Kernels on the Riemannian Manifold of Symmetric Positive Definite Matrices

vorlegt von

Gabriel Leander Wagner vom Berg, M. Sc.
ORCID: 0009-0009-3675-641X

an der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor*in der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Konrad Rieck
Gutachter: Prof. Dr. Benjamin Blankertz
Gutachter: Prof. Dr. Sylvain Chevallier (Université Paris-Saclay, France)
Gutachter: Fabien Lotte, PhD, HDR (Inria Centre at the University of Bordeaux, France)

Tag der wissenschaftlichen Aussprache: 22. Mai 2025

Berlin 2025

Zusammenfassung.

Diese Dissertation präsentiert eine Formalisierung von Kernel-Methoden für Riemannsche Mannigfaltigkeiten, mit besonderem Fokus auf die Mannigfaltigkeit symmetrisch positiv definiter (SPD) Matrizen. Die Arbeit führt das Konzept der Riemannschen linearen Features ein, einen neuartigen Ansatz zur Linearisierung der Mannigfaltigkeit, der die lineare Struktur des Tangentialraums, gegeben durch die Riemannsche Metrik, einbezieht. Zwei allgemeine Klassen von Kernels werden für häufig verwendete Riemannsche Metriken auf SPD-Matrizen beschrieben und verglichen: kanonische Kernels und Kernels in metrischen Räumen.

Die Dissertation bietet einen umfassenden theoretischen Rahmen, einschließlich einer detaillierten Untersuchung der Riemannschen Geometrie, metrischer Räume und Distanzgeometrie. Anschließend werden diese theoretischen Konzepte angewendet, um neue Formalisierungen der Riemannschen Mannigfaltigkeit positiv definiter Matrizen zu entwickeln. Es werden die Standardmetriken auf SPD Matrizen beschrieben und gezeigt, wie die neuen Formalisierungen in jede Metrik integriert werden.

Die Leistung der vorgeschlagenen Kernel-Methoden wird durch zwei Vergleiche evaluiert: Klassifikation und Dimensionsreduktion. Diese werden anhand einer zwei-Klassen Motor-Imagery-Aufgabe mit Elektroenzephalographie (EEG)-Aufzeichnungen validiert.

Diese Arbeit trägt zum Forschungsfeld bei, indem sie eine einheitliche Formalisierung von Kernel-Methoden für Riemannsche Mannigfaltigkeiten und insbesondere SPD-Matrizen bietet. Sie liefert Einblicke in die Konstruktion und Anwendung von Kernels in diesem Rahmen und eröffnet neue Forschungsmöglichkeiten für Brain-Computer-Interfaces und andere Bereiche, in denen SPD-Matrizen als Features für maschinelle Lernmethoden genutzt werden können.

Abstract.

This thesis presents a formalization of kernel methods for Riemannian manifolds, with a focus on the manifold of symmetric positive definite (SPD) matrices. The work introduces the concept of Riemannian linear features, a novel approach to linearizing the manifold that incorporates the linear structure of the tangent space given by the Riemannian metric. Two general classes of kernels are described and compared for commonly used Riemannian metrics on SPD matrices: canonical kernels and metric space kernels.

The thesis provides a comprehensive theoretical framework, including a detailed exploration of differential and Riemannian geometry, metric spaces, and distance geometry. It then applies this framework to develop new formalizations of the Riemannian manifold of positive definite matrices, describing standard metrics and showing how the new concepts are incorporated into each metric.

The performance of the proposed kernel methods is evaluated through two comparisons: classification and dimensionality reduction. These are validated using a two-class motor imagery task with electroencephalography (EEG) recordings, employing a cross-session within-subject leave-one-session-out paradigm.

This work contributes to the field by offering a unified formalization of kernel methods for Riemannian manifolds and SPD matrices in particular. It provides insights into the construction and application of kernels in this framework, opening up new avenues for research in areas such as Brain-Computer Interfaces and other domains where SPD matrices can be extracted as features for machine learning methods.

Contents

1 Introduction

The field of Riemannian geometry for symmetric positive definite (SPD) matrices with an application to methods for Brain-Computer-Interfaces has gained increasing popularity and attention over the past decade [15]. Many new methods have been developed, and different attempts have been made to transfer from methods in Euclidean space. One field of research is kernel methods, which pose challenges when considering Riemannian geometry, as many results from Euclidean space cannot be transferred directly. There are several individual methodological descriptions of kernel methods in a variety of publications in the field, cf., e.g., [3, 27, 33]. A review of kernel methods was done in 2013 [67], but did not include any formalization of kernel methods and was only limited to two kernels. In several publications, the Gaussian kernel for geodesic distances on Riemannian manifolds is considered, and the authors give prerequisites that a manifold needs to fulfill so that the Gaussian kernel is positive definite [22, 32]. None of the publications have attempted a joint formalization of kernel methods for Riemannian manifolds and the manifold of SPD matrices in particular, making it difficult to compare methods between Riemannian metrics and compare different methods within the same Riemannian metric.

This thesis presents such a formalization. We introduce the concept of Riemannian linear features, a method for linearly approximating the manifold that incorporates the linear structure of the tangent space rather than using the tangent vectors as vectors in standard Euclidean space. We test and compare the linearizations by Riemannian linear features and tangent vectors in their ability to retain the distance geometry of the manifold.

We describe two general classes of kernels and compare these methods for commonly used Riemannian metrics on SPD matrices.

The first kernel class is the class of the so-called canonical kernels combining the Riemannian metric with the logarithmic map following the procedure described in [3] but generalized to arbitrary Riemannian metrics. In the notation of Riemannian linear features, the canonical kernels can be written as linear kernels on the Riemannian linear features. These kernels are always positive definite, regardless of the metric.

The second class of kernels is based on the set of SPD matrices viewed as a metric space. This class is not necessarily positive definite for all metrics. We leverage the formalization of Riemannian linear features to allow for the construction of positive definite metric space kernels for all manifolds. Although this thesis is focused on kernels, we can also transfer the formalization to other methods.

To test the performance of the proposed kernel methods, we employ two different comparisons: classification and dimensionality reduction. We use a two-class motor imagery task validated in a cross-session within-subject leave-one-session-out paradigm with electroencephalography (EEG) recordings. All datasets are retrieved using the Mother-of-all-BCI-Benchmarks (MOABB, [31]) library that provides the general comparison framework. The general description of the datasets and preprocessing steps is found in section 8.4.

In the chapter *Mathematical Preface*, we introduce the mathematical concepts referenced throughout the thesis. It starts with a simplified introduction to differential and Riemannian geometry. Due to the importance of metric spaces in constructing kernel methods, we give a short overview of the field of metric spaces and dis-

tance geometry. This chapter does not contain original work and serves simply as a reference to mathematical concepts within the thesis to prevent having to look up the concepts in various books.

The third chapter *Theory* introduces new formalizations of the Riemannian manifold of positive definite matrices. It describes standard metrics on SPD matrices and shows how the new concepts are incorporated into each metric.

The fourth chapter *Linearization of the Manifold* compares the linearizations of the manifold by Riemannian linear features and the tangent space.

The fifth chapter *Kernels on Riemannian Manifolds* describes and formalizes kernels on Riemannian manifolds. It differentiates the two classes of kernels, the inner product kernels, and metric space kernels. Here, we specify under which circumstances the results from standard Euclidean space can be transferred to the various Riemannian manifolds of SPD matrices built from each Riemannian metric. In cases where the results can not be transferred directly, we describe approximations of these methods.

In the sixth chapter *Principal Geodesic Analysis*, we describe the Principal Geodesic Analysis (PGA, [24]), the manifold equivalent of Principal Component Analysis (PCA, [29]). We derive a new way of linearizing the manifold by PGA and compare it to the linearization with Riemannian linear features. Additionally, we compare PGA with kernel PCA in a dimensionality reduction task.

The seventh chapter concludes the thesis with a discussion of its limitations and an outlook on various new fields of research that are opened up by the new formalization, followed by the closing words.

2 Mathematical Preface

This chapter contains a theoretical introduction to the general mathematical concepts referenced throughout the thesis. We briefly introduce the theoretical concepts of Riemannian geometry, which are the foundation for the analysis of kernel methods in this thesis. In the section *Metric Spaces* (section 2.2), we discuss the relationship of points in space as given by their distances and relate Riemannian manifolds to metric spaces (section 2.2.3). Readers familiar with these subjects may skip over the following derivations and use them more as a look-up section, as all relevant topics are cross-referenced throughout the thesis.

2.1 Riemannian Geometry

We will give a short overview of Riemannian geometry concepts relevant to the description and formalization of kernels on Riemannian manifolds. At the heart of Riemannian geometry lie two key elements that fully describe its geometric structure: A differentiable manifold M and an inner product g on the so-called tangent space of M. In the following derivation, we lean on the work of do Carmo [11].

2.1.1 Differentiable Manifolds

This section is included for the sake of completeness of the derivation and is kept very technical. Readers familiar with the concept of a differentiable manifold and the tangent space may continue to the next section.

Definition 2.1.1 ([11], Def. 2.1)**.** *A differentiable manifold of dimension n is a set M and a family of injective mappings $\mathbf{x}_\alpha : U_\alpha \subset \mathbb{R}^n \to M$ of open sets U_α of $\mathbb{R}^n$ into M such that:*

1. $\bigcup_\alpha \mathbf{x}_\alpha (U_\alpha) = M$.
2. *for any pair α, β, with $\mathbf{x}_\alpha (U_\alpha) \cap \mathbf{x}_\beta (U_\beta) = W \neq \phi$, the sets $\mathbf{x}_\alpha^{-1}(W)$ and $\mathbf{x}_\beta^{-1}(W)$ are open sets in $\mathbb{R}^n$ and the mappings $\mathbf{x}_\beta^{-1} \circ \mathbf{x}_\alpha$ are differentiable.*

We use $(M, \mathbf{x})$ to denote the manifold. We say that $\mathbf{x}$ equips M with a differentiable structure.

For the sake of simplicity, we will henceforth use M to refer to the manifold rather than the set of points. The mapping that equips M with a differentiable structure should be evident in all cases.

