i^k\}_{i=1...N} \subset T_{\hat{r}}M$.

return $\{\tilde{x}_i\}_{i=1..N} := \{\exp_{\hat{r}}(\bar{x}_i^k)\}_{i=1..N}$.

3.3.4.1 Computational aspects of tangent p-PCA

The potentially most costly part of computing tangent p-PCA is the root-estimation, Algorithm 2, specifically the computation of logarithms. In each iteration of the gradient descent, N Riemannian logarithms must be computed. So the computational cost of the algorithm is proportional to the cost of computing the log map. On a manifold with closed form logs, such as Kendall's shape space, the cost is negligible. In the worst case, however, computing the log map is an optimization problem with parameter-space $\mathbb{R}^{\dim(M)}$ and an objective function that is expensive to evaluate. However, the log-computations in the algorithm are independent of each other, so can be done in parallel. Computing the Riemannian logarithm is a typical bottleneck in (Riemannan) geometric statistics, see Appendix A.1 for a discussion on this as well as a suggestion for how to speed up computation of the log in certain settings.

3.4 Simulations and applications

In this section, we investigate the behaviour of tangent p-PCA via simulations on the sphere (Kendall's shape space of triangles in the plane) and by applying it to a data set of mammal jaws, represented as landmark shapes in the LDDMM landmark manifold. All implementations are based on the Python library *jaxgeometry* which provides code for computational Riemannian geometry based on automatic differentiation (for more information see *https://bitbucket.org/stefansommer/jaxgeometry/* and [Kühnel, Sommer, and A. Arnaudon 2019]).

3.4.1 Landmark shape spaces

The methods we have developed in this chapter works on any finite-dimensional Riemannian manifold, but is motivated by applications to *shape spaces*. As examples of such manifolds we consider *Kendall's shape space* and the *LDDMM* landmark manifold. In this section we introduce these two manifolds. See also [Salili-James et al. 2022] for an overview and evaluation of various shape frameworks, including geometric morphometrics and LDDMM, applied to the problem of classifying shapes.

3.4.1.1 Kendall's shape space

As mentioned in the introduction in Chapter 1, the starting point of Kendall's shape space is the Euclidean space $\mathbb{R}^{d\cdot k}$, which can be identified with the space of d by k matrices $\mathbb{R}^{d\times k}$. We interpret a matrix $x = [q_1, \dots, q_k] \in \mathbb{R}^{d\times k}$ as consisting of k points $q_i \in \mathbb{R}^d$, $i = 1, \dots, k$, called *landmarks*. We will call such a matrix a *landmark matrix*. The landmarks could for example be points along a curve outlining some 'shape' in the plane, $\mathbb{R}^2$. If all landmarks in a matrix are translated by some common vector, scaled by a common factor and/or rotated by a commmon rotation, the collection of transformed points intuitively still looks like the same 'shape' - just placed differently relative to the coordinate system. By identifying matrices $x \in \mathbb{R}^{d\times k}$ differing only by a translation, a scaling or a rotation we can therefore construct a space in which each point is a *shape*, and different points represent different shapes. Specifically, we say that $x, x' \in \mathbb{R}^{d\times k}$ are equivalent, $x \sim x'$, if there exists some $p \in \mathbb{R}^d$, $c \in \mathbb{R}$ and a rotation matrix $R \in \mathrm{SO}(d)$ such that $x' = c \cdot R\,[q_1 + p, \dots, q_k + p]$. This equivalence relation defines a quotient space which is a smooth manifold for $d \leq 2$, but has singular points of non-smoothness for $d > 2$, corresponding to certain degenerate landmark matrices. This manifold can be equipped with a natural Riemannian metric derived from the Frobenius metric on $\mathbb{R}^{d\times k}$, resulting in *Kendall's shape space*.

Kendall's shape space can be constructed in two steps. The first step is to quotient $\mathbb{R}^{d\times k}$ out by translation and scaling by considering only the following subset,

$$\mathbb{S}_d^k = \left\{ x = [q_1, \dots, q_k] \in \mathbb{R}^{d\times k} \,\middle|\, \|x\|_F = 1, \sum_{i=1}^{k} q_i = 0 \right\},$$

where $\|A\|_F = \sqrt{\mathrm{Tr}(AA^T)}$, for $A \in \mathbb{R}^{d\times k}$, is the Frobenius metric. $\mathbb{S}_{k,d}$ is called the *pre-shape* space and is a manifold diffeomorphic to a hypersphere of dimension $d(k-1)-1$. Kendall's shape space is then the pre-shape space

quotiented out by the rotation group, $\Sigma_d^k = \mathbb{S}_d^k / SO(d)$, equipped with a quotient Riemannian metric derived from the Frobenius metric. The induced geodesic distance d can be expressed in terms of the sphere distance $d_{\mathbb{S}_d^k}$ on the preshape space as

$$d(\pi(x), \pi(y)) = \min_{R \in SO(d)} d_{\mathbb{S}_d^k}(x, Ry),$$

where $\pi : \mathbb{S}_d^k \to \Sigma_d^k$ is the projection of a point on the pre-shape sphere to the equivalence class containing it in the quotient manifold Σ_d^k.

Kendall's shape space is convenient from a computational point of view: closed form expressions exist for the exponential and the logarithm, and parallel transport can be computed fairly cheaply (see [Guigui and Pennec 2022]). All such computations are implemented in Geomstats.

Procrustes alignment Given a preshape x, its Procrustes alignment to preshape y is given by

$$x_y = \min_{R \in SO(d)} \|x - Ry\|_F,$$

that is, x is rotated so as to minimize the sum of squared distances between each landmark in x with the corresponding landmark in y. In the geometric morphometrics literature it is argued that Procrustes alignment is an approximate way of mapping (the equivalence class containing) x to the tangent space of Kendall's shape space at (the equivalence class containing) y. See [Dryden and K. V. Mardia 2016], Chapter 4, for details. We discuss this practice in geometric morphometrics further in Section 3.5.1.

3.4.1.2 The LDDMM landmark manifold

For our purposes, the *Large deformation diffeomorphic metric mapping* (LDDMM) framework can be seen as a way to define a Riemannian metric g_m on $\mathbb{R}^{d \times k}$ that seems reasonable when a point in $\mathbb{R}^{d \times k}$ is interpreted as a collection of k landmarks, i.e. a landmark matrix. What we refer to as the *LDDMM landmark manifold* is the Riemannian manifold $(\mathbb{R}^{d \times k}, g_m)$. We give more details on the metric below, but for now we mention that this manifold is geodesically complete, and so the induced geodesic distance metric d_{g_m} on $\mathbb{R}^{d \times k}$ is complete (see [Younes 2010] and [Micheli, Michor, and Mumford 2012]). Since this is a distance metric on $\mathbb{R}^{d \times k}$ it assigns non-zero distance between landmark matrices $x, x' \in \mathbb{R}^{k \times d}$ that represent the same shape according to Kendall's shape space. In this sense, d_{g_m} is not a distance metric on shapes. Therefore, in practice, a given set of observed landmark matrices are first mapped to the

corresponding pre-shapes and then Procrustes aligned before treating them as points on the LDDMM landmark manifold $(\mathbb{R}^{k\times d}, g_m)$.

We now give a brief description of the Riemannian metric g_m and the induced distance metric d_{g_m}. The idea of the LDDMM metric is to consider the space of *diffeomorphisms* on $\mathbb{R}^d$, i.e. smooth maps from $\mathbb{R}^d$ to itself, and a Riemannian metric on this space. This induces a Riemannian metric g_m on $\mathbb{R}^{d\times k}$, and thereby also a geodesic distance metric d_{g_m} on $\mathbb{R}^{d\times k}$. The distance between two landmark matrices $x, y \in \mathbb{R}^{d\times k}$, defined in this way, can be interpreted as the least 'energy' needed to deform $\mathbb{R}^d$ in such a way that each landmark in x coincides with the corresponding landmark in y.

As opposed to the metric on Kendall's shape space, the LDDMM Riemannian metric can be expressed fairly explicitly. More specifically, we can explicitly express the *cometric*, $g_m^\star$, whose matrix representation is the inverse of the metric matrix. The cometric at $x \in \mathbb{R}^{d\times k}$ is a block matrix

$$g_m^\star(x) = [(K_x)_{i,j}]_{i,j=1..k}, \qquad \text{for} \quad (K_x)_{ij} = \varphi(q_i, q_j) I_d \in \mathbb{R}^{d\times d}$$

where $x = [q_1, \ldots, q_k]$ is a matrix consisting of k landmarks $q_i \in \mathbb{R}^d$, I_d is the d by d identity matrix and φ is a *kernel function*. See [Michor 2020] for a derivation. A standard choice is a Gaussian kernel $\varphi(q_i, q_j) := \beta \exp^{-\|q_i - q_j\|^2/2\sigma^2} \in \mathbb{R}$ with parameters $\beta, \sigma > 0$ (see [Pennec, Sommer, and T. Fletcher 2019], [Younes 2010], [Miller, Trouvé, and Younes 2002]). The kernel determines how correlated nearby landmarks are under low-energy deformations, in the following sense: for larger σ's, nearby points $x, y \in \mathbb{R}^d$ are mapped to nearby points $\phi(x), \phi(y) \in \mathbb{R}^d$ by low-energy deformations ϕ. For smaller σ's, the points $\phi(x), \phi(y)$ can be further apart and each depends less on the value of the other. The σ parameter adds modelling flexibility, but also implies that there is no canonical metric on the LDDMM landmark manifold, as there is on Kendall's shape space.

In the case of LDDMM there are no closed-form expressions for any of the Riemannian operations that are used in geometric statistics. We therefore rely on numerical integration and optimization methods, leading to increased computational costs. The operations are implemented, based on automatic differentiation, in jaxgeometry (*http://bitbucket.org/stefansommer/jaxgeometry*).

3.4.2 Simulations on the sphere

We generate observations on the unit sphere, $\mathbb{S}^2$, by simulating isotropic Brownian motions on $\mathbb{S}^2$ structured according to a p-tree. Note that Kendall's shape space for 3 landmarks in $\mathbb{R}^2$ can be identified with $\mathbb{S}^2$. The topology of the tree is the one shown on Figure 1a, with branch lengths that are scaled up by a common

factor. We simulate isotropic Brownian motions using the Euler-Maruyama-like scheme described in Section 3.3.1. The true root is the north pole, and the convergence criterion parameter is set to $\epsilon = 10^{-5}$. For each of 1000 simulated trees, we estimate the root node via algorithm 2 and compute its error as the Riemannian distance to the true root. A single realization of a p-tree is shown in Figure 3. The error-distribution is shown in Figure 2. We note that it resembles a χ^2-distribution. When the leaf nodes are Euclidean, the root is the mean of a Normal distribution, and it is well known that the exact error-distribution in this case is χ^2-distributed.

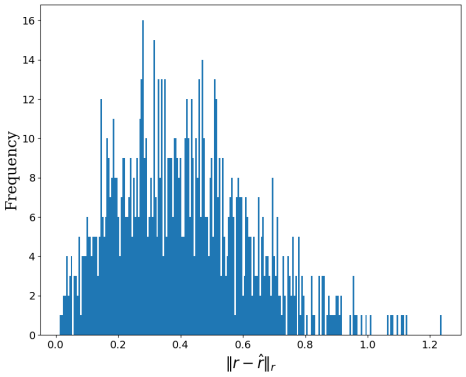


Figure 3.2: Distribution of the Riemannian distance $\|r - \hat{r}\|_r$, i.e. the geodesic distance between the true root and the root estimate, based on 1000 simulated trees on the sphere. For reference, the geodesic distance from the north pole (the true root) to a point on the equator is $\pi/2 \approx 1.57$.

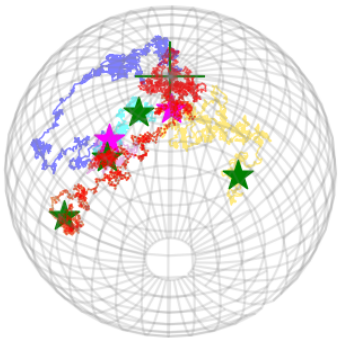


Figure 3.3: A realization of the tree in Figure 1a, generated from Brownian motions on the sphere. The green '+' is the root node. The two pink stars are the inner nodes. The 4 green stars are the leaf nodes.

Figure 4 shows a histogram of the number of iterations until convergence for two different choices of initial root-value: the Fréchet mean and the south pole, respectively. We see that when the initial point is the Fréchet mean, the root estimate converges in 2-5 iterations for most datasets. When the initial point is the south pole, the root estimate converges in 4-11 iterations for most datasets.

3.4.3 Mammal jaws data set

In this section, we analyse a data set of mammal jaws. For each of 113 mammal species, the data set contains from 1 to 8 jaws, each represented by 14 landmarks in $\mathbb{R}^2$. Figure 3.6 shows the 113 landmark shapes represented in $\mathbb{R}^2$. For a

full description of the data, see [Conith, Meagher, and Dumont 2018]. The phylogenetic tree is estimated in [Nyakatura and Bininda-Emonds 2012].

The first step of our analysis is to map landmark matrices to preshapes by scaling and centering them and then perform Procrustes alignment, as described in Section 3.4.1.1. Then we compute the Euclidean mean of the jaw observations within each species, and proceed with these 113 points instead of the full data set. Then, we consider these mean values as points in the LDDMM landmark manifold with a Gaussian kernel $k(q_i, q_j) := \beta \exp^{-\|q_i - q_j\|^2/2\sigma^2} \in \mathbb{R}$, $q_i \in \mathbb{R}^2$. In our analysis, we set $\beta = 1$, and $\sigma = 1.5\omega$. Here, $\omega = \frac{1}{113}\sum_{i=1}^{113} \psi_i$, where ψ_i is the average Euclidean distance between the landmarks in shape i. This yields $\sigma = 0.208$, which is thus proportional to the typical distance between landmarks within a typical jaw shape.

We start out by estimating the root node using algorithm 2. The results are presented in Figure 3.6. As initial estimate we use the root estimate from Euclidean p-PCA. The algorithm converges in 8 iterations with a threshold of $\epsilon = 10^{-4}$. The root estimate deviates visibly from the Euclidean root estimate.

We then perform tangent p-PCA according to Algorithm 3. We compute the phylogenetic covariance matrix $\hat{R}$ in $T_{\hat{r}}M$ and plot its eigenvalues to determine a dimension k which describes most of the variability, while reducing the dimension significantly. The eigenvalues are plotted in Figure 5. The first 6 eigenvectors describes a large proportion of the (non-phylogenetic) variation, thus we choose to represent the observations in $k = 6$ dimensions. Figure 3.7 below shows the projected observations for both tangent p-PCA and for Euclidean p-PCA on Procrustes aligned landmark shapes.

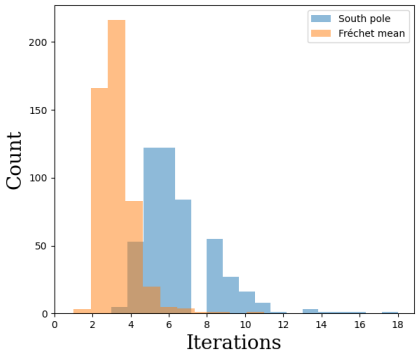


Figure 3.4: Histogram showing the number of iterations until convergence of the root-estimate on the sphere, for two different choices of initial value: the Fréchet mean (yellow) and respectively the south pole (blue). For each initial value, 500 trees was simulated.