Definition 2.1.2. *Let $P \in \mathbf{x}_\alpha (U_\alpha)$. The pair $(U_\alpha, \mathbf{x}_\alpha)$ is called a parameterization, or system of coordinates, of M at P.*

As do Carmo [11, p. 4] pointed out, the Euclidean space $\mathbb{R}^n$ equipped with the differentiable structure given by the identity function is a trivial differentiable manifold.

The structure of SPD matrices demonstrates that we can use the matrix entries as coordinates. As the matrices are symmetric, it suffices to extract the triangular part. Thus, SPD matrices can easily be given a differentiable structure with the identity function as the parametrization.

By definition, every differential manifold M can be linearly approximated at each point $p \in M$, giving rise to the concept of the tangent space.

Definition 2.1.3 ([11], Def. 2.6)**.** *Let M be a differentiable manifold. A differentiable function $\alpha : (-\varepsilon, \varepsilon) \to M$ is called a (differentiable) curve in M. Suppose that $\alpha(0) = p \in M$, and let $\mathcal{D}$ be the set of functions on M that are differentiable at p. The tangent vector to the curve α at $t = 0$ is a function $\alpha'(0) : \mathcal{D} \to \mathbb{R}$ given by*

$$\alpha'(0)f = \left.\frac{d(f \circ \alpha)}{dt}\right|_{t=0}, \quad f \in \mathcal{D}.$$

A tangent vector at p is the tangent vector at $t = 0$ of some curve $\alpha : (-\varepsilon, \varepsilon) \to M$ with $\alpha(0) = p$. This definition as an operator might be contrary to the intuition of a vector. However, recalling that linear functions can be identified with matrices merges these two concepts.

Definition 2.1.4. *[11, Def. 2.6] The set $T_P M = \{\alpha'(0)f | \alpha$ is a curve on M at $P\}$ of all tangent vectors to M at P is called the tangent space.*

A differentiable manifold can be considered similarly to differentiable functions. As differentiable functions are defined by the possibility to approximate the function at any given point by a linear function, differentiable manifolds can be thought of in the same way. The linear approximation of any differentiable manifold at any point is again a (linear) manifold, which we call the tangent space. $T_P M$ forms a vector space of dimension n. We can think of the tangent vectors as the derivative of curves $\alpha : \mathbb{R} \to M$ on M, and tangent vectors span the tangent space.

Definition 2.1.5. *cf. [11, Def. 5.1] A differentiable vector field X on a differentiable manifold M is a map that associates to each point $P \in M$ a point $X(P) \in T_P M$. Considering any parametrization $\mathbf{x} : U \subset \mathbb{R}^n \to M$, we can define smooth vector fields via differentiable functions $a_i : U \to \mathbb{R}$ by*

$$X(P) := \sum_{i=1}^{n} a_i(P) \frac{\partial}{\partial x_i},$$

where $\{\frac{\partial}{\partial x_i}\}$ is the basis associated with $\mathbf{x}$.

We now have all the necessary constructs from the differential manifold at our disposal, allowing us to take a step further toward defining a Riemannian manifold.

The last ingredient for this is the so-called Riemannian metric tensor, often just called the Riemannian metric.

2.1.2 Riemannian Metric

Definition 2.1.6. *[11, Def. 2.1] A Riemannian metric on a differentiable manifold M is a family $(g_P)_{P \in M}$ of inner products $g_P =< ., . >_P$ on the tangent space $T_P M$ that varies differentiably with $P \in M$.*

Varying differentiably can be understood in the following sense: For any two vector fields X and Y on M, the derivative $\frac{d}{dP}$ of the function $h(P) := g_P(X(P), Y(P))$ exists. This definition allows us to equip the tangent space with an inner product that differs from the standard Euclidean inner product.

Now we are ready to define a Riemannian Manifold, the core structure of our analysis.

Definition 2.1.7. *[11, p. 38] Let M be a differentiable manifold and g a Riemannian metric defined on T_PM. Then (M, g) is called a Riemannian manifold.*

We note that the structure of (M, g) and all of the following properties depend on the choice of the iemannian metric g.

Remark 2.1.8. *The tangent space is a finite-dimensional inner product space where the inner product is the Riemannian metric, i.e., for any $P \in M$ where M is an n-dimensional manifold, (T_PM, g_P) is an inner product space. By theorem 8.1.4 it is isometric to standard Euclidean space $(\mathbb{R}^n, < ., . >)$. However, these spaces have different inner products unless $g_P =< ., . >$.*

We will see in section 3.3 how we can find these isometries for SPD matrices and use them to construct a feature map from the manifold to standard Euclidean space in definition 3.6.1.

2.1.3 Geodesic

Definition 2.1.9. *Let $\gamma_{A\to B} : \mathbb{R} \to M$, $\gamma(0) = A, \gamma(1) = B$ be a piecewise differentiable curve on a given manifold connecting two points $A, B \in M$. The length of γ is defined as*

$$L(\gamma) = \int_0^1 g_{\gamma(t)}(\gamma'(t), \gamma'(t))^{\frac{1}{2}} dt.$$

A geodesic is defined as a shortest curve connecting two points on the manifold. Geodesics are not necessarily unique on general Riemannian manifolds. However, all Riemannian manifolds considered in this thesis admit unique geodesics between any two points, so we define a geodesic accordingly.

Definition 2.1.10. *[56, Sec. 1.3.2] Let $A, B \in M$ be points on a Riemannian manifold and denote by $\Gamma := \{\gamma | \gamma$ is a piecewise differentiable curve connecting A and $B\}$ the set of curves connecting A and B. A geodesic $\gamma^*_{A\to B}$ between A and B is the minimizer of the curve length functional, i.e.*

$$\gamma^* = \arg\min_{\gamma\in\Gamma} L(\gamma).$$

Geodesics can alternatively be defined by a starting point and an initial velocity given by a tangent vector V of the tangent space at the starting point $P \in M$. As geodesics satisfy a second-order ordinary differential equation that can be derived from the condition of minimizing length, cf. [30, Eq. 25], it can be shown that there exists only one geodesic $\gamma_{P,V}$ starting from a point $P = \gamma_{P,V}(0) \in M$ with tangent vector $V = \gamma'_{P,V}(0) \in T_PM$ that defines the so-called initial velocity of the geodesic. We will see how these two concepts are related in section 3.1. For geodesics connecting two points, we will use the notation $\gamma_{A\to B}$ whereas, for a geodesic starting at P with velocity V, we will use the notation $\gamma_{P,V}$. Regardless of the notation, we can state a property that can be derived from the differential equation of geodesics that will become relevant later:

Lemma 2.1.11. *[30, Corr. 4.3] Let (M, g) be a Riemannian manifold. Let $\gamma(t)$ be a geodesic on M. Then for all $t \in \mathbb{R}$ we have*

$$g_{\gamma(t)}(\gamma'(t), \gamma'(t)) = c$$

where c is a positive scalar.

This can be thought of as a geodesic having a constant velocity vector $V := \gamma'(t) \in T_{\gamma(t)}M$ on the tangent spaces of points along the geodesic.

Geodesic completeness is a concept that guarantees that each geodesic is entirely included on the manifold M. This means that when extending geodesics to infinity, the resulting points all still lie on M. Formally:

Definition 2.1.12. *[56, Sec. 1.3.5] A manifold is called geodesically complete if for any geodesic $\gamma_{A\to B}$ between two points on $A, B \in M$ we have $\gamma(t) \in M$ for all $t \in \mathbb{R}$.*

Geodesics on connected Riemannian manifolds give rise to a distance measure derived from the underlying geometry of the space.

Definition 2.1.13. *[56, Eq. 1.6] Let $\gamma_{A\to B}(t)$ be a geodesic connecting the points $A, B \in M$. Then*

$$d(A, B) := L(\gamma)$$

is the distance between points A and B.

The Karcher mean (definition 8.1.12) under the geodesic distance is not necessarily unique for general Riemannian metrics and it is not always possible to obtain an analytic solution [15]. The mean is unique for all Riemannian manifolds considered in this thesis and will henceforth be referred to as the Frechét mean (see definition 8.1.12).

2.1.4 Exponential and Logarithmic Maps

The exponential and logarithmic maps allow us to map points back and forth between manifold and tangent space. The naming of these maps arises from a generalization of matrix Lie group theory, where the exponential map corresponds to the matrix exponential map [37, Ch. 20]. Both are parametrized by a reference point P at which a given tangent space is considered.

Definition 2.1.14. *[36, p. 128] Let (M, g) be a Riemannian manifold. For every $P \in M$, let $D(P) \subset T_PM$ be the set of vectors $V \in T_PM$ such that the unique geodesic $\gamma_{P,V}$ with initial conditions $\gamma_{P,V}(0) = P$ and $\gamma'_{P,V}(0) = V$ is defined at least on the interval $[0, 1]$. The exponential map at P is the map $\mathrm{Exp}_P : D(P) \to M$ defined by:*

$$\mathrm{Exp}_P(V) = \gamma_{P,V}(1) =: Q$$

In other words, $\mathrm{Exp}_P(V)$ is the point on M obtained by following the geodesic starting at P with initial velocity V for one unit of time.

The logarithmic map on a Riemannian manifold (M, g) is defined as the local inverse of the exponential map.

Definition 2.1.15. *[56, Sec. 1.3.6] Let (M, g) be a Riemannian manifold and $P \in M$. There exists an open neighborhood U of 0 in T_PM such that $\mathrm{Exp}_P(.) : U \to \mathrm{Exp}_P(U) \subset M$ is a diffeomorphism. Let $V \in U$. For $Q \in \mathrm{Exp}_P(U)$, the logarithmic map of Q at P, denoted $\mathrm{Log}_P(Q) : U \subset M \to T_PM$, is defined as:*

$$\mathrm{Log}_P(Q) = \mathrm{Exp}_P^{-1}(Q) = V$$

Lemma 2.1.16. *Let $P \in M$ and $\mathrm{Log}_P(.)$ the logarithmic map at point P. Then $\mathrm{Log}_P(P) = 0$*

Proof. The geodesic $\gamma_{P,V}$ in the tangent direction $V = 0$ does not change, i.e., $\gamma_{P,0}(t) = P$ for all t. Thus, $\mathrm{Exp}_P(0) = P$ and $\mathrm{Log}_P(P) = 0$ per definition. □

2.1.5 Pullback Metrics

Pullback metrics are a type of metric that transfer the Riemannian metric structure from one Riemannian manifold to a differentiable manifold leveraging a so-called pullback function. The isometry ϕ and its differential are the key elements of pullback metrics, which define the explicit form of the new resulting metric.