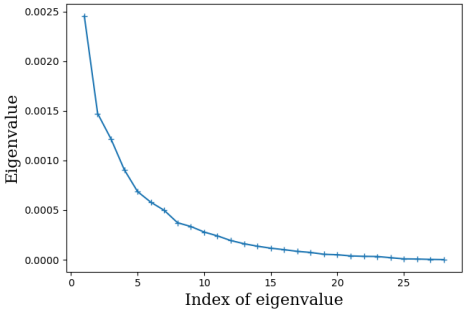


Figure 3.5: Eigenvalues from tangent p-PCA on the Mammal jaws data set.

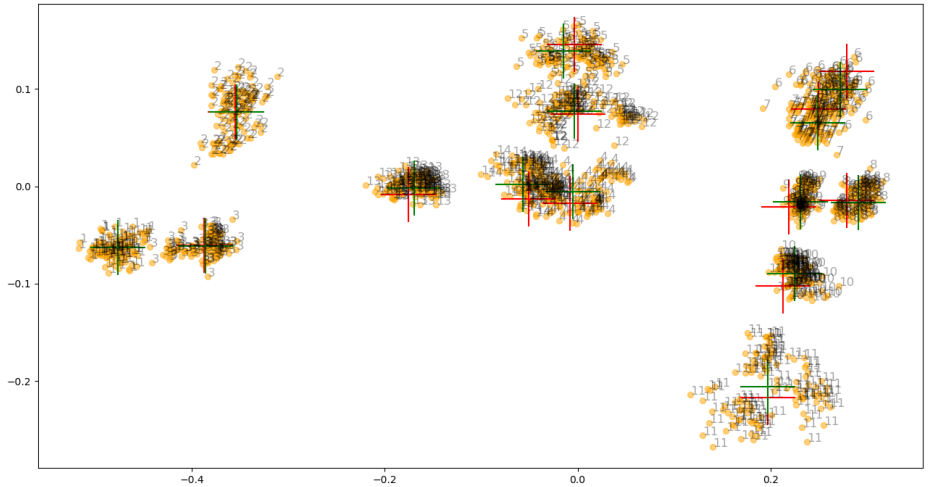


Figure 3.6: Illustration of the Mammal jaws data set (yellow dots) and 2 root estimates, one by algorithm 2 (green '+'-signs) and one by Euclidean p-PCA (red '+'-signs) after Procrustes alignment. Each shape consists of 14 landmarks (dots), numbered in the figure.

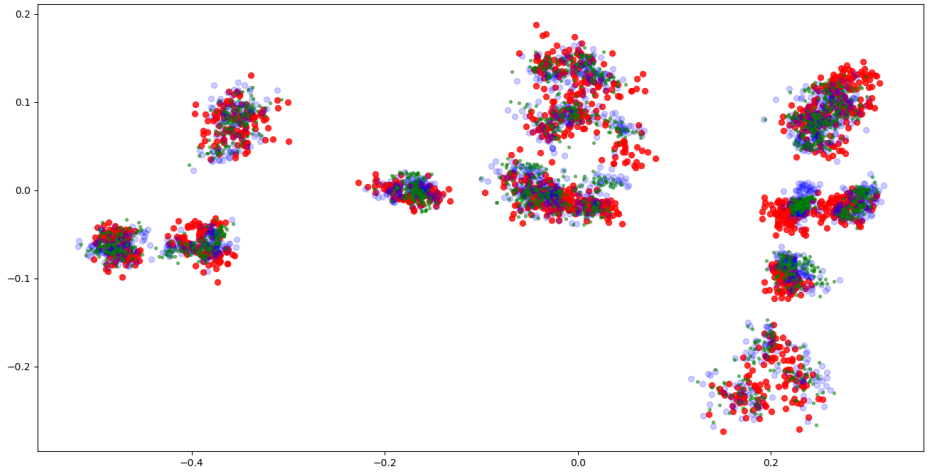


Figure 3.7: Blue dots are the Mammal jaws observations after Procrustes alignment. Red dots are the observations in $T_{\hat{r}}M$ projected to the first 6 tangent p-PCA eigenvectors and then mapped back to M via the Riemannian exponential. Green dots are the observations projected to the first 6 eigenvectors of Euclidean p-PCA.

3.5 Discussion and conclusions

3.5.1 Relations between our approach and that of geometric morphometrics

The field of *geometric morphometrics* (GM) ([Mitteroecker and Schaefer 2022], [Adams, Rohlf, and Slice 2013])) studies morphological features of biological species. One such morphological feature is *shape*, often in the sense of Kendall's shape space, i.e. landmark matrices are analysed in a way that removes the effects of translation, scale and rotation (or subsets of these transformations). A typical approach in the field amounts to representing the observed shapes in a tangent space of Kendall's shape space and then applying Euclidean formulas. This tangent space representation is often done in an approximative way, via Procrustes alignment of an observed set of landmark matrices, as described in Section 3.4.1.1. This is the approach taken by [Polly et al. 2013] for phylogenetic PCA on landmark shapes.

Our contribution to the problem of applying phylogenetic PCA on shapes, relative to the GM approach, is to define the estimators of the root and covariance in an intrinsic way on a general Riemannian manifold and give algorithms for their computation based on the exact geometric objects - e.g. the actual tangent space instead of an approximation. We sum up the practical advantages of this as follows,

A) the estimators can be applied on any finite-dimensional shape manifold, not only Kendall's shape space,

B) the root estimator can handle larger data variation.

In point B) we don't mention the covariance estimator, since this is based on the log and therefore on a tangent space representation of the data. As opposed to this, our definition of the root as a weighted Fréchet mean does not rely on a tangent space.

A survey of the GM literature has led us to conclude that there seems to be a consensus in the field that working in the tangent space of Kendall's shape space, or approximations thereof, is sufficient for the data sets that are encountered in practice (see [Rohlf 1999] [Klingenberg 2020], the section *Tangent Space Approximation to Shape Spaces* and references therein). I.e. that the observed variation is usually sufficiently small for the distances in the tangent space to

approximate the intrinsic distances well. [1] This seems to imply that point B) above is not of practical relevance to the field, when the shapes representation is Kendall's shape space. We elaborate on this point in the next section.

3.5.1.1 The root estimator

In [Polly et al. 2013], the method for finding the root given a data set of landmark matrices consists of the following two steps,

1) Compute generalized Procrustes alignment, that is, compute the Procustes mean (see [Dryden and K. V. Mardia 2016]) and Procrustes align the pre-shapes to this.

2) Use the Euclidean root formula, (3.2.3), on the aligned pre-shapes.

Step *1)* corresponds approximately to representing the observations in the tangent space at the Fréchet mean in Kendall's shape space (see [Dryden and K. V. Mardia 2016], Chapter 4). This corresponds to line 3 of our Algorithm 2, if the algorithm is initialized at the Fréchet mean. Step *2)* corresponds to computing the gradient of the objective function for the root estimator, (3.3.2), i.e. line 4 of our algorithm. All in all, the method in [Polly et al. 2013] corresponds, approximately, to taking a single step of our gradient descent algorithm when the manifold is Kendall's shape space and the initial point is the Fréchet mean.

Due to curvature, the further away the Fréchet or Procrustes mean of the observations is from the Phylogenetic root, the more likely it is that taking a single step in Algorithm 2 is not sufficient. In Kendall's Shape space the computational cost of taking a step in Algorithm 2 is negligible - less than a few seconds. Indeed, computing the Procrustes mean (commonly used in GM) already involves iteration of the same type of computation. Therefore, we see no reason not to continue the gradient descent until convergence with respect to some desired level of accuracy. This holds even in the case where the data is concentrated enough for the tangent space to be a good approximation - indeed, if this is the case, the algorithm will converge faster. In Section 3.4.2 we investigated the convergence behaviour of the root estimator on the sphere (i.e. Kendall's shape space for 3 landmarks in $\mathbb{R}^2$) and in the LDDMM landmark space. On both manifolds, convergence was only very rarely achieved in a single step from the Fréchet mean.

[1] Furthermore, it is concluded that the singularities occuring for an ambient dimension $d > 2$ does not pose a problem in practice (see [Klingenberg 2020]).

3.5.2 Conclusions

We have constructed a version of phylogenetic PCA for data taking values in a finite-dimensional Riemannian manifold, e.g. landmark shape spaces. We have argued that there is a need in evolutionary biology for doing exactly this. We have illustrated the method on a data set of landmark shapes, represented as points in the LDDMM landmark manifold. The method is based on intrinsically defined estimators, and we have discussed the relation of these to the current practice in geometric morphometrics (GM). In particular, we have shown that the root-estimator used in GM corresponds to taking a single step of a Riemannian gradient descent algorithm on Kendall's shape space. Our simulations show that one step is rarely enough, for reasonable convergence criterions.

Chapter 4

Principal subbundles for dimension reduction

Contents

4.1 Introduction

In this chapter, we present a framework for learning an unknown, lower dimensional geometry from a set of observations $\{x_1, \ldots, x_N\}$ on a Riemannian manifold $\mathcal{N}$. In the majority of our presentation we will assume that $\mathcal{N}$ is simply Euclidean space $\mathbb{R}^d$, since the framework is a contribution in this setting, and since the principles of the method then stand out more clearly. The more general formulation for manifold-valued data is presented in Section 4.5.

The framework provides concrete methods for solving the following three problems,

Figure 4.1: *Top*: noisy observations on the S-surface, embedded in $\mathbb{R}^{100}$ but projected to $\mathbb{R}^3$ for the purpose of visualization. The turquoise surface shows the 2-dimensional manifold reconstructed using the principal subbundle. *Below*: a 2D tangent space representation of the observations. Experiment described further in Section 4.6.2.

(A) *Metric learning*, i.e. learning a distance metric, $d(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, expressing the unknown underlying geometry (see [Bellet, Habrard, and Sebban 2015] for an overview).

(B) *Manifold reconstruction*, i.e. estimating a k-dimensional smooth submanifold $M \subset \mathbb{R}^d$ around which the data might be assumed to be distributed, an assumption known as the *manifold hypothesis* [Cayton 2005]. This includes surface reconstruction for observations in $\mathbb{R}^3$ [Huang et al. 2022].

(C) *Dimension reduction*, in the specific sense of learning a representation of the data in $\mathbb{R}^k$, $k < d$, that preserves various chosen local properties, e.g. pairwise distances and angles between neighbouring points. This problem is often called *manifold learning* [Ma and Fu 2012], referring to the fact that the manifold hypothesis is often assumed, although most such methods do not reconstruct the manifold in $\mathbb{R}^d$.

Each of these problems constitutes a whole field of research in itself. Indeed, their assumptions on the data can differ; while methods for (B) and (C) assume a lower dimensional structure of the data, this is not necessarily the case in (A). However, the framework described in this chapter can be used to do both (A), (B) and (C). Our basic assumption is that the data is locally linear, i.e. locally well approximated by k-dimensional affine linear subspaces. This assumption holds under the manifold hypothesis, where the tangent space at each point is a good approximation. However, the assumption may also hold even if the manifold hypothesis fails, due to the phenomenon of *non-integrability* (see Section 4.3.4). In this sense, the framework of principal subbundles relaxes the manifold assumption.

At each point in $\mathbb{R}^d$ we estimate a k-dimensional linear approximation of the data by an eigenspace of a local principal component analysis (PCA). Technically, the collection of these eigenspaces forms a *subbundle* on $\mathbb{R}^d$. In this work we exploit the fact that such a subbundle determines a *sub-Riemannian metric* on $\mathbb{R}^d$. Under such a metric a curve in $\mathbb{R}^d$ has finite length if and only if it is *horizontal*, i.e. if its velocity vector lies within the subbundle at all time points. Due to the nature of the chosen subbundle, a horizontal curve initialized within the point cloud is expected to evolve along the point cloud. Thus, our framework provides a method for *metric learning* (A) in the sense that it estimates a sub-Riemannian metric on $\mathbb{R}^d$, which, under certain assumptions, induces a distance metric on $\mathbb{R}^d$. In particular, it is a geodesic distance, meaning that the distance between $p, q \in \mathbb{R}^d$ equals the length of the shortest horizontal curve connecting p and q. A sub-Riemannian metric can be thought of as a Riemannian metric of lower rank $k \leq d$. To the best of our knowledge, the low-rank (i.e. sub-Riemannian) case has not yet been explored in Riemannian approaches to metric learning (e.g. [Hauberg, Freifeld, and Black 2012], [Perrault-Joncas and Meila 2013]). But it is exactly this property that enables the metric to also provide solutions to problems (B) and (C). It yields a method for *manifold reconstruction* (B) since the sub-Riemannian metric induces a diffeomorphism, $\phi_\mu : \mathbb{R}^k \supset U \to \phi_\mu(U) \subset \mathbb{R}^d$, whose image is a smooth k-dimensional submanifold M^k approximating the data around a chosen base point $\mu \in \mathbb{R}^d$. Technically, ϕ_μ is a restriction of the

sub-Riemannian exponential map at μ. Finally, the framework yields a method for dimension reduction (C) since $U \subset \mathbb{R}^k$ is a coordinate chart for the manifold, so that, after projection of the observations to M^k, each projected observation x_i can be represented as $\phi_\mu^{-1}(x_i) \in \mathbb{R}^k$.

Methods for manifold reconstruction (B) and dimension reduction (C) often deal with the problem of how to combine local linear approximations into a global, non-linear representation. In the field of surface reconstruction from 3D point clouds, state-of-the-art methods such as *Poisson surface reconstruction* (PSR) [Kazhdan, Bolitho, and Hoppe 2006] and *Implicit Geometric Regularization* (IGR) [Gropp et al. 2020] are based on estimation of tangent spaces, which is done via estimation of normals (see [Huang et al. 2022] for a survey and benchmarking). A fundamental obstacle to this strategy of reconstructing a submanifold from tangent space approximations, e.g. reconstructing a surface from a normal field, is that the subspaces determine a submanifold if and only if they form an *integrable* subbundle, cf. the Frobenius theorem (see Section 4.3.4 below). If the subspaces are estimated from a (finite) set of observations, integrability cannot be assumed to hold, even in the absence of noise. PSR and IGP deal with this problem by finding a surface whose normals minimize the distance to the empirical (noisy) normals. This surface is constructed by solving a Poisson equation (PSR) or by fitting a neural network (IGR). However, the approach of fitting normals does not generalize to the case of codimension greater than one, since normals are not defined in this case. Likewise, within manifold learning, methods based on alignments of local linear approximations (e.g. [Teh and Roweis 2002], [Zhang and Zha 2004], [Singer and Wu 2012], [Koelle et al. 2022], [Myhre et al. 2020]), can be thought of as different ways to deal with non-integrability. Such methods are often based on eigendecomposition of a kernel-type matrix, or other linear-algebraic computations. This strategy is useful for finding a representation in $\mathbb{R}^k$ (problem (C)) but not for reconstructing an underlying manifold (problem (B)).

The approach presented in this chapter is different. We combine the local linear approximations into a global representation by integrating a system of second-order ordinary differential equations, the sub-Riemannian geodesic equations. For $k = 1$, this integration yields the flow of the first eigenvector field, called the *principal flow* in [Panaretos, Pham, and Yao 2014]. There are, however, important differences between principal flows and our framework for $k = 1$, see the discussion in Section 4.4.3.1 and numerical results in Section 4.6.4. A follow-up work to the principal flows is the *Principal submanifolds* [Yao, Eltzner, and Pham 2023], where the aim is to leverage k eigenvectors to

construct a k-dimensional submanifold approximating the data. This method is closely related to ours, in that it is based on horizontal curves. A crucial difference, however, is that the curves in [**princSubmYaoEltzner**] are defined by an algorithmic procedure with no theoretical guarantees and the output of the method is a subset of the ambient space whose properties are largely unknown, such as whether it is in fact a submanifold.

A basic motivation and justification for our method is the following observation: if one had access to the true tangent spaces, e.g. via a frame of vector fields spanning them, then the Riemannian geodesic equation w.r.t. the corresponding Riemannian metric will generate an open subset (a normal chart) of the true manifold. I.e. it will generate an exact reconstruction, locally. When the frame is non-integrable, which is likely the case when it is estimated from data, the more general sub-Riemannian framework is needed. We show that, surprisingly, we can still generate a submanifold in this setting, and thereby give solutions to problems (B) and (C). Our framework thus offers a new way to form a global representation from local linear ones that seems natural from the point of view of differential geometry.

4.1.1 Chapter organization and contributions

Our main contribution is the idea of collecting eigenspaces of local PCA's into a tangent subbundle and showing how the induced sub-Riemannian structure can be used to model the data. In Section 4.2, we define principal subbundles on $\mathbb{R}^d$ and prove smoothness properties. In Section 4.3, we present sub-Riemannian geometry on $\mathbb{R}^d$. A large part of this section is devoted to background theory, with some exceptions, e.g. subsection 4.3.5 where we prove that a certain restriction of the sub-Riemannian exponential map is a diffeomorphism, thus generating a submanifold even if the subbundle is non-integrable. This is the crucial result showing the usefulness of sub-Riemannian geometry for manifold reconstruction. In Section 4.4, we discuss the particular sub-Riemannian geometry induced by the principal subbundle. In Section 4.5, we show how the framework generalizes to the case of observations on an a priori known Riemannian manifold. Section 4.6 presents numerical solutions to examples of problems (A) (metric learning), (B) (manifold reconstruction) and (C) (dimension reduction) for observations in $\mathbb{R}^d$ and on the sphere.

Relation of chapter to publications The material presented in this chapter was submitted in July 2023 to the journal *Foundations of computational*

mathematics and to ArXiv [Akhøj, Benn, et al. 2023]. Minor modifications has been made, mainly removal of parts describing background material that has already been presented earlier in the thesis.

4.2 Principal subbundles

In this section, we define the *principal subbundle* as a collection of eigenspaces of local PCAs. Recall that the tangent bundle on $\mathbb{R}^d$, $T\mathbb{R}^d$, can be identified with $\mathbb{R}^d \times \mathbb{R}^d$. For some subset $U \subset \mathbb{R}^d$, the tangent bundle on U, $TU \subset T\mathbb{R}^d$, can be identified with $U \times \mathbb{R}^d$. A rank k subbundle $\mathcal{D}$ of TU is a collection of k-dimensional subspaces associated to points in U, that is

$$\mathcal{D} = \{(x, v) \,|\, x \in U, v \in \mathcal{D}_x\},$$

where each $\mathcal{D}_x$ is a k-dimensional subspace of $\mathbb{R}^d$. Given a data set $\{x_i\}_{i=1..N} \subset \mathbb{R}^d$, we will define the principal subbundle as the subbundle for which each $\mathcal{D}_x$ is the span of the first k eigenvectors of a centered local PCA computed at $x \in \mathbb{R}^d$. We detail this construction below.

4.2.1 Local PCA at the local mean

Let $x_1, \ldots, x_N$ be observations in $\mathbb{R}^d$. By *local PCA* at $p \in \mathbb{R}^d$ we mean the extraction of eigenvectors of the following weighted and centered second moment.

Definition 4.2.1 (Weighted, centered first and second moments)**.** Let $K_\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ be a smooth, decaying kernel function with range parameter $\alpha > 0$. At a point $p \in \mathbb{R}^d$, the normalized weight of observation x_i is

$$w_i(p) := \frac{K_\alpha(\|x_i - p\|)}{\sum_{j=1}^N K_\alpha(\|x_i - p\|)},$$

where $\|\cdot\|$ is the standard norm on $\mathbb{R}^d$. The weighted first moment (the local mean) and the centered weighted second moment (the local covariance matrix) are then:

$$m(p) = \sum_{i=1}^N w_i(p)\, x_i \quad , \quad \Sigma_\alpha(p) := \sum_{i=1}^N w_i(m(p))(x_i - m(p))(x_i - m(p))^T \in \mathbb{R}^{d \times d}.$$

Remark 4.2.2. To save computational time, instead of using $w_i(m(p))$ in $\Sigma_\alpha(p)$ we suggest to use $w_i(p)$, i.e. not recomputing the weights at $m(p)$. This cheaper version is used for the experiments in Sections 4.6.1-4.6.3.

For K_α constantly equal to 1 (or $\alpha = \infty$), $\Sigma_\alpha(p)$ is the ordinary mean-centered covariance matrix, independent of p. In our experiments we use a gaussian kernel with standard deviation α. A motivation for using local PCA's is the following. Under the manifold hypothesis, with an underlying manifold of dimension k, the k-dimensional eigenspace of a local PCA at an observation x_i converges to the true tangent space of that submanifold at x_i in the limit of zero noise and the number of observations going to infinity (see e.g. [Singer and Wu 2012], Theorem B.1, for a convergence result).

4.2.2 Eigenvector fields and the principal subbundle

We define the principal subbundle at $p \in \mathbb{R}^d$ as a k-dimensional eigenspace of the weighted second moment at p. For it to be well-defined at p, the k'th and *k+1*'th eigenvalues of the second moment at p should be different. I.e. the subbundle is defined only outside the following set of points, which we will call singular,

$$\mathcal{S}_{\alpha,k} := \left\{ p \in \mathbb{R}^d \;\middle|\; \lambda_k(p) = \lambda_{k+1}(p) \right\}, \quad 1 \le k \le d \tag{4.2.1}$$

where $\lambda_1(p) \ge \cdots \ge \lambda_d(p)$ are the eigenvalues of $\Sigma_\alpha(p) \in \mathbb{R}^{d\times d}$.

Definition 4.2.3 (Principal subbundle)**.** Let $\lambda_1(p) \ge \cdots \ge \lambda_d(p)$ be the eigenvalues of $\Sigma_\alpha(p) \in \mathbb{R}^{d\times d}$ with associated eigenvectors $e_1(p), \ldots, e_d(p)$. Let $\mathcal{S}_{\alpha,k}$ be the set of singular points (Eq. (4.2.1)). Then the *principal subbundle* on $\mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}$ is defined as

$$\mathcal{E}^{k,\alpha} = \left\{ (p, v) \mid p \in \mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}, v \in \operatorname{span}\{e_1(p), \ldots, e_k(p)\} \right\} \subset T(\mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}).$$

Remark 4.2.4. We will assume the data and the chosen parameters satisfies that $\lambda_k(p) \neq \lambda_{k+1}(p)$ at all points where we want to evaluate the principal subbundle. In our computations we have not encountered points where this assumption was violated.

Remark 4.2.5. Cf. the proof of Proposition 4.2.7 (below), if $\lambda_k(p) > \lambda_{k+1}(p)$ at some $p \in \mathbb{R}^d$, then this property holds on an open set around p.

Note that the principal subbundle only depends on the *eigenspaces*, not the choice of eigenvectors. The latter are not uniquely determined, they depend on a choice of sign and, in the case of repeated eigenvalues, a rotation within a subspace. In order to define a sub-Riemannian structure from this subbundle it needs to be smooth, which is satisfied cf. Proposition 4.2.7 below. A closely related result, Lemma 4.2.6 below, states that if an eigenvalue λ' at $p \in \mathbb{R}^d$ has multiplicity 1, then there exists a smooth vector field on an open subset $O \subset \mathbb{R}^d$

around p which is an eigenvector for $\Sigma_\alpha(x)$ at each $x \in O$. We call this vector field an *eigenvector field*.

Lemma 4.2.6 (Existence of smooth eigenvector fields)**.** *Let e' be an eigenvector of $\Sigma_\alpha(p)$ at $p \in \mathbb{R}^d$ with eigenvalue λ' of multiplicity 1. Then there exists an open subset $O(p) \subset \mathbb{R}^d$ around p and smooth maps $e : O(p) \to \mathbb{R}^d$ and $\lambda : O(p) \to \mathbb{R}_{\geq 0}$ satisfying $e(p) = e', \lambda(p) = \lambda'$, $\|e(x)\| = 1$ and $\Sigma_\alpha(x)e(x) = \lambda(x)e(x)$ for all $x \in O(p)$.*

This result follows directly from [J.-G. Sun 1985], Theorem 2.3, since Σ_α is a smooth map. From this result on eigen*vectors*, one can conclude that the eigen*spaces* are smooth at p if either the eigenvalues $\lambda_1(p), \ldots, \lambda_{k+1}(p)$ are distinct, or $\lambda_k(p), \ldots, \lambda_d(p)$ are distinct at $p \in \mathbb{R}^d$. However, we can in fact show smoothness of the subbundle under the milder, indeed minimal, condition that $\lambda_k(p) > \lambda_{k+1}(p)$ (Proposition 4.2.7). Appendix B.1 contains the proof of this and all other results in the chapter.

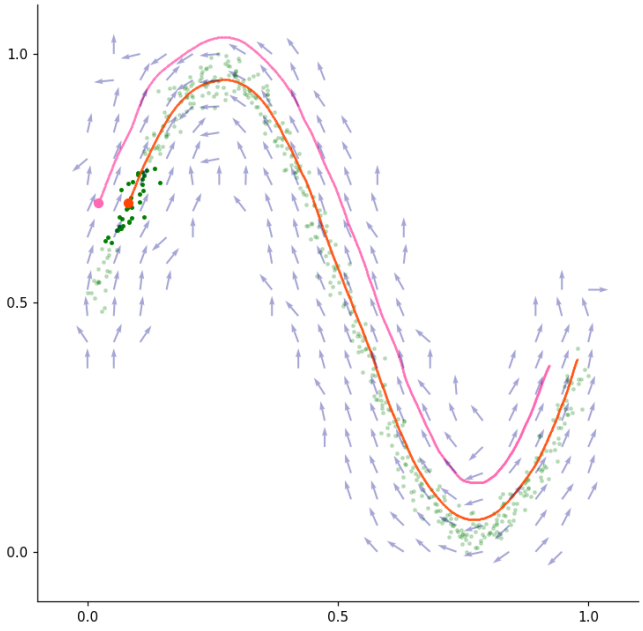
Proposition 4.2.7. *The principal subbundle, defined on $\mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}$, is smooth.*

Figure 4.2 illustrates the principal subbundle (blue arrows) induced by point clouds in $\mathbb{R}^2$ and $\mathbb{R}^3$, including the effect of centering the second moment at the local mean.

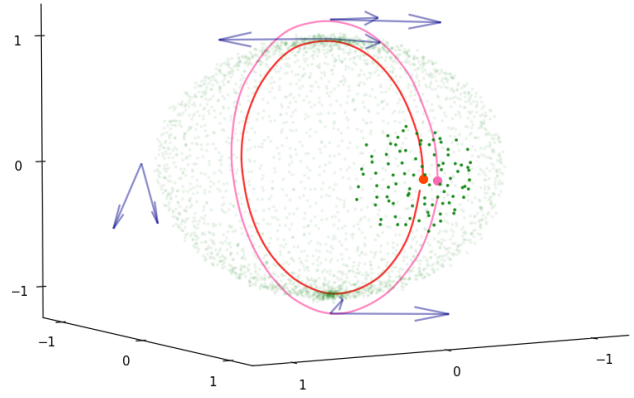
We are interested in studying curves whose velocity vectors are constrained to lie in the principal subbundle (i.e. eigenspaces of local PCA's). This can be done using sub-Riemannian geometry, which we introduce next.

4.3 Sub-Riemannian geometry

We now introduce basic notions of sub-Riemannian geometry on $\mathbb{R}^d$. We focus on the special case that we need, where the sub-Riemannian metric is a restriction of the standard Euclidean inner product. This viewpoint is not presented in sources that we know of, so we devote some space to it. For more comprehensive introductions see e.g. [Agrachev, Barilari, and Boscain 2019] or [Jean 2014]. We strive to make the presentation accessible to someone with only a slight knowledge of differential geometry. Most of the sub-Riemannian computations described in this section has been implemented by the authors in the software library geomstats, for public availability.



(a)



(b)

Figure 4.2: Two illustrations of a principal subbundle induced by point clouds (green points) and sub-Riemannian geodesics (red and pink curves). On Figure *(a)* the subbundle rank is $k = 1$, on *(b)* it is $k = 2$. Blue arrows span the principal subbundle subspace at the basepoint of the arrows; on Figure *a)* each subspace is a line (spanned by one arrow), on Figure *(b)* they are planes (spanned by two arrows). The geodesics are initialized at the red, respectively pink, dots, which are inside, respectively outside, the point clouds. The observations colored in a darker green are those with an assigned normalized weight w_i, w.r.t. the position of the red dot, larger than 10^{-5} - thus the lighter green observations has only a negligible effect on the local PCA computation. On Figure *(b)*, the duration of integration, and thus the curve length, is $T = 2\pi$ (red geodesic) and $T = 2.3\pi$ (pink geodesic), respectively - these are the circumferences of great circles centered at the origin and passing through the initial points.

4.3.1 Horizontal curves and the sub-Riemannian distance

In the special case that we consider, a sub-Riemannian structure on $\mathbb{R}^d$ is fully determined by a rank k subbundle $\mathcal{D} \subset T\mathbb{R}^d$. The subbundle can be represented as a smoothly varying orthogonal projection matrix,

$$g^\star : \mathbb{R}^d \to \mathbb{R}^{d\times d} : p \mapsto F(p)F^T(p), \tag{4.3.1}$$

where $F : \mathbb{R}^d \to \mathbb{R}^{d\times k}$ is a smooth map s.t. $F(p)$ is a rank k matrix whose columns form an orthonormal basis for $\mathcal{D}_p$ at any $p \in \mathbb{R}^d$. The map $g^\star$ is called the *cometric*. If $g^\star(p)$ has full rank d at every $p \in \mathbb{R}^d$, then the map $p \mapsto g^\star(p)^{-1}$ is called a Riemannian metric. We discuss relations between Riemannian and sub-Riemannian geometries below.

A basic intuition behind sub-Riemannian geometry is that, at each point $p \in \mathbb{R}^d$, $\mathcal{D}_p$ contains the allowed velocity vectors of a curve passing through p. If a curve $\gamma : [0, 1] \to \mathbb{R}^d$ satisfies

$$\frac{d}{dt}\gamma(t) \coloneqq \dot\gamma(t) \in \mathcal{D}_{\gamma(t)}$$

for almost all $t \in [0, 1]$ it is called *horizontal*. This class of curves induces a *distance metric* on $\mathbb{R}^d$, the *Carnot-Carathéodory metric*,

$$d^{\mathcal{D}}(p, q) = \inf\left\{ L(\gamma) \;\middle|\; \begin{array}{c} \gamma : [0, 1] \to \mathbb{R}^d \text{ is horizontal} \\ \gamma(0) = p, \gamma(1) = q \end{array} \right\} \in \mathbb{R}_{\geq 0} \cup \{\infty\}, \tag{4.3.2}$$

for any $p, q \in \mathbb{R}^d$, where $L(\gamma) \coloneqq \int_0^1 \|\dot\gamma(t)\|dt$ is the curve length functional. An important property of a sub-Riemannian geometry is whether any two points p, q can be connected by a horizontal curve, or, equivalently, whether $d(p, q)$ is finite for all $p, q \in \mathbb{R}^d$. A sufficient condition for this is that $\mathcal{D}$ is *bracket-generating* (cf. the Chow-Rashevski theorem, [Chow 2002], [Agrachev, Barilari, and Boscain 2019]). This means that, for all $p \in \mathbb{R}^d$, $\operatorname{Lie}\mathcal{D}_p$ equals $\mathbb{R}^d$, where $\operatorname{Lie}\mathcal{D}_p$ consists of the span of all $\mathcal{D}$-valued vector fields and all of their iterated Lie brackets (see e.g. [Lee 2013]). In this case, $d^{\mathcal{D}}$ induces the standard topology on $\mathbb{R}^d$.