Definition 2.1.17. *[60] Let (N, g) be a Riemannian manifold where g is a Riemannian metric on N. Let M be a differentiable manifold and $\phi : M \to N$ a smooth map from M to N. The pullback metric $\phi^* g$ on M is defined using the differential (also called pushforward) of ϕ, denoted by $D\phi$. For any two tangent vectors $X, Y \in T_P M$ at a point $P \in M$ the pullback metric $\phi^* g$ on M is given by*

$$(\phi^* g)_P(X, Y) = g_{\phi(P)}([\mathrm{D}_P\, \phi](X), [\mathrm{D}_P\, \phi](Y)).$$

$[\mathrm{D}_P\, \phi] : T_P M \to T_{\phi(P)} N$ is the differential of ϕ at P, mapping tangent vectors from M to N.

By this construction, $\phi : M \to N$ is an isometry between M and N. All manifold properties beyond the inner product can be transferred similarly using the pullback. Pullback metrics are defined via differentials, which need to be evaluated to compute the properties of a Riemannian manifold.

Table 2 of [61] shows a list of the transferred properties. The Euclidean structure is commonly transferred to other manifolds using pullback metrics as it is simple to use. The resulting metrics inherit key properties from the Euclidean metric.

The Frechét mean of data points $(x_i)_{i=1,\dots,n} \in M$ according to a pullback metric under an isometry $\phi : M \to N$ can be calculated as

$$\mu = \phi^{-1}(\arg\min_{q \in M} \sum_{i=1}^{n} d_N(q, \phi(x_i))^2).$$

where d_N is the geodesic distance on N.

2.1.6 Curvature

An essential concept of Riemannian manifolds is curvature, which allows us to relate them to Euclidean space. The complete derivation of the notion of curvature in all its mathematical complexity can be found in [11, Ch. 4]. As this thesis requires only a limited understanding of the topic, we will formalize curvature in a simpler way.

Consider a Riemannian manifold (M, g) and three points $X_1, X_2, X_3 \in M$. The geodesics connecting these points, or vertices, form a geodesic triangle, analogous to a triangle with straight lines in Euclidean space.

In Euclidean space, the angles of a triangle sum to π. For a geodesic triangle, the angles can be calculated using the inner product of the tangent vectors at each vertex X_i. Let θ_i denote the angle at vertex X_i, calculated as $\cos(\theta_i) = \frac{g(\gamma' X_i \to X_j, \gamma' X_i \to X_k)}{|\gamma' X_i \to X_j||\gamma' X_i \to X_k|}$. The sum of these angles is denoted as $\sum_{i=1}^{3} \theta_i$.

Definition 2.1.18. *Let (M, g) be a Riemannian manifold. Let $X_1, X_2, X_3 \in M$ form a geodesic triangle. We say that a geodesic triangle has*

- *zero curvature if $\sum_{i=1}^{3} \theta_i = \pi$,*

- *positive curvature if* $\sum_{i=1}^{3} \theta_i < \pi$,
- *negative curvature if* $\sum_{i=1}^{3} \theta_i > \pi$

for any points $X_1, X_2, X_3 \in M$. *We say a manifold* M *has*

- *zero curvature if all geodesic triangles in* M *have zero curvature,*
- *positive curvature if all geodesic triangles in* M *have positive curvature,*
- *negative curvature if all geodesic triangles in* M *have negative curvature,*
- *mixed curvature if the geodesic triangles in* M *have positive and negative curvature,*

We say that it has constant curvature if for any points $X_1, X_2, X_3 \in M$ *forming a geodesic triangle with constant side length* $d(X_i, X_j) = c$ *there exists a positive scalar* k *such that*

$$\sum_{i=1}^{3} \theta_i = k\pi.$$

It follows from the definition of geodesic triangles that for a Riemannian manifold (M, g) these two statements are equivalent:

- The manifold is locally isometric to Euclidean space
- The manifold has zero curvature

Remark 2.1.19. *A manifold* (M, g) *with constant curvature of zero is called a flat manifold. The Riemannian metric* g *is then called a flat metric. If the manifold* (M, g) *is geodesically complete and simply connected, it is globally isometric to standard Euclidean space.*

2.1.7 Sections of the Orthonormal Frame Bundle

Definition 2.1.20. *Let* (M, g) *be a* k*-dimensional Riemannian manifold. Then the orthonormal frame bundle of* M *at a point* $P \in M$ *is denoted by* $O(M)_P$ *and defined by*

$$O(M)_P := \{L : (\mathbb{R}^k, <.,.>) \to (T_P M, g), L \text{ is a linear isometry}\}$$

for all $P \in M$. $f_P \in O(M)_P$ *is called a section of the orthonormal frame bundle at* P.

This technical definition is the foundation of the formalization of Riemannian kernel methods and will be described in more detail in section 3.3. A section is a linear isometry from standard Euclidean space to the tangent space viewed as an inner product space under the Riemannian metric.

As it will become important later, one property to remember is the relationship of linear isometries between two spaces from proposition 8.1.9. In the context of sections this means that for any two sections $f, g \in O(M)_x$ we can find an orthogonal matrix U such that $Uf = g$.

2.1.8 Parallel Transport

Parallel transport is a topic of Riemannian geometry, which proposes a simple idea but requires extensive mathematical description. A parallel transport describes transporting a tangent vector S from one tangent space to another along a given curve according to an underlying Riemannian metric. It is concerned with differential equations, the so-called covariant derivative, connections, Christoffel symbols and more complex subjects. It is often viewed as a more theoretical concept, and analytic solutions to finding parallel transport often remain elusive.

The most prominent example is the so-called Levi-Civita parallel transport along a geodesic. This is also the parallel transport of interest in this thesis, so we will focus on describing this parallel transport without delving too deep into the mathematical theory and direct the interested reader to [11, Ch. 2] for a complete derivation. We will give an axiomatic definition of the Levi-Civita parallel transport along a geodesic. This definition is more approachable and comes without the notion of more advanced theoretical concepts than those already discussed.

We can describe the Levi-Civita parallel transport along a geodesic axiomatically. We will go through the necessary axioms individually and discuss their mathematical interpretation. At the end, we will give a concise definition.

First, we need to define the definition and image set of the function in question. We will denote it by $\mathrm{PT}_{A\to B} : T_AM \to T_BM$. This notation indicates that PT is a function that maps from the tangent space of point A to the tangent space of point B. We assume that the geodesic $\gamma_{A\to B}$ is unique.

Linear Isometry: We want PT to be a linear isometry between the two tangent spaces. This means that it needs to fulfill two properties. First, it needs to be a linear function, i.e., for two tangent vectors $X, Y \in T_AM$ we require

$$\mathrm{PT}_{\gamma_{A\to B}}(X+Y) = \mathrm{PT}_{\gamma_{A\to B}}(X) + \mathrm{PT}_{\gamma_{A\to B}}(Y)$$

to hold. Second, the isometry requirement indicates that the inner product of two tangent vectors $X, Y \in T_AM$ with the inner product given by the Riemannian metric tensor on T_AM is the same as the inner product of the parallel transported vectors on T_BM. Or, formally,

$$g_A(X,Y) = g_B(\mathrm{PT}_{\gamma_{A\to B}}(X), \mathrm{PT}_{\gamma_{A\to B}}(Y)).$$

These two properties define a linear isometry. As noted in proposition 8.1.9, linear isometries are unique up to orthogonal rotation. If we find an isometry, we can change it by multiplication with an orthogonal matrix and recover a valid linear isometry. Thus, this requirement of PT being a linear isometry gives us a potential group of functions where the orthogonal transformation determines the degrees of freedom.

Path Composition Consistency: We require PT to be consistent with the composition of geodesics. This means that if we have two geodesics $\gamma_{A\to B}$ and $\gamma_{B\to C}$ the compositions of parallel transports along the geodesics is the same as parallel transport along the composition of the paths. To describe this formally, we will once introduce a different notation and will identify $\mathrm{PT}_{A\to B} = \mathrm{PT}_{\gamma_{A\to B}}$. Now, the path composition consistency can be formalized as

$$\mathrm{PT}_{\gamma_{A\to B}\circ\gamma_{B\to C}} = \mathrm{PT}_{\gamma_{A\to B}} \circ \mathrm{PT}_{\gamma_{B\to C}}\,.$$

Preservation of Geodesics: Preservation of geodesics means that the parallel transported tangent vector of the geodesic at $\gamma(0) = A$ should be the same as the tangent vector at $\gamma(1) = B$. Or, formally, we require

$$\mathrm{PT}_{\gamma_{A\to B}}(\gamma'(0)) = \gamma'(1).$$

The requirement reduces the degrees of freedom of the permissible orthogonal transformations for linear isometries by one. This is because the orthogonal transformations that distinguish the linear isometries now only allow for transformations that rotate perpendicular to the tangent vector of the geodesic.

Minimal Rotation: The last requirement removes the last degrees of freedom and uniquely describes a single parallel transport PT. It is equivalent to the notion of the transport being so-called torsion-free. We again need to fall back onto the theory of unique solution of differential equations. Let V be a vector field that, contrary to the definition in definition 2.1.5, is parametrized by a scalar t in the way that we write

$$V(t) = V(\gamma_{A\to B}(t)).$$

Then we get that there exists a unique vector field that fulfills for a given vector S

$$V(0) = S \in T_A M, \text{ and } \frac{d}{dt}V(t) = 0.$$

This describes the condition that V introduces no extra rotation other than that given by the curvature along the path of the geodesic.