4.3.2 Sub-Riemannian geodesics

We now turn to horizontal curves that are 'locally length-minimizing', i.e. any local perturbation of the curve increases its length. For our purposes, the most important class of such curves is called *normal sub-Riemannian geodesics*.

Normal geodesics are solutions to a system of equations on the cotangent bundle $T^\star\mathbb{R}^d$, which, in our setting, can be identified with $\mathbb{R}^d \times \mathbb{R}^d$. A curve

$\gamma : [0, T] \to \mathbb{R}^d$ is a normal geodesic if and only if it is the projection to $\mathbb{R}^d$ of a curve in $T^\star\mathbb{R}^d$, $\psi : [0, 1] \to T^\star\mathbb{R}^d$, that satisfies the *sub-Riemannian Hamiltonian equations*. Let H denote the *sub-Riemannian Hamiltonian*,

$$H : T^\star\mathbb{R}^d \to \mathbb{R}_{\geq 0} : (p, \eta) \mapsto \frac{1}{2}\eta^T g_p^\star \eta.$$

We will write H_p if we consider it as a function on $T_p^\star\mathbb{R}^d$ only. The Hamiltonian equations are then given by

$$\begin{aligned} \dot{p} &= \frac{\partial H}{\partial \eta}(p, \eta) = g_p^\star \eta, \\ \dot{\eta} &= -\frac{\partial H}{\partial p}(p, \eta). \end{aligned} \tag{4.3.3}$$

A solution $\psi(t) := (p(t), \eta(t))$ with initial value (p_0, η_0) is called a *normal extremal*. The associated normal geodesic is the curve $\gamma_{p_0}^{\eta_0}(t) := \pi(p(t), \eta(t)) := p(t)$, i.e. the projection of ψ to the first component $\mathbb{R}^d$. Notice that the horizontality of γ is apparent from the fact that $g_p^\star$ projects η to $\mathcal{D}_p$ in (4.3.3). In the Riemannian case the Hamiltonian equations are equivalent to a system of ODE's on the tangent bundle called the *geodesic equations*. This parameterizes geodesics by their initial tangent vector instead of, as in the sub-Riemannian case, the initial cotangent vector. We end this section with a few facts about solutions to Hamilton's equations that we will need later on. Firstly, the Hamiltonian is conserved along solutions, i.e. $H(p_t, \eta_t) = H(p_0, \eta_0)$ for all $t \in [0, T]$ (see e.g. [Agrachev, Barilari, and Boscain 2019], Section 4.2.1). This implies that a normal geodesic γ is a constant speed curve, since

$$\|\dot{\gamma}(t)\| = \|g_{p_t}^\star \eta_t\| = \sqrt{2H(p_t, \eta_t)}. \tag{4.3.4}$$

This further implies that $\gamma_{p_0}^{\eta_0}$ has unit speed if $\eta_0 \in H_{p_0}^{-1}(1/2)$, and therefore that its length is given by the duration of integration T. Lastly, we will need the fact that the Hamiltonian equations are time-homogenous in the sense that, for any $\eta_0 \in H^{-1}(1/2)$ and $\alpha > 0$, $\gamma_{p_0}^{\alpha\eta_0}(t) = \gamma_{p_0}^{\eta_0}(\alpha t)$ ([Agrachev, Barilari, and Boscain 2019], Section 8.6).

4.3.3 The sub-Riemannian exponential and logarithm

The sub-Riemannian *exponential* map at $p \in \mathbb{R}^d$ maps a cotangent $\eta \in T_p^\star\mathbb{R}^d \cong \mathbb{R}^d$ to the position at time 1 of the normal geodesic initialized by (p, η), i.e.

$$\exp_p^{\mathcal{D}} \quad : \quad T_p^\star\mathbb{R}^d \to \mathbb{R}^d \quad : \quad \eta \mapsto \exp_p^{\mathcal{D}}(\eta) := \gamma_p^\eta(1).$$

The exponential map will also be denoted simply by exp. The time-homogeneity of the Hamiltonian equations mentioned in the previous section has two important consequences. Firstly, for $\alpha > 0$, it holds that $\exp_p(\alpha\eta) = \gamma_p^\eta(\alpha)$, so scaling η amounts to moving along a single normal geodesic; secondly, γ can be assumed to be unit speed parameterized, and therefore the length of the normal geodesic $\alpha \mapsto \exp_p(\alpha\eta)$, $\alpha \in [0,1]$, is given by $\sqrt{2H(\eta)}$. In the case where this normal geodesic is a global, not just local, length minimizer between its endpoints p and $y \coloneqq \exp_p(\eta)$, we get the formula

$$d^{\mathcal{D}}(p,y) = \sqrt{2H(p,\eta)}. \tag{4.3.5}$$

4.3.3.1 Optimizing for the logarithm

To compute the sub-Riemannian distance between two points, eq. (4.3.5) suggests that one should invert the exponential map. If the exponential map at p is a diffeomorphism (thus invertible) around $0 \in T_p^\star\mathbb{R}^d$, its inverse is called the *logarithmic* map, defined by

$$\log_p^{\mathcal{D}} : \mathbb{R}^d \supset U \to O \subset T_p^\star\mathbb{R}^d \quad \text{satisfying} \quad \gamma_p^{\log_p^{\mathcal{D}}(y)}(1) = y$$

for some open sets U and O with $p \in U$. However, such an open set U on which $\exp_p$ is a diffeomorphism only exists if $rank(\mathcal{D}) = d$ (see [Agrachev, Barilari, and Boscain 2019] Prop. 8.40), in which case the geometry is Riemannian. A simple way to see this is that $\exp_p(H_p^{-1}(0)) = 0$, where $H_p^{-1}(0) = \mathcal{D}_p^\perp$. In the sub-Riemannian case of $rank(\mathcal{D}) < d$ we propose an approximate log map given as a solution to the following optimization problem, for $p, y \in \mathbb{R}^d$,

$$\overline{\log}_p(y) \in \operatorname*{argmin}_{\eta \in \mathbb{A}} \| \exp_p(\eta) - y \|^2 + H(p,\eta), \tag{4.3.6}$$

where $\mathbb{A} = T_p^\star\mathbb{R}^d$. This problem searches for the shortest normal geodesic between p and y. For reasons that will be explained in Section 4.4.4, we will also be interested in the case of $\mathbb{A} = \mathcal{D}_p^\star \subset T_p^\star\mathbb{R}^d$, the metric dual of $\mathcal{D}_p$ (Equation 4.3.7 below). Under certain assumptions on $\mathcal{D}$, notably bracket-generatingness, the image set $\exp_p(T^\star\mathbb{R}^d)$ is dense in $\mathbb{R}^d$ even when $\operatorname{rank}\mathcal{D} < d$ [Rifford 2014], implying that the error $\| \exp_p(\eta) - y \|^2$ in (4.3.6) can be made arbitrarily small. The problem of finding shortest horizontal curves between points is studied in non-holonomic control theory (see e.g. [Jean 2014]). In our current implementations, however, we find (local) solutions via a minimization algorithm based on BFGS [Wright, Nocedal, et al. 1999] and automatic differentation of the exponential map, which is possible using e.g. the python library Jax [Frostig, Johnson, and Leary 2018].

4.3.4 The subbundle induces a foliation

If a bracket generating subbundle $\mathcal{D}$ (i.e. $\operatorname{Lie}\mathcal{D} = T\mathbb{R}^d$) represents one extreme for subbundles on $\mathbb{R}^d$ then its opposite is that of a constant rank integrable subbundle; that is, a constant rank subbundle $\tilde{\mathcal{D}}$ satisfying $\operatorname{Lie}\tilde{\mathcal{D}} = \tilde{\mathcal{D}}$. An important property of integrable subbundles is that they posses integral manifolds which are immersed submanifolds $\mathcal{M} \subset \mathbb{R}^d$ such that $T_p\mathcal{M} = \tilde{\mathcal{D}}_p$ for all points $p \in \mathcal{M}$. Given a constant rank integrable subbundle $\tilde{\mathcal{D}}$, the global Frobenius Theorem tells us that $\mathbb{R}^d$ is foliated, or partitioned, by the collection of all maximal integral manifolds of $\tilde{\mathcal{D}}$ - each integral manifold is called a leaf and has dimension equal to the rank of $\tilde{\mathcal{D}}$ (see Lee, Chapter 19 for full details on integrable subbundles, there called *involutive distributions*, and the Frobenius Theorem). The geometry induced by $\tilde{\mathcal{D}}$ on $\mathcal{M} \subset \mathbb{R}^d$ is Riemannian since $\tilde{\mathcal{D}}_p$ is the full tangent space at each point $p \in \mathcal{M}$, implying that all curves on $\mathcal{M}$ are horizontal; therefore the sub-Riemannian geodesic equations are identical to the Riemannian geodesic equations of the submanifold. If a subbundle $\breve{\mathcal{D}}$ is neither bracket generating ($\operatorname{Lie}\breve{\mathcal{D}} = T\mathbb{R}^d$) nor integrable ($\operatorname{Lie}\breve{\mathcal{D}} = \breve{\mathcal{D}}$) then the subbundle $\operatorname{Lie}\breve{\mathcal{D}} \subset T\mathbb{R}^d$ is integrable and foliates $\mathbb{R}^d$ by its integral manifolds $\mathcal{M}$, each of dimension $\operatorname{rank}(\operatorname{Lie}\breve{\mathcal{D}})$. The induced geometry on each integral manifold $\mathcal{M}$ is sub-Riemannian (not all curves are horizontal).

In relation to problem A, mentioned in the introduction, the previous discussion implies that the induced distance metric is finite, $d^{\mathcal{D}}(p,q) < \infty$, for all points p, q in the same leaf, whereas it is infinite for points belonging to different leaves - a horizontal curve is constrained to move within a single leaf. In relation to problem B, we are interested in generating a k-dimensional submanifold of $\mathbb{R}^d$ from a rank k subbundle $\mathcal{D}$ whose integrability or bracket generation is a priori unknown. In Proposition 3.1 below we show how this can be done via sub-Riemannian geometry. The generated submanifold is tangent to $\mathcal{D}$ in 'radial' directions, but not in all directions, as will be explained below.

4.3.5 The exponential image of the dual subbundle

The content of the previous sections implies the following. If $\mathcal{D}$ is integrable, then there exists an open set $U \subset \mathcal{D}_p$ s.t. $M := \exp_p(U)$ is a k-dimensional embedded submanifold of $\mathbb{R}^d$ whose tangent space a every point $q \in M$ equals $\mathcal{D}_q$. In this case, $\exp_p$ is a diffeomorphism from U to this submanifold. On the other hand, if $\mathcal{D}$ is not integrable, there exists no submanifold that is tangent to $\mathcal{D}$, in particular $\exp_p(U)$ does not satisfy this. However, in the following we show that $\exp_p(U)$ is still a k-dimensional embedded submanifold.

Let

$$\mathcal{D}_p^\star := \{\langle v, \cdot\rangle | v \in \mathcal{D}_p\} \subset T_p^\star \mathbb{R}^d \tag{4.3.7}$$

be the dual space of $\mathcal{D}_p$ w.r.t. the standard inner product $\langle v, u\rangle := v^T u$. This simply means that $\mathcal{D}_p^\star$ consists of the tangent vectors (column vectors) in $\mathcal{D}_p$ considered as covectors (row vectors). Thus, $\mathcal{D}_p^\star$ is a k dimensional subspace of $T_p^\star \mathbb{R}^d$ which can be identified with $\mathcal{D}_p \subset T\mathbb{R}^d$.

Proposition 4.3.1 ($\exp_\mu$ is a local diffeomorphism from $\mathcal{D}_\mu^\star$)**.** *Let $\mu \in \mathbb{R}^d$ be arbitrary. There exists an open subset $C_\mu \subset \mathcal{D}_\mu^\star$ containing 0 such that $\exp_\mu^{\mathcal{D}}$ restricted to C_μ is a diffeomorphism onto its image. That is,*

$$M_\mu^{\mathcal{D}} := \exp_\mu^{\mathcal{D}}(C_\mu) \subset \mathbb{R}^d$$

is a smooth k-dimensional embedded submanifold of $\mathbb{R}^d$ containing μ.

We postpone the proof to Section 4.5, where Proposition 4.5.1 generalizes the statement to the setting where the ambient space is a Riemannian manifold.

It holds that $T_p(M_\mu^{\mathcal{D}}) = \mathcal{D}_p$ at $p = \mu$, but at a general $p \in M_\mu^{\mathcal{D}}$ these spaces are different if $\mathcal{D}$ is not integrable. They need not even be 'close', as can be seen in e.g. the Heisenberg group where $\exp_0^{\mathcal{D}}(C_0)$ is the xy-plane, to which the Heisenberg subbundle is almost orthogonal at certain points p. But $M_\mu^{\mathcal{D}}$ is 'radially horizontal', in the sense that it is the union of normal geodesics from μ each of which is horizontal w.r.t. $\mathcal{D}$. In particular, if we assume that C_μ is convex and let $\partial C_\mu \subset \mathcal{D}_p^\star$ denote its boundary, then

$$\exp_\mu^{\mathcal{D}}(C_\mu) = \left\{\gamma_p^\eta(t) \;\middle|\; \eta \in \partial C_\mu,\ t \in [0,1]\right\}, \tag{4.3.8}$$

where each geodesic $t \mapsto \gamma_p^\eta(t)$ is tangent to $\mathcal{D}$.

Note that, since the exponential map restricted to C_p is a diffeomorphism, the log-optimization problem (4.3.6) with $\mathbb{A} = \mathcal{D}_p^\star$ has a unique solution for $p = \mu$ and any $y \in M_\mu^{\mathcal{D}}$.

4.4 Sub-Riemannian geometry of the principal subbundle

In this section, we present a sub-Riemannian (SR) structure on $\mathbb{R}^d$ based on local PCA's, namely, the SR structure determined by the principal subbundle. Moving horizontally with respect to the principal subbundle means to move within a k-dimensional subspace of maximum local variation at each step. Therefore, geodesics that are horizontal w.r.t. this structure follow the point cloud, and the associated *exp* and *log* maps can be used for representing the data. The image of the dual subbundle under the exponential map, described in Proposition 4.3.1

above, will be called a *principal submanifold* when the principal subbundle is used. Such a submanifold approximates the data for well-chosen hyperparameters. This is described in Section 4.4.3 where we also give an algorithm to compute it. Furthermore, we discuss the use of the log optimization problem (4.3.6) for giving a representation of the observations in $\mathbb{R}^k$ (Section 4.4.4) and for computing distances between observations (Section 4.4.5).