Definition 2.1.21. *Let $A, B, C \in M$ and $\gamma_{A\to B}, \gamma_{B\to C}$ the geodesic connecting A and B and B and C. The unique map $\mathrm{PT}_{\gamma_{A\to B}} : T_A M \to T_B M$ that fulfills the following conditions is called the Levi-Civita parallel transport along a geodesic:*

- ***Linear Isometry:***
  $g_A(X, Y) = g_B(\mathrm{PT}_{\gamma_{A\to B}}(X), \mathrm{PT}_{\gamma_{A\to B}}(Y))$ *and* $\mathrm{PT}_{\gamma_{A\to B}}(X + Y) = \mathrm{PT}_{\gamma_{A\to B}}(X) + \mathrm{PT}_{\gamma_{A\to B}}(Y)$
- ***Preservation of Geodesics:*** $\mathrm{PT}_{\gamma_{A\to B}}(\gamma'(0)) = \gamma'(1)$
- ***Path Composition Consistency:*** $\mathrm{PT}_{\gamma_{A\to B}\circ\gamma_{B\to C}} = \mathrm{PT}_{\gamma_{A\to B}} \circ \mathrm{PT}_{\gamma_{B\to C}}$
- ***Minimal Rotation Requirement:*** $\frac{d}{dt}\mathrm{PT}_{\gamma_{A\to B}(t)}(S) = 0$ *for any* $S \in T_A M$

Remark 2.1.22. *We denote the parallel transport function that fulfills the axioms of definition 2.1.21 by* $\mathrm{PT}^{\times}$.

2.2 Metric Spaces

The concept of a metric space provides a foundational framework for discussing and analyzing the notion of distance. Metric space theory is closely connected to the field of distance geometry. Distance geometry focuses on the study of geometric properties and relationships based on the concept of distance. It involves determining the positions of points in space given the distances between them. This field is particularly concerned with analyzing distance matrices and the embedding of metric spaces into Hilbert spaces. The embeddability of metric spaces into Hilbert spaces plays a key role in analyzing positive definite kernel functions. We will see later that we can describe Riemannian manifolds as metric spaces using the geodesic distance. This enables us to apply the theory of metric spaces regarding kernels to Riemannian manifolds.

Definition 2.2.1. *Let* M *be a set. A function* $d : \mathrm{M} \times \mathrm{M} \to \mathbb{R}$ *is called a metric on* M *if it fulfills the following properties for all points* $x, y, z \in \mathrm{M}$*:*

1. $d(x, x) = 0$
2. *if* $x \neq y$, $d(x, y) > 0$ *(positivity)*
3. $d(x, y) = d(y, x)$ *(symmetry)*
4. $d(x, y) + d(y, z) \geq d(x, z)$ *(triangle inequality)*

The pair (M, d) *is called metric space.*

The distance function d endows a space M with a metric structure. A metric space is called finite if the set M consists of a finite number of elements. For a given set M, we can define several metrics that potentially produce different properties of the resulting metric space. For a given metric d, there are functional changes that we can apply to it and still recover a valid metric.

In the study of metric spaces, special attention is given to a class of functions f that transform a given metric $d(a, b)$ by $f(d)$ so that $\tilde{d}(a, b) = f(d(a, b))$ is again a metric. This class of functions is called the metric transforms.

Definition 2.2.2. *[16] A function* $f : [0, \infty) \to [0, \infty)$ *is called metric transform if for all metric spaces* (M, d) $f(d)$ *is a metric.*

Metric transforms change the structure of the metric space and can transform the distance geometry of the space. It may be possible to embed a given transformed metric space into Hilbert space, a key point of metric spaces that is the topic of the next section.

2.2.1 Hilbert Space Embeddability

A large field of research in metric spaces and distance geometry is to determine whether a given metric space can be isometrically embedded in Hilbert space. This embeddability allows simplified analysis of the given space as the linear structure of Hilbert spaces can be transferred to the metric space. Additionally, this enables to transfer results for positive definite functions from Hilbert space to a given metric space.

Definition 2.2.3. *Let* $\mathfrak{H}$ *be an infinite dimensional Hilbert space. A metric space* (M, d) *is isometrically embeddable in* $\mathfrak{H}$ *if there exists an isometry* $\phi : \mathrm{M} \to \mathfrak{H}$ *such that for all* $x, y \in \mathrm{M}$ $d(x, y) = d_{\mathfrak{H}}(\phi(x), \phi(y)) = ||\phi(x) - \phi(y)||$.

The direct search for the isometry ϕ is usually unfeasible. We do not need to know the exact form of ϕ, but rather, it suffices to prove its existence. Thus, we can turn to other methods to show embeddability.

There is a general theorem for embeddability:

Theorem 2.2.4. *[48, Section 3] A separable metric space* (M, d) *is embeddable in* $\mathfrak{H} \iff$ *for any* $n + 1$ *points* $x_i \in M$ *we have*

$$\sum_{i,j=0}^{n} d(x_i, x_j)^2 c_i c_j \leq 0$$

for all c_i satisfying

$$\sum_{i=0}^{n} c_i = 0.$$

When working with discrete sets, it suffices to show that any number n of points in (M, d) can be isometrically embedded in $\mathfrak{H}$. This merits the definition of the so-called n-point property:

Definition 2.2.5. *[65] Let $\mathfrak{H}$ be an infinite dimensional Hilbert space. A metric space M has the n-point property if any n points from M are isometrically embeddable in $\mathfrak{H}$.*

By virtue of the triangle inequality, we note that every metric space has the three-point property by definition.

2.2.2 Positive Definite Functions

Schoenberg first developed the theory of positive definite functions [50], which is closely related to metric spaces and distance geometry. It is connected to positive definite kernels on metric spaces, as seen in section 5.3.

Definition 2.2.6. *[48, Eq. 3] Let (M, d) be a metric space. A real continuous function $f(t)$ which is defined on values of d is said to be positive definite in (M, d) if*

$$\sum_{i,j=1}^{n} f(d(x_i, x_j))c_i c_j \geq 0$$

for any n points $x_i \in \mathrm{M}$ for arbitrary real c_i.

There is a one-to-one mapping of metrics that are embeddable in Hilbert space to positive definite functions given by

Theorem 2.2.7. *[49] Let (M, d) be a metric space, f a function defined on values of d and $x_i, x_j \in \mathrm{M}$. $f(d(x_i, x_j))$ is embeddable in $\mathfrak{H} \iff e^{-\lambda f(d(x_i,x_j))^2}$ is a positive definite function in (M, d).*

Thus, finding a function f for a given metric d such that $f(d)$ is embeddable in $\mathfrak{H}$ allows us to directly construct a positive definite function from $f(d)$.

Two areas of research can be followed from here on: Firstly, we can assume the embeddability on d and find conditions on a function f such that $f(d)$ retains its embeddability. Secondly, we can start with a metric space (M, d) that is not embeddable and try to find a function f such that $(\mathrm{M}, f(d))$ is embeddable.

We are unaware of any general result in the second area of research, so we will focus on formalizing functions that retain embeddability. A comprehensive theorem describes the permissible functions f to retain embeddability. For this, let us introduce completely monotone functions of order q, extending the definition given by [49, Sec. 3] as done in [52]:

Definition 2.2.8. *[52, Def. 2.30] [49, Section 3.1] A real function $f(t)$ is said to be completely monotone (cm) of order q for $t > 0$ if*

$$(-1)^n \frac{d^n}{dt^n} f(t) \geq 0 \text{ for } 0 \leq t < \infty.$$

for all $n \geq q$ where $n = 0, 1, \ldots$ and

$$f(0) = \lim_{t \to 0} f(t)$$

There is a link between positive definite functions and completely monotone functions of order 0.

Theorem 2.2.9. *[49, Thm. 3] A function $f(t)$ is completely monotone of order 0 for $t \geq 0$ if and only if $f(d^2)$ is positive definite in Hilbert space.*

The relationship between completely monotonic functions and embeddable metric transforms is given as

Theorem 2.2.10. *[49, Thm. 6'] Let d be a metric in a Hilbert space. $f(d)$ is embeddable in Hilbert space if and only if $\frac{d}{dt} f^2(\sqrt{t})$ is completely monotone of order 0 for $t \geq 0$.*

We can thus describe permissible continuous metric transforms f in terms of completely monotone functions, assuming that d is already embeddable.

2.2.3 Riemannian Manifolds as Metric Spaces

Riemannian manifolds can be viewed as metric spaces: All connected Riemannian manifolds (M, g) give rise to a geodesic distance d (see definition 2.1.13). This results in a metric in the sense of metric spaces, i.e., (M, d) is a metric space [36, Thm. 2.55]. Thus, Riemannian manifolds form a subset of metric spaces and can be viewed as such; all metric space theory can be directly applied.

For Riemannian manifolds, the theory of embeddability can be viewed in the context of curvature. If a Riemannian manifold is globally isometric to Euclidean space, it is embeddable in Hilbert space. Non-flat Riemannian manifolds can not be embedded in Hilbert space.

3 Theory

In this section, we derive the underlying theory for formalizing kernel methods on Riemannian manifolds, building on the theory of sections of the orthonormal frame bundle. We construct a linear feature space that incorporates the linear structure provided by the Riemannian metric on the tangent space. We start by clarifying concepts from the general Riemannian geometry from section 2.1 and advance to build the theory. We then state all important properties of relevant Riemannian metrics on the manifold of symmetric positive definite matrices.

3.1 Concerning Geodesics

There are two different approaches to expressing a geodesic found in the literature. This might be confusing, but both concepts are directly related. The first is to define a geodesic in terms of a point P on the manifold and a tangent vector $V \in T_P M$. The so-defined geodesic describes the path on the manifold that a curve starting from P follows in the direction of a tangent vector V

$$\gamma_{P,V}(t) = \mathrm{Exp}_P(tV).$$

The second definition describes the path between two points $P, Q \in \mathcal{P}^n$. We can leverage the logarithmic map and show that these definitions are connected by $V = \mathrm{Log}_P(Q)$, i.e.

$$\gamma_{P \to Q}(t) = \mathrm{Exp}_P(t\,\mathrm{Log}_P(Q)).$$

The first definition is the standard way of describing geodesics in differential geometry. In some settings, we are more interested in the shortest paths between two points on the manifold. The second definition is more commonly used when working directly on the manifold, for example, in applied topics like machine learning with Brain-Computer Interfaces.

For the remainder of this thesis, we will assume that geodesics are defined between two points on a manifold unless explicitly stated otherwise.

3.2 Concerning Curvature

In this thesis, we only distinguish two classes of Riemannian manifolds: flat and non-flat (curved). As the curvature depends on the metric, we refer to the metrics resulting in a flat manifold as flat metrics. We can show that the resulting metric space is globally isometric to Euclidean space for flat metrics on the manifold of SPD matrices denoted by $\mathcal{P}^n$.

Lemma 3.2.1. *Let $(\mathcal{P}^n, g)$ be a flat Riemannian metric on the SPD matrices $\mathcal{P}^n$. Then it is globally isometric to Euclidean space.*

Proof. $\mathcal{P}^n$ is a simply connected open subset of Euclidean space. Thus, the local isometry extends to a global isometry by the Cartan-Ambrose-Hicks theorem [44]. □

This means that in the context of SPD matrices, we can construct every flat manifold as a pullback metric (see definition 2.1.17) from standard Euclidean space using a global isometry ϕ.

3.3 Concerning Sections

For any Riemannian manifold $(\mathcal{P}^n, g)$, every section $f_P : (\mathbb{R}^{\frac{n(n+1)}{2}}, <.,.>) \to (T_P\mathcal{P}^n, g_P)$ of the orthonormal frame bundle $O_P(\mathcal{P}^n)$ at any point $P \in \mathcal{P}^n$ is a global isometry between standard Euclidean space $\mathbb{R}^{\frac{n(n+1)}{2}}$ and the tangent space $(T_P\mathcal{P}^n, g_P)$.

The general use of sections is to leverage this isometry and state the Riemannian metric on each tangent space in terms of the standard Euclidean inner product.

Proposition 3.3.1. *Let g be a Riemannian metric on $\mathcal{P}^n$. Then there exists a family of functions f_P parametrized by $P \in \mathcal{P}^n$ such that $g_P(a,b) =< f_P^{-1}(a), f_P^{-1}(b) >$ where $<.,.>$ is the standard Euclidean inner product.*

Proof. f can be written as $f : \mathcal{P}^n \to O(\mathcal{P}^n)$ with the property that $f_P : \mathbb{R}^{\frac{n(n+1)}{2}} \to T_P\mathcal{P}^n$, i.e., in every point P the map f gives an isometry from $\mathbb{R}^{\frac{n(n+1)}{2}}$ to $T_P\mathcal{P}^n$. As an open subset of the Euclidean space of symmetric matrices that is contractible, the manifold of SPD matrices admits a global section f. As f is a global isometry, its inverse is well-defined. □

We can construct the metric simply as the standard Euclidean inner product by identifying any section of the orthonormal frame bundle of any given Riemannian metric g. We note that by proposition 8.1.9, f is not uniquely defined.

Remark 3.3.2. *For every f for which $g_P(a,b) =< f_P^{-1}(a), f_P^{-1}(b) >$ holds, we can construct $\tilde{f}^{-1} = T(f^{-1})$, where T is an orthogonal transformation (see definition 8.1.7), and still recover the Riemannian metric, i.e., $g_P(a,b) =< f_P^{-1}(a), f_P^{-1}(b) >=< \tilde{f}_P^{-1}(a), \tilde{f}_P^{-1}(b) >$.*

The tangent space of a Riemannian metric on $\mathcal{P}^n$ is commonly described as the set of symmetric matrices $\mathcal{S}^n$ and the Riemannian metric on the tangent space is given in terms of the Frobenius inner product $<.,.>^F$ cf. e.g. [8, 15].

To translate this identification to sections, we introduce the notion of symmetric sections.

Definition 3.3.3. *Let g be a Riemannian metric on $\mathcal{P}^n$ and let $f_P \in O_P(\mathcal{P}^n)$. Then we define the symmetric section as $f_P^{sym} : (\mathcal{S}^n, <.,.>^F) \to (T_P\mathcal{P}^n, g_P)$ mapping between the symmetric matrices $\mathcal{S}^n$ and $T_P\mathcal{P}^n$ by identifying the tangent space with the symmetric matrices $\mathcal{S}^n$.*

A section and its symmetric counterpart produce the same Riemannian metric, i.e., $g_P(a,b) =< f_P^{-1}(a), f_P^{-1}(b) >=< (f_P^{sym})^{-1}(a), (f_P^{sym})^{-1}(b) >^F$ by the same argument that lets us identify the tangent space of metrics on SPD matrices with symmetric matrices.

Regarding sections and orthogonal transformations, identifying $T_P\mathcal{P}^n$ with symmetric matrices reaches its limits. Although the description is equivalent in terms of the result of the inner product, there are differences

in how we describe orthogonal transformations for the inner products (see proposition 8.1.28). By formulating the Riemannian metric on $T_P\mathcal{P}^n$ in terms of matrices, we assume the possible orthogonal transformations T of an $n\times n$-dimensional symmetric matrix A to be of the type $T(A) = UAU^T$, where $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix. We denote these orthogonal transformations by T^{sym}, see definition 8.1.27. However, as shown in proposition 8.1.28, these symmetric orthogonal transformations only cover a subset of possible orthogonal transformations on a vector $v \in \mathbb{R}^{\frac{n(n+1)}{2}}$ which are given by $T(v) = Uv$, where $U \in \mathbb{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}$ is again an orthogonal matrix, but of a larger dimensionality. Thus, by describing sections in terms of symmetric matrices, we implicitly limit the possible orthogonal transformations that the inner product allows.

Except for the difference in possible orthogonal transformations, there is no formal difference in the description of a section and its symmetric counterpart in this thesis. To simplify notation and adhere to the tangent space identification with the symmetric matrices, we always assume a section f to be symmetric and omit writing f^{sym} unless explicitly stated otherwise.

3.4 Concerning Parallel Transport

As introduced in section 2.1.8, parallel transport can be used to transport a set of points from one tangent space T_PM from any given reference point $P \in M$ along a curve γ towards the tangent space of a new reference point Q. In applied fields, parallel transport is often used for domain adaptation to find a common reference tangent space to compare data from different sources. Thus, it is advantageous to have an analytic description of parallel transport.

The parallel transport axiomatically described in section 2.1.8 denoted by $\mathrm{PT}^{\times}$ often lacks an analytic expression.

In the following, we will show how this can be facilitated using the concept of sections.

Proposition 3.4.1. *Let $(\mathcal{P}^n, g)$ be a Riemannian manifold. For any $f_Q \in O(\mathcal{P}^n)_Q$ and $f_P \in O(\mathcal{P}^n)_P$, the function $\Phi_{Q,P}$ defined by*

$$\Phi_{Q,P}(V) = f_P(f_Q^{-1}(V))$$

is a linear isometry between $(T_Q\mathcal{P}^n, g_Q)$ and $(T_P\mathcal{P}^n, g_P)$.

Proof. $O(\mathcal{P}^n)_Q$ and $O(\mathcal{P}^n)_P$ consist of all linear isometries between $T_Q\mathcal{P}^n$ or $T_P\mathcal{P}^n$ to $\mathbb{R}^{\frac{n(n+1)}{2}}$. Because $(T_Q\mathcal{P}^n, g_Q)$ and $(T_P\mathcal{P}^n, g_P)$ are both linearly isometric to $(\mathbb{R}^{\frac{n(n+1)}{2}}, <.,.>)$, $(T_Q\mathcal{P}^n, g_Q)$ and $(T_P\mathcal{P}^n, g_P)$ are also linearly isometric. □

By this simple proposition, which only uses the definition of the orthonormal frame bundle and the concatenation of isometries, we have constructed a candidate function for parallel transport which is an isometry between tangent spaces.

Proposition 3.4.2. *Let $(\mathcal{P}^n, g)$ be a Riemannian manifold. Then there exist $f_Q^{\times} \in O(\mathcal{P}^n)_Q$ and $f_P^{\times} \in O(\mathcal{P}^n)_P$ such that for the concatenation of sections $\Phi_{Q,P}$ defined by*

$$\Phi_{Q,P}(V) = f_P^{\times}((f_Q^{\times})^{-1}(V)),$$

we have $\Phi_{Q,P}(V) = \mathrm{PT}^{\times}_{Q\to P}$.

Proof. Because $(T_Q\mathcal{P}^n, g_Q)$ and $(T_P\mathcal{P}^n, g_P)$ are both linearly isometric to $(R^k, <\ .,.\ >)$, $(T_Q\mathcal{P}^n, g_Q)$ and $(T_P\mathcal{P}^n, g_P)$ are also linearly isometric. $O(\mathcal{P}^n)_Q$ and $O(\mathcal{P}^n)_P$ include all isometries. Per definition, $\mathrm{PT}^{\times}_{Q\to P}$ is a linear isometry between tangent spaces. Then for any $f_P \in O(\mathcal{P}^n)_P$, $f_Q^{-1} := \mathrm{PT}^{\times}_{Q\to P} \circ f_P^{-1}$ is a linear isometry from $T_Q\mathcal{P}^n$ to $\mathbb{R}^{\frac{n(n+1)}{2}}$. Evidently, $f_Q \in O(\mathcal{P}^n)_Q$. It is clear that we can now write $\mathrm{PT}^{\times}_{Q\to P} = f_Q^{-1} \circ f_P$. Thus, $\mathrm{PT}^{\times}_{Q\to P}$ can be constructed as stated above for correct choices of f_P and f_Q. □

Proposition 3.4.2 shows that it is possible to choose a candidate function that defines the metric from the orthonormal frame bundle and try to find the orthogonal transformation that constructs $\mathrm{PT}^{\times}$. Given a metric, there is usually an evident candidate function. It remains to find the corresponding orthogonal transformation to get the parallel transport $\mathrm{PT}^{\times}$. The following lemma is stated for completeness:

Lemma 3.4.3. *Let $\Phi_{Q,P}(V)$ as defined in proposition 3.4.1. Let* vech *be defined as the operator extracting a vector from a symmetric matrix as defined in definition 8.1.23. Then there exists an orthogonal matrix U such that* $\mathrm{vech}^{-1}(U\,\mathrm{vech}(\Phi_{Q,P}(V))) = \mathrm{PT}^{*}_{Q\to P}$.