4.4.1 Properties of the sub-Riemannian structure

The sub-Riemannian structure that we will use to model the data is the one determined by the principal subbundle $\mathcal{E}^{k,\alpha}$, also denoted simply by $\mathcal{E}$. Proposition 4.2.7 about smoothness of the subbundle implies smoothness of the cometric $g^\star$. For any $p \in \mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}$ the cometric can be represented as $g^\star_p = F(p)F(p)^T \in \mathbb{R}^{d\times d}$, where $F = [e_1(p), \ldots, e_k(p)]$ is a matrix whose columns are the first k eigenvectors of the weighted second moment $\Sigma_\alpha(p)$ (Definition 4.2.1).

We know that $\mathcal{E}$ is of constant rank k, but we do not know if $\operatorname{Lie}\mathcal{E}$ is of constant rank, let alone if it is bracket-generating (i.e. $\operatorname{rank}(\operatorname{Lie}\mathcal{E}) = d$). Under the manifold hypothesis, in the limit of zero noise and the number of observations going to infinity, the convergence result of [Singer and Wu 2012] (Theorem B.1) suggests that the subbundle is everywhere tangent to a submanifold and thus integrable.

4.4.2 Computing geodesics

We compute geodesics w.r.t. the chosen sub-Riemannian structure by numerically integrating the sub-Riemannian Hamiltonian equations (4.3.3), see Appendix B.2 for notes on the implementation and approximations allowing speed-up. In [J.-G. Sun 1985], Theorem 2.4, formulas are given for the derivatives of eigenvector fields. This enables computation of derivatives of the Hamiltonian,

$$\begin{aligned} H(p,\eta) &= \frac{1}{2}\eta^T g^\star_p \eta \\ &= \frac{1}{2}\eta^T F(p)F(p)^T\eta \\ &= \frac{1}{2}\eta^T [e_1(p),\ldots,e_k(p)][e_1(p),\ldots,e_k(p)]^T\eta, \end{aligned}$$

via automatic differentiation libraries such as Jax [Frostig, Johnson, and Leary 2018]. The formulas in [J.-G. Sun 1985] hold under the assumption that the first $k+1$ eigenvalues, $\lambda_1(p), \ldots, \lambda_{k+1}(p)$, are distinct (cf. Lemma 4.2.6). Note that

our basic assumption on the observations is that they are well approximated locally by a k-dimensional linear space, implying that the first k eigenvalues are relatively close, possibly equal. Two comments on this: *1.* Using the results in [J.-g. Sun 1990] (see also Proposition 4.2.7 and its proof), it is possible to compute derivatives of the Hamiltonian under the milder assumption of only $\lambda_k(p)$ and $\lambda_{k+1}(p)$ being distinct - however, in practice we have not had the need to pursue this. *2.* Since the differences between $\lambda_1, \ldots, \lambda_k$ are likely to be relatively small, the ordering and rotation of the eigenvectors is effectively random. However, this does not affect the Hamiltonian equations, since the Hamiltonian depends only on the cometric, a projection matrix, which is invariant to rotations and permutations of the basis $F(p)$ within $\mathcal{E}_p$.

Figure 4.2 illustrates sub-Riemannian geodesics with respect to the metric induced by two different point clouds. The surfaces (principal submanifolds) presented in figures 4.1 and 4.5 are likewise composed of many such geodesics, cf. the next section.

4.4.3 Principal submanifolds (Problem *B*)

As the first use of principal subbundles, we define the *principal submanifold* from a base point $\mu \in \mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}$, given a set of observations in $\mathbb{R}^d$. This choice of data representation implicitly assumes that the data is locally well-described by a submanifold, i.e. the 'manifold hypothesis'.

Definition 4.4.1 (Principal submanifold at μ)**.** Let $\{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ be a set of observations. Let $\mu \in \mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}$ be a chosen base point, let α be the kernel range and let $k \in \{1, \ldots, d-1\}$ be the rank of the principal subbundle, $\mathcal{E} = \mathcal{E}^{k,\alpha} \subset T\mathbb{R}^d$. Let $\mathcal{E}^\star_\mu$ be the dual subbundle at μ, and $B_r \subset \mathcal{E}^\star_\mu$ a k-dimensional open ball of radius r containing 0. The *principal submanifold* of radius r is given by

$$M^k_\mu(r) \coloneqq \exp^{\mathcal{E}}_\mu(B_r) \subset \mathbb{R}^d, \tag{4.4.1}$$

Remark 4.4.2. We will assume that r is sufficiently small for $M^k_\mu(r)$ to actually be a submanifold, cf. Proposition 4.3.1. If we write simply M^k_μ, we will assume that r takes the largest such value.

Algorithm 4 describes how to compute a point set representation of a principal submanifold, up to arbitrary resolution. For hyperparameters, μ, k, α (the base point, dimension and range, respectively), the principal submanifold, M^k_μ, is an estimate of the true underlying submanifold, M, locally around μ. As described in Section 4.3.3, M^k_μ cannot be expected to be exactly tangent to $\mathcal{E}$ since $\mathcal{E}$ might not be integrable. However, since $\mathcal{E}^{k,\alpha}$ approximates the tangent spaces

of the true submanifold our expectation is that the subbundle is 'close' to being integrable and therefore that the difference between $\mathcal{E}_p$ and $T_p(M_\mu^k)$ is small for $p \in M_\mu^k$. The approximation $M_\mu^k \approx M$ comes with the following guarantee: if $\mu \in M$ and the principal subbundle contains the true tangent spaces to M around μ, then the principal submanifold is an open subset of the true submanifold M. In particular, the ball $B_r \subset \mathcal{E}_\mu^\star \subset T_p^\star M \cong \mathbb{R}^k$ is a (normal) coordinate chart for M. Figure 4.3 illustrates the effect of noise on the geodesics, and therefore on the principal submanifold, for points distributed around the unit sphere. In the noiseless case, Figure 4.3 *a)*, the computed geodesic paths are identical to the exact Riemannian geodesics on the sphere, up to numerical error, and the resulting principal submanifold is thus identical to the sphere (the mean norm of each generated point is 0.9992 with standard deviation 0.0014). In Figure 4.3 *b)* the observations on the sphere have been added isotropic Gaussian noise in $\mathbb{R}^3$ with marginal standard deviation $\sigma = 0.1$. In this case the geodesics still evolve very close to the sphere (the mean norm of each generated point is 1.0299 with standard deviation 0.0162), but they start to cross after some integration steps, so that the manifold property of $M_\mu^k(r)$ seems to hold for a smaller value of the radius r compared to the noiseless case.

4.4.3.1 Relation to principal flows

We end this subsection with a discussion on the relation between a principal submanifold for $k = 1$ and the principal flow, described in [Panaretos, Pham, and Yao 2014]. For $k = 1$, integrating the Hamiltonian equations (4.3.3) yields the flow of the first eigenvector field e_1 starting from μ. This is called the *principal flow* in [Panaretos, Pham, and Yao 2014], but the methods differ in important ways. Firstly, the principal flow at p is based on a second moment which is centered around p, not at the local mean around p. The span of the first eigenvector of such an uncentered second moment will be 'orthogonal' to the point cloud when evaluated at points outside of it. This causes the principal flow to stray away from the observations if it reaches such a point. As opposed to this, the first eigenvector of the centered second moment stays tangential to the point cloud when evaluated outside of it, as illustrated by the pink curve in Figure 4.2 *a)*. This behaviour arguably makes it more stable, see simulation results in section 4.6.4 and Figure 4.6. Secondly, to handle the fact that eigenvectors are determined only up to their sign, the principal flow is computed by solving a variational problem and integrating an associated system of ODE's. This system of ODE's has to be integrated for a range of candidate values of a Lagrange multiplier, in the end choosing the value for which the corresponding curve

minimizes an energy functional. As opposed to this, we formulate the problem as a Hamiltonian system of ODE's which is invariant to the sign of the vector field (only the corresponding rank 1 subbundle matters), removing the need for the variational formulation and the ODE integration for multiple values of a Lagrange multiplier. It is this reformulation of principal flows as solutions to a set of geodesic (Hamiltonian) equations that also allows us to generalize the concept to higher dimensions.

4.4.3.2 Projection to a principal submanifold

An observation $x_i \in \mathbb{R}^d$ can be projected to M_μ^k by

$$\pi_{M_\mu^k}(x_i) = exp_\mu^{\mathcal{E}}(\overline{\log}_\mu(x_i))$$

where $\overline{\log}_\mu(x_i)$ is a solution to (4.3.6) with search space $\mathbb{A} = \mathcal{E}_\mu$. Alternatively, given a discrete representation $\widetilde{M_\mu^k}$ of M_μ^k, computed using Algorithm 4, one can use the discrete projection $\pi_{\widetilde{M_\mu^k}}(x_i) := \operatorname{argmin}_{p \in \widetilde{M_\mu^k}} \|x_i - p\|$, which can be solved numerically as a Euclidean 1-nearest neighbours problem.

Algorithm 4 Point set representation of a principal submanifold

Input:

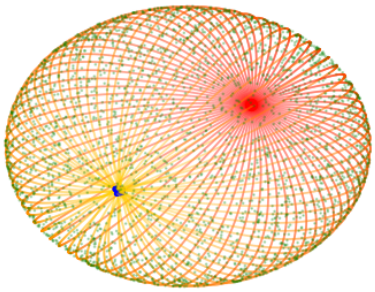
- *Geometric parameters:* kernel range $\alpha \in (0, \infty)$, submanifold dimension $k \in \{1, \dots, d-1\}$, base point $\mu \in \mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}$, radius $r > 0$.
- *Numerical parameters:* a number of geodesics $L \in \mathbb{N}$, the stepsize $\Delta > 0$.

Output:

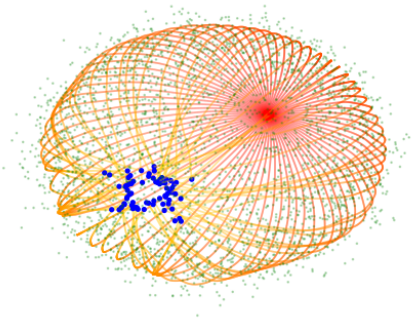
$sL+1$ points in $M_\mu^{\mathcal{E}^{\alpha,k}}(r) \subset \mathbb{R}^d$, where $s = \lfloor r/\Delta \rfloor$ is the number of integration steps.

1: *Initialization*: Generate L cotangents η_i on the k-dimensional unit sphere, $\eta_i \in \mathcal{S}^k \subset \left(\mathcal{E}^{k,\alpha}\right)^\star \cong \mathbb{R}^k, i = 1 \dots L$.

2: **for** $i = 1$ **to** L **do**
3: Integrate Hamiltonian equations (4.3.3) with initial condition (μ, η_i) over $s = \lfloor r/\Delta \rfloor$ steps of stepsize Δ.
4: Store the points along the trajectory; $p_{ij} = \exp_\mu^{\mathcal{E}}(j\Delta\eta_i), j = 1 \dots s$.
5: **end for**
6: **return** Points $\left\{p_{ij} = \exp_\mu^{\mathcal{E}}(j\Delta\eta_i) \middle| i = 1 \dots N, j = 1 \dots s\right\}$



(a)



(b)

Figure 4.3: Two illustrations of geodesics w.r.t. the sub-Riemannian metric induced by points clouds distributed around the unit sphere. On Figure *a)* 2000 points (green points) x_i are sampled uniformly on the sphere. On Figure *b)* noise is added to the observations, which are now sampled from $y_i \cong N(x_i, I_3 \cdot \sigma)$, i.e. isotropic Gaussian distributions with marginal standard deviation $\sigma = 0.1$ (green points). On each figure, 75 geodesics with initial cotangents on a grid in the dual subbundle at the basepoint $\mu = (0, -1, 0)$ are generated (these are the curves with a color gradient from red to yellow). The duration, and thus length of each geodesic is $T = \pi$, which theoretically corresponds to half a round on the unit sphere. The endpoint of each geodesic is marked by the blue dots. For points on the geodesics on Figure *a)*, the mean norm is 0.9992 with standard deviation 0.0014, thus the geodesics stay close to the true submanifold (the sphere). For points on the geodesics on Figure *b)*, the mean norm is 1.0299 with standard deviation 0.0162 - thus the geodesics still stay close to the sphere, but they now deviate somewhat from great arcs, as illustrated by the endpoints not being exactly at the opposite pole.

4.4.4 Representation of observations in *k*-dimensional Euclidean space (Problem *C*)

The ball $B_r \subset \mathcal{E}_\mu^\star \cong \mathbb{R}^k$ forms a coordinate chart for the principal submanifold, i.e. any point $p \in M_\mu^k(r)$ can be represented as $\bar{p} := \exp_\mu^{-1}(p) \in \mathbb{R}^k$. It behaves like a normal chart, in the sense that the SR distance between the base point μ and $p \in M_\mu^k$ is preserved, $d^{\mathcal{E}}(\mu, p) = \|\bar{p}\|$, while the distances between arbitrary points $p, q \in M_\mu^k$ are distorted in a way that depends on the curvature of $M_\mu^{\mathcal{E}}$. If $\{x_1, \ldots, x_N\}$ are observations distributed around M_μ^k, then the projections $\pi_{M_\mu^k}(x_i) \in M_\mu^k, i = 1 \ldots N$, can be represented in this chart by solving the log problem (4.3.6) with $\mathbb{A} = \mathcal{E}_\mu$, yielding lower dimensional representations $\overline{\pi_{M_\mu^k}(x_i)} := \overline{log}_\mu(\pi_{M_\mu^k}(x_i)) \in \mathbb{R}^k, i = 1..N$. Computing this is less complex than it looks; in fact, solving the projection problem (either the continuous or the discrete version, c.f. Section 4.4.3.2) already involves solving the log-problem, so computing a projection also yields the representation in $\mathbb{R}^k$. See Figure 4.1 and Section 4.6.2 describing a 2D representation of the S-surface embedded in $\mathbb{R}^{100}$.

4.4.5 Computing the SR distance between points (Problem *A*)

As discussed in Section 4.3.3, we can combine Equations (4.3.5) and (4.3.6) to approximate the SR distance between two points $x, y \in \mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}$ by

$$d^{\mathcal{E}}(x, y) \approx \sqrt{2H(\overline{\log}_x(y))},$$

with log search space $\mathbb{A} = T_\mu^\star \mathbb{R}^d$. As mentioned, we cannot expect to find the exact SR distance, i.e. the length of the globally shortest curve joining x and y, even in the case of a bracket-generating subbundle for which $d^{\mathcal{E}}$ is in fact finite for all x, y. When the points are observations, i.e. $x, y \in \{x_i\}_{i=1,\ldots,N}$, this might not be desirable either since the error in the log minimization problem (4.3.6) can be interpreted as an effect of random noise. In practice, we have found it useful to approximate the log by the *discrete log*, presented in Appendix A.1.1.

See Section 4.6.3 for a numerical evaluation of estimated distances $d^{\mathcal{E}}$ based on a dataset in $\mathbb{R}^{50}$. Figure 4.4 illustrates a computation of $\overline{\log}_x(y)$ based on a dataset distributed around the S-surface. The base point, x, is the blue dot and the target point, y, is the pink dot. The red curve is the geodesic $t \mapsto \exp_x^{\mathcal{E}}(t \cdot \overline{\log}_x(y)), t \in [0, 1]$, the length of which constitutes our estimate of the distance between x and y. As expected, the endpoint $\exp_p^{\mathcal{E}}(\overline{\log}_x(y))$ doesn't match y exactly. On Figure 4.5, the color gradient and concentric circles on the face illustrate the SR distance to the base point on the nose.

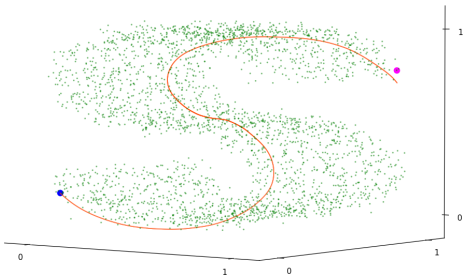


Figure 4.4: Illustration of a computation of $\overline{\log}_x(y)$ based on observations (green dots) distributed around the S-surface. The base point, x, is the blue dot and the target point, y, is the pink dot. The red curve is the geodesic $t \mapsto \exp_x^{\mathcal{E}}(t \cdot \overline{\log}_x(y)), t \in [0, 1]$.

4.4.6 Hyperparameters

The kernel range α and the dimension k are hyperparameters that are common to many methods and there is a significant body of literature about how to select them. See Appendix B.3 for our comments and references. Regarding the base point $\mu \in \mathbb{R}^d$ of a principal submanifold, we suggest to use a local mean around a well-chosen observation x_0. Which particular x_0 will be application specific, but a general purpose option is a within-sample *Fréchet mean*,

$$\hat{\mu} \in \underset{\mu \in \{x_i\}_{i=1..N}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} d(\mu, x_i),$$

where d is either the Euclidean distance or $d^{\mathcal{E}}$ of the principal subbundle.

4.5 Generalization to observations on a Riemannian manifold

In this section, we generalize the framework of principal subbundles to the setting where the observations are points on an a priori known Riemannian manifold. A numerical application of the method to such data is presented in Section 4.6.4. These two sections assume a deeper knowledge of differential geometry than elsewhere, but they can be skipped without loss of continuity by the reader who wish to focus on the case of Euclidean valued data. It turns out that the formulation of principal subbundles for Euclidean valued data, given above, is based only on operations that generalize naturally to the setting of manifold valued data, as we show below.

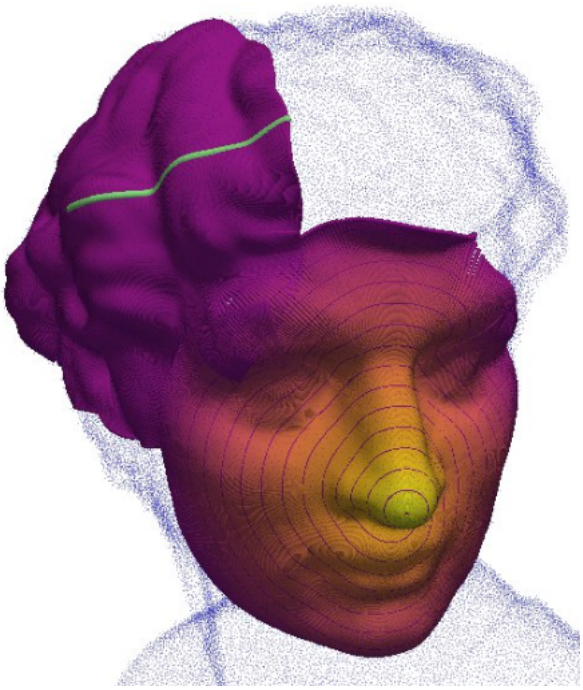


Figure 4.5: Illustration of the experiment described in Section 4.6.1. Two principal submanifolds recontructing the 'head sculpture' surface from a noisy point cloud (blue points). One submanifold is centered approximately at the tip of the nose, the other is centered at the left end of the green line. The figure shows the raw points generated by Algorithm 4 - 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 have the same sub-Riemannian distance to the base point. Likewise, the colors of the face depends on the SR distance to the base point at the tip, a lighter color signifies shorter distance. The green line on the top submanifold highlights a single SR geodesic - each of the two submanifolds consists of $L = 2500$ such geodesics.

We now assume that $\{x_i\}_{i=1\dots N}$ are points on an a priori known smooth manifold $\mathcal{N}$ of dimension $d < \infty$, equipped with a known Riemannian metric h. This is a generalization of the theory presented above, where $\mathcal{N}$ was $\mathbb{R}^d$ and h was the Euclidean metric. Our aim is now to find a lower dimensional geometric structure (e.g. a submanifold) *within* this given manifold $\mathcal{N}$.

4.5.1 Sub-Riemannian structures on a general smooth manifold

This section introduces sub-Riemannian geometry on a smooth manifold $\mathcal{N}$ of dimension d, not necessarily $\mathbb{R}^d$. A rank k sub-Riemannian structure on $\mathcal{N}$ is

determined by a rank k subbundle $\mathcal{D}$ of $T\mathcal{N}$ and a metric tensor g on $\mathcal{D}$. We will assume that the sub-Riemannian metric tensor g is the restriction $h|_{\mathcal{D}}$ of a given Riemannian metric tensor h on $T\mathcal{N}$ to $\mathcal{D}$, i.e. $g_x(u,v) = h_x(u,v)$ for all $x \in \mathcal{N}$ and $u, v \in \mathcal{D}_x$. The pair $(\mathcal{D}, g)$ is equivalent to a rank k cometric tensor $g^\star$ on $T^\star\mathcal{N}$. The triple $(\mathcal{N}, g, \mathcal{D})$, or equivalently the pair $(\mathcal{N}, g^\star)$, is called a sub-Riemannian manifold. The version of sub-Riemannian geometry we described and used in the previous sections corresponds to $\mathcal{N} = \mathbb{R}^d$ and the ambient Riemannian metric h being the Euclidean metric.

In this general setting, a curve $\gamma : [0,T] \to \mathcal{N}$ is still called horizontal if its velocities satisfy $\dot{\gamma}_t \in \mathcal{D}_{\gamma_t} \subset T_{\gamma_t}\mathcal{N}$ for all $t \in [0,T]$. And this again induces the Carnot-Carathéodory distance metric $d^{\mathcal{D}}$ (equation 4.3.2) on $\mathcal{N}$. The discussion in Section 4.3.4 about integrability and foliations carries over directly; the subbundle $\mathcal{D}$ partitions $\mathcal{N}$ into a foliation of submanifolds of dimension $\operatorname{Lie}\mathcal{D}$, and the distance metric $d^{\mathcal{D}}(x,y)$ is finite only between points on the same leaf. The Hamiltonian equations, exp and log are also defined exactly as in Section 4.3, and the relationship between the sub-Riemannian distance and the Hamiltonian (Eq. (4.3.5)) still holds. One difference from the previous Euclidean setting, however, is that the cometric cannot be expressed as a projection matrix, as we did in Equation (4.3.1). Therefore it is more convenient to represent the Hamiltonian in the following equivalent way (see [Agrachev, Barilari, and Boscain 2019], Proposition 4.22 for a derivation,

$$H : T\mathcal{N} \to \mathbb{R}_{\geq 0} : H(x,\eta) = \frac{1}{2}\sum_{i=1}^{d} \left(\eta(f_i(x))\right)^2, \tag{4.5.1}$$

where $\{f_i\}_{i=1..k}$ is an orthonormal frame for $\mathcal{D}$ w.r.t. g and $\eta(f_i(x))$ denotes the cotangent $\eta \in T_x^\star\mathcal{N}$ evaluated at the tangent $f_i(x) \in T_x\mathcal{N}$. The derivatives of the Hamiltonian that enter into the Hamiltonian equations can be expanded in a way that is suitable for implementation (see Equation (4.38) in [Agrachev, Barilari, and Boscain 2019]).

To construct a k-dimensional submanifold from a k-dimensional *non-integrable* subbundle we still need a result such as Proposition 4.3.1, which luckily holds in this general setting. The result carries over verbatim, with the dual subbundle now being the dual w.r.t. our (general) Riemannian metric h on $\mathcal{N}$, i.e.

$$\mathcal{D}_x^\star := \{h_x(v,\cdot) | v \in \mathcal{D}_x\} \subset T_x^\star\mathcal{N}.$$

Proposition 4.5.1 (The exponential is a local diffeomorphism on the dual subbundle)**.** *Let $p \in \mathcal{N}$ be arbitrary. There exists an open subset $C_p \subset \mathcal{D}^\star$*

containing 0 *such that* $\exp_p^{\mathcal{D}}|_{C_p}$ *is a diffeomorphism onto its image. That is,*

$$M_p^{\mathcal{D}} \coloneqq \exp_p^{\mathcal{D}}(C_p) \subset \mathcal{N}$$

is a smooth k*-dimensional embedded submanifold of* $\mathbb{R}^d$ *containing* p*.*

Proof. We will show that $\exp_p^{\mathcal{D}}$ is a local immersion by showing that $d_0 \exp_p^{\mathcal{D}}$ is injective ([Lee 2013], Proposition 4.1). For any $\eta \in T_0\mathcal{D} \cong \mathcal{D}$ it holds that

$$\begin{aligned} d_0\left(\exp_p^{\mathcal{D}}\right) \circ \eta &= \left.\frac{d}{ds}\right|_{s=0} \exp_p^{\mathcal{D}}(0 + s\eta) \\ &= \left.\frac{d}{ds}\right|_{s=0} \gamma_p^{\eta}(s) \\ &= g^{\star}(p)\eta, \end{aligned}$$

where the second equality uses the fact that the sub-Riemannian exponential satisfies $\exp_p^{\mathcal{D}}(s\eta) = \gamma_p^{\eta}(s)$, see corollary 8.36 in [Agrachev, Barilari, and Boscain 2019]. Viewed as a map $g_p^{\star} : \mathcal{D}^{\star} \to \mathcal{D}_p \subset \mathcal{N}$ (i.e. as the sub-Riemannian sharp map), $g_p^{\star}$ is injective on $\mathcal{D}_p^{\star}$ by construction of $\mathcal{D}_p^{\star}$. Thus $\exp_p^{\mathcal{D}}$ is an immersion. This implies the existence of a set $C_p \subset \mathcal{D}_p^{\star}$ containing 0 s.t. $\exp_p^{\mathcal{D}}\big|_{C_p}$ is an embedding ([Lee 2013] Proposition 4.25). Which implies that $M_p^{\mathcal{D}} \coloneqq \exp_p^{\mathcal{D}}(C_p)$ is an embedded k-dimensional submanifold of $\mathcal{N}$. $p \in M_p^{\mathcal{D}}$ since $\exp_p^{\mathcal{D}}(0) = p$, by definition. □

4.5.2 Principal subbundles on a Riemannian manifold

We now generalize local PCA to the setting of observations on a Riemannian manifold. In this setting, local PCA is exchanged for local *tangent* PCA, by which we mean the extraction of eigenvectors from the following second moment.

Definition 4.5.2 (Non-centered weighted tangent second moment on a Riemannian manifold)**.** Let $\{x_1, \ldots, x_N\}$ be observations on a Riemannian manifold $(\mathcal{N}, h)$. Let $K_\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a smooth, decaying kernel function with range parameter $\alpha > 0$. At a point $p \in \mathcal{N}$, we denote by $\log_p^h(x_i)$ the Riemannian log of the observation point x_i w.r.t. metric h. The *weighted tangent second moment* is defined by

$$\Sigma_\alpha(p) = \sum_{i=1}^{N} w_i(p) \left(\log_p^h(x_i) \otimes \log_p^h(x_i)\right),$$

with normalized weight functions

$$w_i \ : \ \mathcal{N} \to \mathbb{R}_{\geq 0} \ : \ p \mapsto w_i(p) = \frac{K_\alpha(\| \log_p^h(x_i)\|_p)}{\sum_{j=1}^N K_\alpha(\| \log_p^h(x_j)\|_p)}. \tag{4.5.2}$$

Remark 4.5.3. Recall that $\| \log_p^h(x_i)\|_p = d^h(p, x_i)$ since the length of the shortest geodesic from p to x_i is precisely the length of the vector in $T_p\mathcal{M}$ that exponentiates to x_i.

For any $v, u \in T_p\mathcal{N}$, the tensor product $v \otimes u$ can be identified with a linear map on $T_p\mathcal{N}$ (an endomorphism), whose coordinate representation is a $d \times d$ matrix, as explained in Section 3.3 and Appendix A.2. The coordinate expression of the second moment is

$$[\Sigma_\alpha(p)]_a = \sum_{i=1}^N w_i(p) \left[\log_p^h(x_i)\right]_a \left(\left[\log_p^h(x_i)\right]_a\right)^T [h_p]_a.$$

As in the case of Euclidean valued data, we want the principal subbundle of $T\mathcal{N}$ to be based on local PCA's centered around local means. For that purpose, the principal subbundle subspace at point p will be based on the eigendecomposition of the weighted second moment at the weighted mean $m(p)$ defined below:

Definition 4.5.4 (Weighted tangent mean map on a Riemannian manifold)**.** Let $\{x_1, \ldots, x_N\} \subset \mathcal{N}$ be observations on a Riemannian manifold $(\mathcal{N}, h)$, let the normalized weight functions w_i be defined as in (4.5.2), and let $\exp_p^h$ be the Riemannian exponential map at p w.r.t. metric h. The *weighted tangent mean map* is defined by

$$m \ : \ \mathcal{N} \to \mathcal{N} \ : \ p \mapsto m(p) = \exp_p^h \left(\sum_{i=1}^N w_i(p) \log_p^h(x_i) \right). \tag{4.5.3}$$

The eigenvectors of $\Sigma_\alpha(m(p))$ belong to the tangent space at $m(p)$, not the tangent space at p. Thus, the extracted eigenvectors needs to be mapped back to the tangent space at p, which we do by parallel transport, as described in the definition below.

The principal subbundle on $(\mathcal{N}, h)$ can only be defined at points p s.t. both p and $m(p)$ is outside the cut locus of every observation and of each other, since we need to compute the corresponding logarithms. We therefore define the set of singular points as follows,

$$S'_{\alpha,k} = \left\{ p \in \mathcal{N} \mid p, m(p) \in \bigcup_{q \in \{x_1,\ldots,x_N,p\}} \mathrm{Cut}(q) \quad \text{or} \quad \lambda_k(m(p)) = \lambda_{k+1}(m(p)) \right\}, \tag{4.5.4}$$

where $\lambda_i(m(p))$ is the i'th eigenvalue of $\Sigma_\alpha(m(p))$ of Definition 4.5.2.

Definition 4.5.5 (Principal subbundle on a Riemannian manifold)**.** Let $\lambda_1(q) \geq \cdots \geq \lambda_d(q)$ be the eigenvalues of $\Sigma_\alpha(q)$, at $q \in \mathcal{N}$, with associated eigenvectors $e_1(q), \ldots, e_d(q)$. Let $\Pi_x^y(v)$ denote parallel transport of $v \in T_x\mathcal{N}$ to $T_y\mathcal{N}$ along the length-minimizing geodesic between x and y. Then the *principal subbundle* $\mathcal{E}^{k,\alpha} \subset T\mathcal{N}$ is defined as

$$\mathcal{E}^{k,\alpha} = \left\{ (p, v) \mid p \in \mathcal{N} \setminus \mathcal{S}'_{\alpha,k}, \right. \\ \left. v \in \mathrm{span} \left\{ \Pi^p_{m(p)} e_1(m(p)), \ldots, \Pi^p_{m(p)} e_k(m(p)) \right\} \right\}$$

Remark 4.5.6. If $(\mathcal{N}, h)$ is Euclidean space, the above definition reduces to the Euclidean Definition 4.2.3 since $\left(\log_q^h(x_i) \otimes \log_q^h(x_i) \right) = (x_i - q)(x_i - q)^T$ and Π_q^p is the identity map for $q \in \mathbb{R}^d$.