Proof. $\Phi_{Q,P}(t)(V)$ and $\mathrm{PT}^{*}_{Q\to P}$ are linear isometries between $(T_Q\mathcal{P}^n, g_Q)$ and $(T_P\mathcal{P}^n, g_P)$. Orthogonal matrices represent orthogonal transformations on vector spaces. By proposition 8.1.9 linear isometries are unique up orthogonal transformations. $\mathrm{vech}^{-1}(U\,\mathrm{vech}(\Phi_{Q,P}(t)(V)))$ denotes the orthogonal transformation on the $\frac{n(n+1)}{2}$-dimensional vector obtained from the transported V, concluding the proof. □

Theorem 3.4.4. *Let f be the sections such that $\mathrm{PT}^{\times}_{P\to I} = f_I(f_P^{-1})$ and $\mathrm{PT}^{\times}_{Q\to I} = f_I(f_Q^{-1})$. Then $f_Q(f_P^{-1}) = \mathrm{PT}^{\times}_{\gamma_{P\to I}\circ\gamma_{I\to Q}}$ is the parallel transport $\mathrm{PT}^{\times}$ along the joint geodesics from P to I and from I to Q.*

Proof. The path composition axiom in the definition of $\mathrm{PT}^{\times}$ provides the proof. □

The fact that the concatenation of any two sections is an isometry between tangent spaces is an isometry gives a direct relation to $\mathrm{PT}^{\times}$. We only need an orthogonal transformation to align the concatenated sections and $\mathrm{PT}^{\times}$. This connection between sections and $\mathrm{PT}^{\times}$ may facilitate finding analytic expressions for $\mathrm{PT}^{\times}$.

3.5 Manifold Transport

In some settings, it may be desirable to transport points on the manifold from one place to another rather than between tangent spaces (see, e.g., [68, 47, 66]). This is sometimes also referred to as parallel transport. To avoid confusion and to be able to distinguish between the two concepts, we will introduce the notion of manifold transport, similar to the map Ψ defined in [66, Eq. 8]:

Definition 3.5.1. *Let $\mathrm{PT}_{A\to B}(X)$ be a parallel transport of a tangent vector X from T_AM to T_BM. Then $\mathrm{MT}_{A\to B}(P) := \mathrm{Exp}_B(\mathrm{PT}_{A\to B}(\mathrm{Log}_A(P)))$ defines the manifold transport of $P \in M$ from A to B according to an underlying parallel transport* PT.

If the underlying parallel transport is not mentioned explicitly, we assume $\mathrm{PT}^{\times}$.

3.6 Riemannian Linear Features

We established that the tangent space of a differentiable manifold is a linear space in section 2.1.1. Points on a Riemannian manifold (M, g) are mapped to the tangent space using the logarithmic map. Thus, we can view the logarithmic map as a way to create a linear feature from a point on the manifold. The notion of this linear tangent space in this context is often replaced by the common Euclidean space under the inner product structure of the Frobenius inner product [24, 14, 18]. This means that distances and inner product on the tangent space are evaluated using the standard Euclidean inner product. Although this approach is not wrong, it neglects the Riemannian structure of the tangent space. The tangent space is a finite-dimensional inner product space, thus complete and isomorphic to Euclidean space, i.e., has a linear structure. However, the linear structure of $T_P M$ is given by the specific inner product at this point that corresponds to the Riemannian metric g_P and not by the standard Euclidean inner product.

We aim to describe a linear feature map that incorporates the Riemannian linear structure of the tangent space $T_P M$. This linear map can be found for all Riemannian metrics on SPD matrices and relies on sections of the orthonormal frame bundle, see section 3.3. We call this map the Riemannian linear feature map and the resulting linearized vectors are the Riemannian linear features, denoted by RLF.

Definition 3.6.1. *Let $(\mathcal{P}^n, g)$ be a Riemannian manifold. Let f_P be a section of the orthonormal frame bundle at P, such that $g_P(A, B) =< f_P^{-1}(A), f_P^{-1}(B) >$. Let $\mathrm{Log}_P(Q)$ be the corresponding logarithmic map at point P. Then $\mathrm{RLF}_P(Q) := f_P^{-1}(\mathrm{Log}_P(Q))$ defines the Riemannian Linear Feature of a point $Q \in \mathcal{P}^n$ parametrized by $P \in \mathcal{P}^n$.*

Remark 3.6.2. *$\mathrm{RLF}_P(Q)$ describes a feature map that maps from M to Euclidean space under the standard Euclidean inner product. Its inverse is given by $\mathrm{RLF}_P^{-1}(S) = \mathrm{Exp}_P(f_P(S))$.*

This space has been implicitly described for the Affine Invariant metric in [3] and a description of the map has been given in [62]. However, it has not been given a general name or explicitly been defined as a space of interest for generic Riemannian metrics on $\mathcal{P}^n$.

Remark 3.6.3. *We call the space described by Riemannian linear features the Riemannian linear feature space.*

In the definition of Riemannian linear features, we did not specify the choice of the section f_P to construct the space. Any choice of section produces a valid map to standard Euclidean linear space. As the sections are linear isometries and thus unique up to orthogonal transformation (see proposition 8.1.9), the Riemannian linear feature spaces under any of the sections are related by the same transformation. Let f_P and $\tilde{f}_P$ be two sections that give rise to the Riemannian linear features RLF_P and $\tilde{\mathrm{RLF}}_P$ respectively. Then for any point $A \in \mathcal{P}^n$ we can find an orthogonal transformation T such that $\mathrm{RLF}_P(A) = T(\tilde{\mathrm{RLF}}_P(A))$. Thus, Riemannian linear features inherit the uniqueness property up to orthogonal transformations from the underlying sections. Depending on the goal, this gives independence of the section f to construct the Riemannian linear feature space, as the inner products and distances between vectors on Riemannian linear feature space are the same for all sections. We can thus choose the most apparent section and construct a genuinely standard Euclidean space on any point $P \in \mathcal{P}^n$.

Similar to constructing symmetric sections, we can choose whether to construct the Riemannian linear features as vectors or symmetric matrices. Again, the only difference is in permissible orthogonal transformations

on the features. As we assume sections to map to symmetric matrices, we will assume that the underlying Riemannian linear features are also described in terms of symmetric matrices unless explicitly noted otherwise.

We can connect Riemannian linear features given a specific section with tangent space features parallel transported to the identity by the following lemma:

Lemma 3.6.4. *Let f be the section such that $\mathrm{MT}_{P\to I}(X) = \mathrm{Exp}_I(f_I(f_P^{-1}(\mathrm{Log}_P(X))))$ is the manifold transport from P to I under the parallel transport $\mathrm{PT}^{\times}$ and let $\mathrm{RLF}_P(X)$ the Riemannian linear feature of X at P using the same section f. Then $f_I(\mathrm{RLF}_P(X))$ is the tangent vector of the manifold transported X from P to the identity.*

Proof.

$$\begin{aligned}\mathrm{Log}_I(\mathrm{MT}_{P\to I}(X)) &= \mathrm{Log}_I(\mathrm{Exp}_I(f_I(f_P^{-1}(\mathrm{Log}_P(X))))) && (3.6.5)\\ &= f_I(f_P^{-1}(\mathrm{Log}_P(X))) && (3.6.6)\\ &= f_I(\mathrm{RLF}_P(X)) && (3.6.7)\end{aligned}$$

□

Corollary 3.6.8. *If $f_I^{-1}(X) = f_I(X) = X$, Riemannian linear features and tangent vectors of data points manifold transported to the identity coincide.*

We can use the Riemannian linear feature space as a common linear space onto which we can project data points from various points of the manifold. Assuming two data clouds $\{X_i\}_{i=1,\dots,n}$ and $\{\tilde{X}_i\}_{i=1,\dots,n}$ centered around P and $\tilde{P}$ respectively, we can construct the Riemannian linear feature spaces RLF_P and $\mathrm{RLF}_{\tilde{P}}$. As these spaces are standard Euclidean space, we can directly compare the data points from both spaces. In this space, we can use any known techniques from Euclidean space to align both datasets, e.g., whitening. In contrast to the tangent space at the identity identified with the symmetric matrices, the so-constructed feature space allows for a larger class of orthogonal transformations, see proposition 8.1.28.

The following lemma shows that Riemannian linear features and tangent space features at the Frechét mean are centered, which will become relevant later.

Lemma 3.6.9. *Let $\{X_i\}_{i=1,\dots,n}$ be a dataset on $\mathcal{P}^n$ with mean P. Let $||.||_A$ be the norm induced by the Riemannian metric on the tangent space of $A \in \mathcal{P}^n$. Let d_E denote the Euclidean distance. Then $\sum_i \mathrm{RLF}_P(X_i) = 0$.*

Proof. By the identification $d(A,B) = ||\mathrm{Log}_A(B)||_A$ we have

$$\begin{aligned}d(A,B) &= ||\mathrm{Log}_A(B)||_A\\ &= g_A(\mathrm{Log}_A(B), \mathrm{Log}_A(B))^{\frac{1}{2}}\\ &= <\mathrm{RLF}_A(B), \mathrm{RLF}_A(B)>^{\frac{1}{2}}\\ &= ||\mathrm{RLF}_A(B)||\\ &= d_E(0, \mathrm{RLF}_A(B))\end{aligned}$$

and $\mathrm{Log}_A(A) = 0$. Thus, if P is the mean of $\{X_i\}_{i=1,\dots,n}$, $\mathrm{Log}_P(P) = 0$ is the mean of the Riemannian linear features. □

Corollary 3.6.10. *Let $\{X_i\}_{i=1,\dots,n}$ be a dataset on M with mean P. Then $\sum_i \operatorname{Log}_P(X_i) = 0$.*

Proof. Let f be any section. As f is linear, $f(0) = 0$. Thus, we have

$$\begin{aligned}
\sum_i \operatorname{RLF}_P(X_i) &= 0 \\
f_P\left(\sum_i \operatorname{RLF}_P(X_i)\right) &= f_P(0) \\
\sum_i f_P(\operatorname{RLF}_P(X_i)) &= 0 \\
\sum_i f_P(f_P^{-1}(\operatorname{Log}_P(X_i))) &= 0 \\
\sum_i \operatorname{Log}_P(X_i) &= 0
\end{aligned}$$

□

We can use Riemannian linear features to construct metric spaces that can be isometrically embedded in Hilbert space for arbitrary metrics. Equipped with the standard Euclidean inner product, the induced metric for Riemannian linear features (see remark 8.1.3) is

$$d_P(X_i, X_j) = ||\operatorname{RLF}_P(X_i) - \operatorname{RLF}_P(X_j)||.$$

This metric depends on the reference point P.