Remark 4.5.7. The above construction of the subbundle subspace at p can be approximated by using the Euclidean definition in the tangent space at p, i.e. by letting $\mathcal{E}_p^{k,\alpha}$ at $p \in \mathcal{N}$ be the span of eigenvectors of $\Sigma_\alpha(0)$ computed from vectors $\{\log_p^h(x_i)\}_{i=1\ldots N} \subset T_p\mathcal{N} \cong \mathbb{R}^d$, where Σ_α is the Euclidean second moment from Definition 4.2.1. In this way, only N log's have to be computed, instead of $2N$ (see Algorithm 5), and the parallel transport operation is omitted. Note that the experiments in Section 4.6.4 uses Definition 4.5.5, not the described approximation.

Algorithm 5 describes how to compute the principal subbundle from data on a Riemannian manifold $(\mathcal{N}, h)$.

As for the Euclidean case (Proposition 4.2.7), we prove that the principal subbundle on $(\mathcal{N}, h)$ is smooth at all points where it is defined.

Proposition 4.5.8. *The principal subbundle, defined on $\mathcal{N} \setminus S'_{\alpha,k}$, is smooth.*

4.5.3 Computing with a principal subbundle on a Riemannian manifold

Given a dataset $\{x_1, \ldots, x_N\} \subset \mathcal{N}$, the associated principal subbundle $\mathcal{E}$ determines a sub-Riemannian structure on $\mathcal{N}$, namely $(\mathcal{N}, h|_{\mathcal{E}}, \mathcal{E})$. Using this

Algorithm 5 Computing the principal subbundle at a point on a Riemannian manifold

Input:

- Observations $\{x_i\}_{i=1\dots N}$ on a Riemannian manifold $\mathcal{N}$ of dimension d, and a point $p \in \mathcal{N}$, satisfying $m(p) \in \mathcal{N} \setminus \mathcal{S}_{\alpha,k}$, at which to compute the subbundle subspace.
- Parameters $\alpha \in (0, \infty)$ (range of the kernel), $k \in \{1, \dots, d-1\}$ (dimension of the subspace).

Output:

- A set of vectors spanning the principal subbundle subspace at p, $\mathcal{E}_p \subset T_p\mathcal{N}$.

1: **for** $i = 1$ **to** N **do**
2: Compute the normalized weight $w_i(p) = \frac{K_\alpha(\| \log_p^h(x_i)\|_p)}{\sum_{j=1}^N K_\alpha(\| \log_p^h(x_j)\|_p)}$.
3: **end for**
4: Compute the weighted mean $m(p) \in \mathcal{N}$ around p (Eq. (4.5.3)).
5: **for** $i = 1$ **to** N **do**
6: Compute the normalized weight at $m(p)$,

$$w_i(m(p)) = \frac{K_\alpha(\| \log_{m(p)}^h(x_i)\|_{m(p)})}{\sum_{j=1}^N K_\alpha(\| \log_{m(p)}^h(x_j)\|_{m(p)})}.$$

7: **end for**
8: Compute the weighted second moment at the weighted mean,

$$\Sigma_\alpha(m(p)) = \sum_{i=1}^N w_i(m(p)) \left(\log_{m(p)}^h(x_i) \log_{m(p)}^h(x_i)^T h_p \right).$$

9: Eigendecompose $\Sigma_\alpha(m(p))$ and select the k eigenvectors $\{e_1, \dots, e_k\}$ with the largest k eigenvalues.
10: Parallel transport each eigenvector from $T_{m(p)}\mathcal{N}$ to $T_p\mathcal{N}$ along the length-minimizing geodesic between $m(p)$ and p, yielding $e_i^\star := \Pi_{m(p)}^p e_i \in T_p\mathcal{N}$.
11: **return** $\{e_1^\star, \dots, e_k^\star\}$.

structure, we can integrate the associated sub-Riemannian Hamiltonian equations in the same way as described in section 4.4.2, except that we use the expression (4.5.1) for the Hamiltonian. This gives us sub-Riemannian exponential and

logarithmic maps on $\mathcal{N}$, so that problems A, B and C can be solved on a general Riemannian manifold, in exactly the same way as in the Euclidean case, described in sections 4.4.3-4.4.5.

A principal submanifold is computed in the same way as in the Euclidean case (Algorithm 4). It assumes that we have a representation of the manifold in a chart. Due to the centering step, computing the subbundle at a point $p \in M$ requires solving the parallel transport equation and computing $2N$ log maps, N logs between the observations and point p (lines 1-3), and N logs between the observations and the local mean around p (lines 5-7). See remark 4.5.7 for an approximation requiring only N log computations and no parallel transport. The run time of the algorithm thus depends heavily on the run time of the log map, or an approximation thereof, on the given Riemannian manifold. Examples of manifolds with computationally cheap log maps are hyperspheres, Kendall´s shape space, Grassmann manifolds, SPD matrices. See the Python library Geomstats [Miolane et al. 2020] for implementations of various manifolds including efficient log maps.

4.5.4 Relations to tangent PCA and PGA

As described in Section 2.4, the most direct generalizations of PCA to manifold-valued data are tangent PCA and PGA. We argue that a principal submanifold constructed from a principal subbundle on $\mathcal{N}$ can be seen as a locally data-adaptive combination of these two methods. In the principal subbundles framework, we compute local tangent PCA's to construct the principal subbundle $\mathcal{E}_\alpha$ of the tangent bundle $T\mathcal{N}$. This determines a data-dependent sub-Riemannian metric and thus sub-Riemannian geodesics on $\mathcal{N}$, with which we can approximate the data. In contrast to PGA, the geodesics forming the principal submanifold are not those of the ambient Riemannian manifold $(\mathcal{N}, g)$, but those of an estimated sub-Riemannian structure on $\mathcal{N}$. In PGA, the approximating submanifold is the exponential image of a k-dimensional tangent subspace, i.e. $\exp_\mu^{\mathcal{E}}(\Delta) \subset \mathcal{N}$ for $\Delta \subset T_\mu\mathcal{N}$. This is similar to a principal submanifold, except that the exponential is now the sub-Riemannian exponential determined by the principal subbundle and the subspace Δ is the metric dual of $\mathcal{E}_\mu$, the principal subbundle at μ. Note that by doing local PCA's (i.e. solving many simple, local least-squared-error problems) we remove the need for the expensive 'global' optimization for the subspace Δ necessary for PGA.

4.6 Applications

We now demonstrate how principal subbundles provide solutions to problems A, B, C, mentioned in the introduction. In particular, we reconstruct 2D submanifolds embedded in $\mathbb{R}^3$ and $\mathbb{R}^{100}$, respectively, and give a 2D tangent space representation of the latter. We furthermore evaluate a sub-Riemannian distance metric on $\mathbb{R}^{50}$ learned from observations distributed around a 4-dimensional sphere embedded in $\mathbb{R}^{50}$. In subsection 4.6.4 we compute a 1D principal submanifold approximating data on the sphere (a Riemannian manifold).

4.6.1 Surface reconstruction in $\mathbb{R}^3$ (problem *B*)

We reconstruct a 2D surface, the 'head sculpture', based on a point cloud contained in the surface reconstruction benchmark dataset from [Huang et al. 2022]. According to the classification in [Huang et al. 2022], the surface is of complexity level 2 out of 3, and the point cloud has been added noise of level 2 out of 3, see [Huang et al. 2022] for details. Note that the evaluations in the benchmarking paper was made after a preliminary denoising step, whereas our reconstruction was done on the raw point cloud. This is to illustrate the potential use of principal submanifolds for denoising. The hyperparameters we use for the principal subbundle are $\alpha = 0.001$, and $k = 2$. See Appendix B.5.2 for a reconstruction of the face using observations distorted by noise level 3 out of 3.

Figure 4.5 shows two principal submanifolds reconstructing the head sculpture locally: one is based around the tip of the nose (radius $r = 0.3$) and one at the top left side of the head ($r = 0.25$). Both base points are computed as the kernel-weighted mean around a chosen observation. The numerical parameters in Algorithm 4, determining the resolution, were $L = 2500$ (the number of geodesics) and $\Delta = 0.001$ (the integration stepsize).

A principal submanifold corresponds to a chart on the surface; in particular, a normal chart. It is a basic fact of differential geometry that a complicated surface such as the head sculpture cannot be covered by a single such chart. One therefore needs to reconstruct the surface based on multiple principal submanifolds corresponding to different base points; however, principal submanifolds based at different points might not overlap in a smooth way due to noise. To construct a smooth surface covering the whole area, we thus need a scheme for combining different principal submanifolds $M^{\mathcal{E}}_{\mu_1}, M^{\mathcal{E}}_{\mu_2}, \ldots$. Many such schemes are conceivable. In appendix B.4, we propose one that combines submanifolds by weighing points according to their sub-Riemannian distance to a set of nearest

base points. The discrepancy between submanifolds in the areas of overlap depends on the level of noise. In the experiment shown in Figure 4.5 we did not find it necessary to use a weighing scheme - see Appendix B.5.1 for a close-up illustration of the overlap.

4.6.2 Unfolding the S-surface in $\mathbb{R}^{100}$ (problem *C*)

In this experiment, we demonstrate the use of principal subbundles to contruct a representation of $\mathbb{R}^d$-valued data in $\mathbb{R}^k$, $k < d$. Let $y_i := ((y_i)_1, (y_i)_2, (y_i)_3)^T \in \mathbb{R}^3$, $i = 1..3000$, be points on the S-surface, scaled such that its height, width and depth is 1. We embed each point in $\mathbb{R}^d$, $d = 100$, by adding zeros, $\tilde{y}_i = ((y_i)_1, (y_i)_2, (y_i)_3, 0, \ldots, 0)^T$. The observations are then generated by adding Gaussian noise, $x_i \sim N(\tilde{y}_i, \sigma^2 I_d) \in \mathbb{R}^d$ for $\sigma = 0.025$.

The upper part of Figure 4.1 shows the observations $\{x_i\}_{i=1..N}$ and an approximating principal submanifold, projected to $\mathbb{R}^3$ for the purpose of visualization. The base point of the principal submanifold is the local mean around the within-sample Fréchet mean w.r.t. Euclidean distance, $\mu = (0.47, 0.47, 0.49)$. The lower part of Figure 4.1 shows the *log* representation of the observations in $\mathcal{E}^\star_\mu \cong T_\mu M^k_\alpha$. The kernel range is $\alpha = 0.01$ and the rank is $k = 2$.

4.6.3 Learning a distance metric on $\mathbb{R}^{50}$ (problem *A*)

We sample $N = 10000$ points, $\{y_i\}_{i=1..N}$, uniformly on the k-dimensional unit sphere embedded in $\mathbb{R}^d$, for $k = 4$, $d = 50$. For each of these points $y_i \in \mathbb{R}^d$ we generate an observation $x_i \in \mathbb{R}^d$ by adding d-dimensional Gaussian noise, $x_i \sim N(y_i, \sigma I_d)$, where $\sigma = 0.01$.

We generate 20 such data sets with associated principal subbundles $\mathcal{E}_j$, $j = 1..20$. For each data set we compute the SR distance $d^{\mathcal{E}_j}(p, q), j = 1 \ldots 20$, where $p = (1, 0, \ldots, 0) \in \mathbb{R}^d$ and $q = (-\sqrt{1/2}, -\sqrt{1/2}, 0, \ldots, 0) \in \mathbb{R}^d$. We find the mean, μ_0, and standard deviation, σ_0, of these 20 computed distances to be $\mu_0 = \frac{1}{20}\sum_{j=1}^{20} d^{\mathcal{E}_j}(p, q) = \frac{3}{4}\pi + 0.023$, $\sigma_0 = 0.025$. This result shows that the learned distances are close to true distance, $d^{\mathbb{S}^4}(p, q) = \frac{3}{4}\pi$, on the 4-dimensional sphere.

4.6.4 Curve approximation on the sphere

In this experiment we randomly generate 20 datasets, each with $N = 100$ points distributed around a random curve on the sphere, $\mathcal{S}^2$. The random curves are

generated as follows. A 4'th order polynomial

$$f : \mathbb{R} \to \mathbb{R} : t \mapsto (t - a_1)(t - a_2)(t - a_3)(t - a_4) \tag{4.6.1}$$

is generated by sampling roots a_1, a_2 from a uniform distribution on $(-1, 0)$, and roots a_3, a_4 from a uniform distribution on $(0, 1)$. Using two such intervals yields polynomials with more complex curvature. The graph of the polynomial, $P := \{t, f(t) | t \in [-1, 1]\}$, is considered a subset of $T_{p_0}\mathcal{S}^2$ and mapped to $\mathcal{S}^2$ by the Riemannian exponential, $\exp_{p_0}$, where $p_0 = (0, 0, 1)$ is the north pole (in extrinsic coordinates). Let $\{t_i\}_{i=1..N} \subset [-1, 1]$ be 100 evenly spaced points. Let $z_i = \exp_{p_0}((t_i, f(t_i)))$, $i = 1 \dots N$, be points on the curve on $\mathbb{S}^2$. The noisy observations are generated as $x_i = \exp_{z_i}(v_i)$, where $v_i \sim \mathrm{N}(0, I_2 \cdot \sigma)$, a 2D isotropic Gaussian with marginal variance σ, assuming a representation of $T_{z_i}\mathbb{S}^2$ in an orthonormal basis. In our experiments we used $\sigma = 5 \cdot 10^{-4}$. Note that the resulting observations on $\mathbb{S}^2$ are non-uniformly sampled along the curve (making the problem more difficult). See Figure 4.6 for an example of such a randomly generated dataset.

For each randomly generated dataset we estimate a base point as the within-sample Fréchet mean w.r.t. the geodesic distance on the sphere. We use as kernel function a Gaussian density with standard deviation $\alpha = 0.045$. This value is hand picked since our aim is to compare the performance of different methods disregarding uncertainty due to estimation of hyperparameters. Using this kernel function, we compute 3 curve approximations of the data set. Firstly, we compute the principal submanifold using Algorithm 4. Secondly, we compute the Principal submanifold without the centering and parallel transport step, i.e. the Principal flow [Panaretos, Pham, and Yao 2014]. Thirdly, we compute as baseline model the first principal geodesic from tangent PCA. For each approximation we compute the sum of squared errors (SSE), where the errors are measured by the length of the geodesic joining observation x_i and its geodesic projection to the given curve. Figure 4.6 shows an example data set and its 3 curve approximations. Figure 4.7 shows boxplots summarizing the 20 SSE's computed for each approximation method.

The SSE's and visual inspection of the corresponding plots shows that the centered version of the Principal submanifold is significantly more stable than the uncentered version (the principal flow). The uncentered version tends to stray away from the data when it reaches positions slightly outside of the point cloud. This is as expected, c.f. our discussion in Section 4.4.3. The principal geodesic has the highest SSE, as expected for this type of data that is distributed around a curve with relatively high curvature.

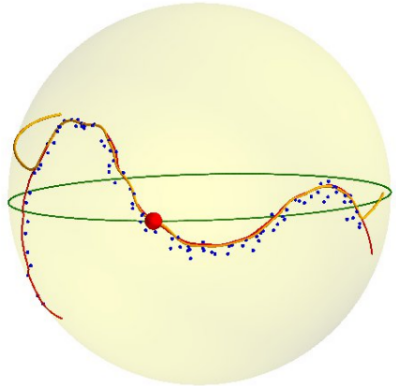


Figure 4.6: Three curves approximating a set of observations (blue points) on the sphere $\mathbb{S}^2$. The green curve is the principal geodesic computed by tangent PCA centered at the red point. The yellow curve is the principal submanifold based on a non-centered second moment (i.e. it is a principal flow). The red curve is the principal submanifold based on our proposed centered second moment. The base point of both principal submanifolds is the red point.

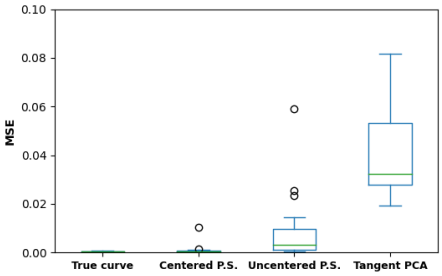


Figure 4.7: A box plot comparing the sum of squared errors (SSE), measured w.r.t. geodesic distance on the sphere, for each curve approximation. The 'True curve' label refers to the SSE for the curve $t \mapsto \exp_{p_0} \circ (t, f(t))$, described in Section 4.6.4, from which noisy samples are generated. The other labels refers to the curves described in the caption of Figure 4.6, with 'P.S.' abbreviating principal submanifold.

4.7 Conclusions

We have introduced the idea of modelling a data set $\{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ by a tangent subbundle consisting of affine subspaces of $\mathbb{R}^d$, and the sub-Riemannian geometry that it induces. We have demonstrated that geodesics w.r.t. this sub-Riemannian structure can be used to solve a number of important problems in statistics and machine learning, such as: reconstruction of submanifolds approximating the observations, finding lower dimensional representations and computing geometry-aware distances. Furthermore, we have shown that the framework generalizes to datasets on a given Riemannian manifold.

It can be considered a drawback of the framework that the point cloud must be relatively well connected, in the sense of not having large 'holes' or disconnected parts, relative to the kernel range. However, we conjecture that this can be somewhat alleviated by introducing a position-dependent range parameter.

Chapter 5

Conclusion and future work

Contents

5.1 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 3, 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 4, 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.

5.2 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 3, 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.

5.3 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 4.2.1. 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 4).

Extensions In this section, we mention two alternative ways to use a principal subbundle. The common theme is to consider other classes 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 4. 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 4.3.2)). 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 4.3.2). 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 4, e.g. when computing

projections to a principal submanifold in Section 4.4.3.2, 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

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

   For each point $y_{t_0}^{v_0} := \exp_p(t_0 v_0) \in \overline{B_p}$, 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 2, Chapter 3).

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 4, 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,

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

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

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 $p_u := \exp_p(u)$. 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 $\nabla_\cdot R$ 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 $\overline{x_j} := \log_\mu(x_j)$, 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 $p \in M$.

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

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

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

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

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

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

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

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

Equation (A.4.4) uses the fact that

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

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 4

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.

Proposition 4.2.7. *The principal subbundle, defined on $\mathbb{R}^d \setminus \mathcal{S}_{\alpha,k}$, is smooth.*

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 $\mathcal{E}^{\alpha,k}_{p'}$ 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 [J.-g. 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 $\bar{\lambda}_i(p) = \lambda_i(p), i = 1 \ldots d$.

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

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

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

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, $i, j, k \in \{1, \dots, d\}$ 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. □

Proposition 4.5.8. *The principal subbundle, defined on $\mathcal{N} \setminus S'_{\alpha,k}$, 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 4.2.7, 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 [J.-g. 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 4.6.

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. (4.3.4). This is desired because the computation of curve length and distance via eq. (4.3.5) 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 4.5, zooming in on the region of overlap between the two principal submanifolds.

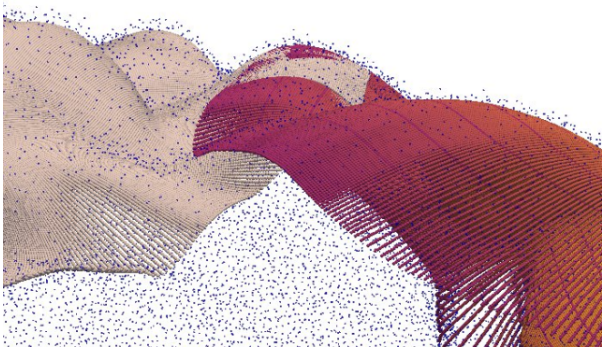


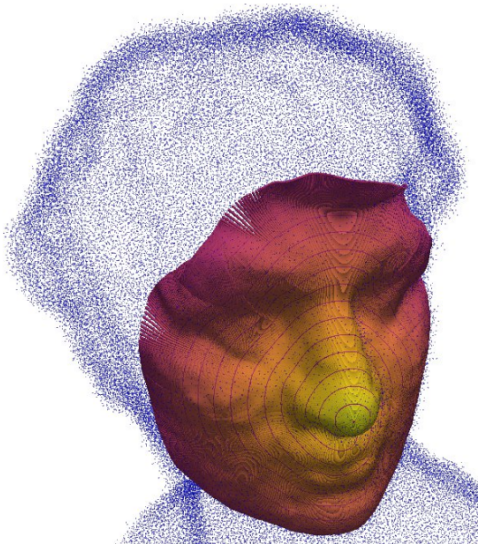
Figure B.1: Supplementary figure to Figure 4.5, zooming in on the region where the two submanifolds overlap. The left, beige submanifold in this figure is the purple one in Figure 4.5, recolored to be able to distinguish more clearly the two submanifolds.

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

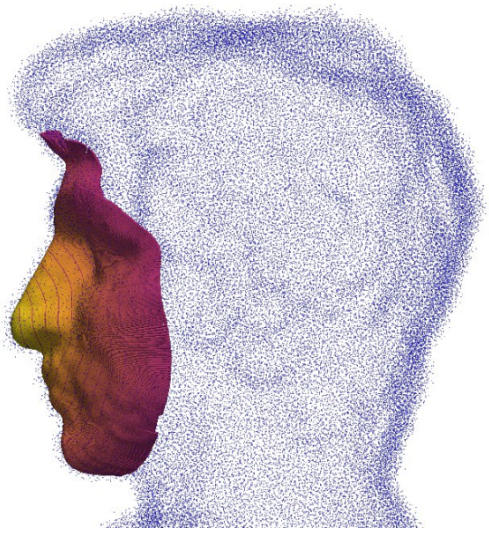
Figure B.2 illustrates the reconstruction of the face of the 'head sculpture' (from the benchmark dataset described in [Huang et al. 2022]), with noise level 3 out of 3. The parameters are the same as for the experiment described in section 4.6.1 except for a slightly larger kernel range.

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 section4.6.3. 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 4 - 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.

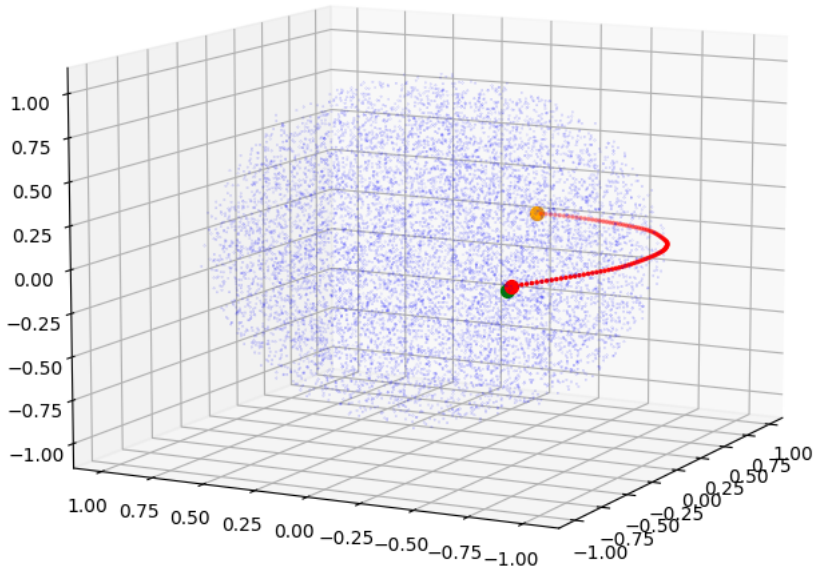


Figure B.3: Illustration of a single computed geodesic found by solving the log problem $\log_p(q)$, for p, q and observations as described in section 4.6.3. The blue points are observations on the 4-dimensional sphere embedded in $\mathbb{R}^{50}$ projected to $\mathbb{R}^3$. The orange point is the 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.

Adams, D., Rohlf, F. J., and Slice, D. (2013). "A field comes of age: geometric morphometrics in the 21st century". In.

Afsari, B. (2011). "Riemannian ˆ{} center of mass: existence, uniqueness, and convexity". In: *Proceedings of the American Mathematical Society* vol. 139, no. 2, pp. 655–673.

Agrachev, A., Barilari, D., and Boscain, U. (2019). *A comprehensive introduction to sub-Riemannian geometry*. Vol. 181. Cambridge 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.

Arnaudon, M. and Miclo, L. (2014). "Means in complete manifolds: uniqueness and approximation". In: *ESAIM: Probability and Statistics* vol. 18, pp. 185–206.

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.

Bellet, A., Habrard, A., and Sebban, M. (2015). "Metric learning". In: *Synthesis lectures on artificial intelligence and machine learning* vol. 9, no. 1, pp. 1–151.

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.

Cavalli-Sforza, L. L. and Edwards, A. W. (1967). "Phylogenetic analysis. Models and estimation procedures". In: *American journal of human genetics* vol. 19, no. 3 Pt 1, p. 233.

Cayton, L. (2005). "Algorithms for manifold learning". In: *Univ. of California at San Diego Tech. Rep* vol. 12, no. 1-17, p. 1.

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

Chow, W.-L. (2002). "Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung". In: *The Collected Papers Of Wei-Liang Chow*. World Scientific, pp. 47–54.

Conith, A. J., Meagher, M. A., and Dumont, E. R. (2018). "The influence of climatic variability on morphological integration, evolutionary rates, and disparity in the Carnivora". In: *The American Naturalist* vol. 191, no. 6, pp. 704–715.

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.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical shape analysis: with applications in R*. Vol. 995. John Wiley & Sons.

Felsenstein, J. (1973). "Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters". In: *Systematic Biology* vol. 22, no. 3, pp. 240–249.

— (1985). "Phylogenies and the comparative method". In: *The American Naturalist* vol. 125, no. 1, pp. 1–15.

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.

Frostig, R., Johnson, M. J., and Leary, C. (2018). "Compiling machine learning programs via high-level tracing". In: *Systems for Machine Learning* vol. 4, no. 9.

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

Gavrilov, A. V. (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.

Gropp, A., Yariv, L., Haim, N., Atzmon, M., and Lipman, Y. (2020). “Implicit geometric regularization for learning shapes”. In: *arXiv preprint arXiv:2002.10099*.

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.

Guigui, N. and Pennec, X. (2022). “Numerical accuracy of ladder schemes for parallel transport on manifolds”. In: *Foundations of Computational Mathematics* vol. 22, no. 3, pp. 757–790.

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.

Hsu, E. P. (2002). *Stochastic analysis on manifolds*. 38. American Mathematical Soc.

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.

Jean, F. (2014). *Control of nonholonomic systems: from sub-Riemannian geometry to motion planning*. Springer.

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.

Karcher, H. (1977). "Riemannian center of mass and mollifier smoothing". In: *Communications on pure and applied mathematics* vol. 30, no. 5, pp. 509–541.

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.

Koelle, S. J., Zhang, H., Meila, M., and Chen, Y.-C. (2022). "Manifold Coordinates with Physical Meaning". In: *Journal of Machine Learning Research* vol. 23, no. 133, pp. 1–57.

Kühnel, L., Sommer, S., and Arnaudon, A. (2019). "Differential geometry and stochastic dynamics with deep learning numerics". In: *Applied Mathematics and Computation* vol. 356, pp. 411–437.

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.

Ma, Y. and Fu, Y. (2012). *Manifold learning theory and applications*. Vol. 434. CRC press Boca Raton.

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.

Martins, E. P. and Hansen, T. F. (1996). “The statistical analysis of interspecific data: a review and evaluation of phylogenetic comparative methods”. In: *Phylogenies and the comparative method in animal behavior*, pp. 22–75.

— (1997). “Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data”. In: *The American Naturalist* vol. 149, no. 4, pp. 646–667.

Micheli, M., Michor, P. W., and Mumford, D. (2012). “Sectional curvature in terms of the cometric, with applications to the Riemannian manifolds of landmarks”. In: *SIAM Journal on Imaging Sciences* vol. 5, no. 1, pp. 394–433.

Michor, P. W. (2020). “Manifolds of mappings for continuum mechanics”. In: *Geometric Continuum Mechanics*, pp. 3–75.

Miller, M. I., Trouvé, A., and Younes, L. (2002). “On the metrics and Euler-Lagrange equations of computational anatomy”. In: *Annual review of biomedical engineering* vol. 4, no. 1, pp. 375–405.

Minka, T. (2000). “Automatic choice of dimensionality for PCA”. In: *Advances in neural information processing systems* vol. 13.

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

Myhre, J. N., Shaker, M., Kaba, M. D., Jenssen, R., and Erdogmus, D. (2020). "A generic unfolding algorithm for manifolds estimated by local linear approximations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 854–855.

Nyakatura, K. and Bininda-Emonds, O. R. (2012). "Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates". In: *BMC biology* vol. 10, no. 1, pp. 1–31.

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.

Perrault-Joncas, D. and Meila, M. (2013). "Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery". In: *arXiv preprint arXiv:1305.7255*.

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

Polly, P. D., Lawing, A. M., Fabre, A.-C., and Goswami, A. (2013). "Phylogenetic principal components analysis and geometric morphometrics". In: *Hystrix* vol. 24, no. 1, p. 33.

Revell, L. J. (2009). "Size-correction and principal components for interspecific comparative studies". In: *Evolution: International Journal of Organic Evolution* vol. 63, no. 12, pp. 3258–3268.

Rifford, L. (2014). *Sub-Riemannian geometry and optimal transport*. Springer.

Rohlf, F. J. (1999). "Shape statistics: Procrustes superimpositions and tangent spaces". In: *Journal of classification* vol. 16, pp. 197–223.

Said, S. and Manton, J. (2012). "Brownian processes for monte carlo integration on compact Lie groups". In: *Stochastic Analysis and Applications* vol. 30, no. 6, pp. 1062–1082.

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.

Singer, A. and Wu, H.-T. (2012). “Vector diffusion maps and the connection Laplacian”. In: *Communications on pure and applied mathematics* vol. 65, no. 8, pp. 1067–1144.

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. (1985). “Eigenvalues And Eigenvectors Of A Matrix Dependent On Several Parameters.” In: *Journal of Computational Mathematics* vol. 3, no. 4, p. 351.

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.

Teh, Y. and Roweis, S. (2002). “Automatic alignment of local representations”. In: *Advances in neural information processing systems* vol. 15.

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

Udriste, C. (1994). *Convex functions and optimization methods on Riemannian manifolds*. Vol. 297. Springer Science & Business Media.

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.

Zhang, Z. and Zha, H. (2004). “Principal manifolds and nonlinear dimensionality reduction via tangent space alignment”. In: *SIAM journal on scientific computing* vol. 26, no. 1, pp. 313–338.