3.7 Riemannian Metrics for Symmetric Positive Definite Matrices

This section will present common Riemannian metrics on the manifold of symmetric positive definite matrices $\mathcal{P}^n$ and describe the Riemannian linear features for each.

3.7.1 Euclidean

The Euclidean metric is the Riemannian metric corresponding to the Euclidean inner product on the tangent space of the manifold of positive definite matrices.

It is considered too simple to merit an explicit derivation. As it forms the basis of many pullback metrics for SPD matrices, we will state all important properties here.

Equipping the differentiable manifold of SPD matrices with this metric has several disadvantages. The resulting Riemannian manifold is not geodesically complete meaning that geodesics cannot be extended infinitely and still recover an SPD matrix, cf. definition 2.1.12. This has implications for optimization algorithms working on geodesics. Additionally, the Euclidean metric suffers from the so-called swelling effect. This means that the determinant of the average SPD matrix is larger than the determinants of any of the original matrices being averaged. The effect is particularly problematic when working with covariance matrices, as it artificially inflates the overall variance. For properties of the Euclidean metric, see table 3.1.

Euclidean	
Inner product	$g_P(A, B) =< A, B >_F$
Section	$f_P^{-1}(A) = A$
Geodesic	$\gamma_{A\to B}(t) = (1-t)A + tB$
Exponential Map	$\mathrm{Exp}_P(S) = S + P$
Logarithmic Map	$\mathrm{Log}_P(Q) = Q - P$
Distance	$d(P, Q) = \|\|P - Q\|\|$
Parallel Transport	$\mathrm{PT}_{P\to Q}(X) = X - P + Q$
Linear Features	$RLF_P(X) = X - P$

Table 3.1: **Euclidean Metric Properties.**

3.7.2 Pullback Metrics

As noted in definition 2.1.17, pullback metrics allow to construct Riemannian metrics on a differentiable manifold M using an isometry $\phi : M \to N$ that maps from a Riemannian manifold (N, g) to M, creating a Riemannian manifold $(M, \phi * g)$. Pullback metrics transfer properties of the manifold leveraging this isometry $\phi : M \to N$ and its differential $[\mathrm{D}_P\,\phi]$. The inverse differential is given as $[\mathrm{D}_P\,\phi]^{-1} = [\mathrm{D}_{\phi(P)}\,\phi^{-1}]$. By theorem 8.1.37, we can efficiently compute the differential for pullback metrics for SPD matrices. We label the properties of the pullback metric $(M, \phi * g)$ with an asterisk $*$. If no asterisk is shown, the property on the original Riemannian manifold (N, g) is indicated. See table 3.2 for properties of pullback metrics.

Proposition 3.7.1. *Let f_X denote a section of the orthonormal frame bundle $O(N)_X$ on the Riemannian manifold N. The manifold transport of a point $A \in M$ from $P \to Q, P, Q \in M$ for pullback metrics is*

$$\mathrm{MT}^*_{P\to Q}(A) = \phi^{-1}(\mathrm{MT}_{\phi(P)\to\phi(Q)}(\phi(A)))$$

Proof.

$$\begin{aligned}
\mathrm{MT}^*_{P\to Q}(A) &= \mathrm{Exp}^*_Q([\mathrm{D}_Q\,\phi]^{-1}(f_{\phi(Q)}(f^{-1}_{\phi(P)}([\mathrm{D}_P\,\phi](\mathrm{Log}^*_P(A))))))\\
&= \phi^{-1}(\mathrm{Exp}_{\phi(Q)}([\mathrm{D}_Q\,\phi][\mathrm{D}_Q\,\phi]^{-1}(f_{\phi(Q)}(f^{-1}_{\phi(P)}([\mathrm{D}_P\,\phi][\mathrm{D}_P\,\phi]^{-1}\,\mathrm{Log}_{\phi(P)}(\phi(A))))))))\\
&= \phi^{-1}(\mathrm{Exp}_{\phi(Q)}(f_{\phi(Q)}(f^{-1}_{\phi(P)}(\mathrm{Log}_{\phi(P)}(\phi(A))))))\\
&= \phi^{-1}(\mathrm{MT}_{\phi(P)\to\phi(Q)}(\phi(A)))
\end{aligned}$$

□

Lemma 3.7.2. *For pullback metrics constructed on M from the Euclidean metric, manifold transport preserves the distances on the manifold, i.e., for $A, B \in M$ $d_*(A, B) = d_*(\mathrm{MT}^*_{P\to Q}(A), \mathrm{MT}^*_{P\to Q}(B))$.*

Proof. Let $\phi : M \to \mathbb{R}^n$ be the pullback map. For the Euclidean metric we have $f_P^{-1}(A) = A$. With the

properties given in table 3.1 we have

$$\begin{aligned}
&d_*(\mathrm{MT}^*_{P\to Q}(A), \mathrm{MT}^*_{P\to Q}(B)) \\
=&||\phi(\phi^{-1}(\phi(Q) + \phi(A) - \phi(P))) - \phi(\phi^{-1}(\phi(Q) + \phi(B) - \phi(P)))|| \\
=&||\phi(Q) + \phi(A) - \phi(P) - (\phi(Q) + \phi(B) - \phi(P))|| \\
=&||\phi(A) - \phi(B)|| \\
=&d_*(A, B)
\end{aligned}$$

□

Lemma 3.7.3. *For pullback metrics constructed on M from the Euclidean metric, the Euclidean distance between Riemannian linear features corresponds to the distance of the points on the manifold, i.e., for $A, B \in M$ we have $d(\mathrm{RLF}^*_P(X), \mathrm{RLF}^*_P(Y)) = d(\phi(X), \phi(Y))$.*

Proof. For the Euclidean metric we have $f_P^{-1}(A) = A$.

$$\begin{aligned}
&d(\mathrm{RLF}^*_P(X), \mathrm{RLF}^*_P(Y)) \\
=&\sqrt{< f^{-1}_{\phi(P)}(\mathrm{Log}_{\phi(P)}(\phi(X))) - f^{-1}_{\phi(P)}(\mathrm{Log}_{\phi(P)}(\phi(Y))), f^{-1}_{\phi(P)}(\mathrm{Log}_{\phi(P)}(\phi(X))) - f^{-1}_{\phi(P)}(\mathrm{Log}_{\phi(P)}(\phi(Y))) >} \\
=&\sqrt{< \mathrm{Log}_{\phi(P)}(\phi(X)) - \mathrm{Log}_{\phi(P)}(\phi(Y)), \mathrm{Log}_{\phi(P)}(\phi(X)) - \mathrm{Log}_{\phi(P)}(\phi(Y)) >} \\
=&\sqrt{< \phi(X) - \phi(P) - (\phi(Y) - \phi(P)), \phi(X) - \phi(P) - (\phi(Y) - \phi(P)) >} \\
=&\sqrt{< \phi(X) - \phi(Y), \phi(X) - \phi(Y) >} \\
=&d(\phi(X), \phi(Y))
\end{aligned}$$

□

Note that the previous lemmas hold independently of the points P and Q. These features are an inherent property of flat metrics and hold by definition.

3.7.2.1 Log-Euclidean

The Log-Euclidean framework introduced by [2] is a pullback metric using the isometry $\phi(X) = \mathrm{Log}(X)$. Although the original paper uses a different approach, the metric can be derived as a pullback metric from the Euclidean metric using the matrix logarithm function as the isometry. For the Log-Euclidean metric, $f_I^{-1}(X) = f_I(X) = X$ thus by corollary 3.6.8 the Riemannian linear features correspond to the tangent space features of points that are transported to the identity using manifold transport. See table 3.3 for properties of the Log-Euclidean metric.

3.7.2.2 Log-Cholesky

The Log-Cholesky metric derived in [40] is a pullback metric but was not constructed as a pullback metric directly from the Euclidean metric for SPD matrices. The metric was first constructed on the space of lower triangular matrices and then transferred to the space of SPD matrices using the Cholesky decomposition. We will give a brief overview of the construction and the relevant properties here but refer the reader to [40] for a complete derivation:

Pullback metrics	
Inner Product	$g^*_P(A,B) = g_{\phi(P)}([\mathrm{D}_P\,\phi](A), [\mathrm{D}_P\,\phi](B))$
Section	$f^{-1}{}^*_P(A) = f^{-1}_{\phi(P)}([\mathrm{D}_P\,\phi](A))$
Geodesic	$\gamma^*_{A\to B}(t) = \phi^{-1}(\gamma_{\phi(A)\to \mathrm{Log}_{\phi(A)}(\phi(B))}(t))$
Exponential Map	$\mathrm{Exp}^*_P(S) = \phi^{-1}(\mathrm{Exp}_{\phi(P)}([\mathrm{D}_P\,\phi](S)))$
Logarithmic Map	$\mathrm{Log}^*_P(Q) = [\mathrm{D}_P\,\phi]^{-1}(\mathrm{Log}_{\phi(P)}(\phi(q)))$
Distance	$d_*(P,Q) = \|\|\,\mathrm{Log}_{\phi(P)}(\phi(Q))\|\|_{\phi(P)} = d(\phi(P),\phi(Q))$
Parallel Transport	$\mathrm{PT}^*_{P\to Q}(S) = [\mathrm{D}_Q\,\phi]^{-1}(f_{\phi(Q)}(f^{-1}_{\phi(P)}([\mathrm{D}_P\,\phi](S))))$
Manifold Transport	$\mathrm{MT}^*_{P\to Q}(A) = \phi^{-1}(\mathrm{Exp}_{\phi(Q)}(f_{\phi(Q)}(f^{-1}_{\phi(P)}(\mathrm{Log}_{\phi(P)}(\phi(A))))))$
Linear Features	$\mathrm{RLF}^*_P(A) = f^{-1}_{\phi(P)}(\mathrm{Log}_{\phi(P)}(\phi(A)))$

Table 3.2: **Pullback Metric Properties.** For this table, let (N, g) denote a Riemannian manifold, M a differential manifold and let $\phi : M \to N$ denote an isometry. $[\mathrm{D}_P\,\phi]$ denotes the differential of ϕ at P. See Table 2 of [61] for reference.

Log-Euclidean	
Inner Product	$g_P(A,B) =< [\mathrm{D}_P\,\mathrm{Log}](X), [\mathrm{D}_P\,\mathrm{Log}](Y)) >$
Section	$f_P^{-1}(A) = [\mathrm{D}_P\,\mathrm{Log}](A)$
Geodesic	$\gamma_{A\to B}(t) = \mathrm{Exp}\left((1-t)\,\mathrm{Log}(A) + t\,\mathrm{Log}(B)\right)$
Exponential Map	$\mathrm{Exp}_P(S) = \mathrm{Exp}(\mathrm{Log}(P) + [\mathrm{D}_P\,\mathrm{Log}](S))$
Logarithmic Map	$\mathrm{Log}_P(Q) = [\mathrm{D}_{\mathrm{Log}(P)}\,\mathrm{Exp}](\mathrm{Log}(Q) - \mathrm{Log}(P))$
Distance	$d(P,Q) = \|\|\,\mathrm{Log}(P) - \mathrm{Log}(Q)\|\|$
Parallel Transport	$\mathrm{PT}^{\times}_{P\to Q}(S) = [\mathrm{D}_{\mathrm{Log}(Q)}\,\mathrm{Exp}]([\mathrm{D}_P\,\mathrm{Log}](S))$
Manifold Transport	$f_{P\to Q}(A) = \mathrm{Exp}(\mathrm{Log}(A) - \mathrm{Log}(P) + \mathrm{Log}(Q))$
Linear Features	$RLF_P(X) = \mathrm{Log}(X) - \mathrm{Log}(P)$

Table 3.3: **Log-Euclidean Metric Properties.**

Let $\mathfrak{L}$ denote the space of lower triangular matrices. It is a differentiable manifold of dimension $\frac{n(n+1)}{2}$. Let $\mathfrak{L}^+$ denote the space of lower triangular matrices with positive entries on the diagonal. As an open subset of $\mathfrak{L}$, it is again a differentiable manifold.

Let $\mathcal{L}$ denote the Cholesky decomposition of a SPD matrix. It is known to exist and be unique. Let $\mathcal{L}^{-1}$ denote its inverse, that is $\mathcal{L}^{-1}(L) = LL^T$. $\mathcal{L}$ is a diffeomorphism between $\mathcal{P}^n$ and $\mathfrak{L}^+$. Define $S_{\frac{1}{2}} = \lfloor S \rfloor + \frac{1}{2}\mathbb{D}(S)$.

Let $\mathbb{D}(.)$ denote the operator that extracts the diagonal part of a matrix and $\lfloor . \rfloor$ denote the operator that extracts the lower triangular part below the diagonal. The idea is now to split a lower triangular matrix L into two parts: $\lfloor L \rfloor$] and $\mathbb{D}(L)$. Matrices in $\mathfrak{L}^+$ have only positive values on the diagonal and otherwise unrestricted elements below the diagonal, $\mathfrak{L}^+$ can be parameterized by a function ϕ such that

$$\phi(X) = \begin{cases} X_{ij} & i \neq j \\ \exp(X_{ij}) & i = j \end{cases}$$

Adhering to this split, an inner product $\tilde{g}_L$ is defined on tangent space $T_L\mathfrak{L}^+$, identified with $\mathfrak{L}$ by considering both parts of a lower triangular matrix independently

$$\tilde{g}_L(X, Y) = < \lfloor X \rfloor, \lfloor Y \rfloor > + < \mathbb{D}(L)^{-1}\mathbb{D}(X), \mathbb{D}(L)^{-1}\mathbb{D}(Y) >$$

Thus, we have defined a Riemannian metric on $\mathfrak{L}^+$ and constructed a Riemannian manifold $(\mathfrak{L}^+, \tilde{g})$. The next step is to transfer this structure to the set of symmetric positive definite matrices. This is done via the pullback $\mathcal{L}$, the Cholesky decomposition.

So the Log-Cholesky metric on $\mathcal{P}^n$ is defined as

$$g_P(X, Y) = \tilde{g}_{\mathcal{L}(P)}([\mathrm{D}_P\,\mathcal{L}](X), [\mathrm{D}_P\,\mathcal{L}](Y))$$

where $[\mathrm{D}_P\,\mathcal{L}](X) = \mathcal{L}(P)(\mathcal{L}(P)^{-1}X\mathcal{L}(P)^{-T})_{\frac{1}{2}}$.

All other properties are transferred in the same way as for general pullback metrics.

The Log-Cholesky metric is flat, i.e., isometric to the Euclidean metric for SPD matrices. The isometry is not given in the original publication but can be constructed as

$$\phi(X)_{ij} = \begin{cases} \log(\mathcal{L}(X)_{ij}) & \text{for } i = j \\ \mathcal{L}(X)_{ij} & \text{for } i \neq j \end{cases} \tag{3.7.4}$$

Note that the resulting matrix is lower triangular, so the Euclidean metric from this space is not pulled back to the SPD matrices from the symmetric matrices but rather from the lower triangular matrices. The inverse of the isometry can be constructed as

$$\phi^{-1}(X) = CC^T, C = \begin{cases} \exp(X_{ij}) & \text{for } i = j \\ X_{ij} & \text{for } i \neq j \end{cases} \tag{3.7.5}$$

ϕ is not a matrix function because it does not act on the eigenvalues after eigendecomposition and its expression cannot be simplified as, e.g., the matrix logarithm for SPD matrices.

The metric does not suffer from the swelling effect and the mean has an analytic solution [40]. See table 3.4 for properties of the Log-Cholesky metric.

Log-Cholesky	
Inner Product	$g_P(A,B) = \tilde{g}_{\mathcal{L}(P)}([\mathrm{D}_P\,\mathcal{L}](A), [\mathrm{D}_P\,\mathcal{L}](B))$
Geodesic	$\gamma_{A\to B}(t) = C(t)C(t)^T, C(t) = (1-t)\mathcal{L}(A) + t\mathcal{L}(B) + \mathbb{D}(\mathcal{L}(A))^{1-t}\mathbb{D}(\mathcal{L}(B))^t$
Exponential Map	$\mathrm{Exp}_P(S) = CC^T,$
	$C = \lfloor \mathcal{L}(P) \rfloor + \lfloor \mathcal{L}(P) \rfloor \lfloor 0.5\mathcal{L}(P)^{-1} S \mathcal{L}(P)^{-T} \rfloor + \exp \mathbb{D}(\mathcal{L}(P)^{-1} S \mathcal{L}(P)^{-T}) \mathbb{D}(\mathcal{L}(P))$
Logarithmic Map	$\mathrm{Log}_P(Q) = CC^T, C = \mathcal{L}(P)(\lfloor \mathcal{L}(Q) \rfloor - \lfloor \mathcal{L}(P) \rfloor + \mathbb{D}(\mathcal{L}(P)) \log(\frac{\mathbb{D}(\mathcal{L}(Q))}{\mathbb{D}(\mathcal{L}(P))}))$
Distance	$d(P,Q) = (\Vert \lfloor \mathcal{L}(P) \rfloor - \lfloor \mathcal{L}(Q) \rfloor \Vert_F^2 + \Vert \log(\mathbb{D}(\mathcal{L}(P))) - \log(\mathbb{D}(\mathcal{L}(Q))) \Vert_F^2)^{\frac{1}{2}}$
Parallel Transport	$\mathrm{PT}^{\times}_{P\to Q}(A) = CC^T, C = \begin{cases} \frac{\mathcal{L}(A)_{ij}\mathcal{L}(Q)_{ij}}{\mathcal{L}(P)_{ij}} & \text{for } i = j \\ \mathcal{L}(A)_{ij} - \mathcal{L}(P)_{ij} + \mathcal{L}(Q)_{ij} & \text{for } i \neq j \end{cases}$
Linear Features	$RLF_P(X) = \begin{cases} \log\left(\frac{\mathcal{L}(X)_{ij}}{\mathcal{L}(P)_{ij}}\right) & \text{for } i = j \\ \mathcal{L}(X)_{ij} - \mathcal{L}(P)_{ij} & \text{for } i \neq j \end{cases}$

Table 3.4: **Log-Cholesky Metric Properties.** $\mathcal{L}$ denotes the Cholesky decomposition of a symmetric positive definite matrix, $\mathbb{D}(.)$ denotes the operator that extracts the diagonal part of a matrix and $\lfloor . \rfloor$ denotes the operator that extracts the lower triangular part below the diagonal.

3.7.3 Affine Invariant

The Affine Invariant (AIR) metric is a widely used metric on the manifold $\mathcal{P}^n$ of symmetric positive definite matrices. The metric is derived as the Fisher-Rao metric from the multivariate Gaussian model $N_p(\mu, \Sigma)$. Thus, it can be seen as the canonical metric for data from multivariate Gaussian processes. The complete mathematical derivation of all properties of the Affine Invariant metric can be followed in [43]. The curvature of the Affine Invariant metric is non-positive and bounded. See table 3.5 for properties of the Affine Invariant metric.

Remark 3.7.6. *As the Affine Invariant metric is not flat, its metric space under the geodesic distance can not be isometrically embedded into Hilbert space.*

Lemma 3.7.7. *For the Affine Invariant metric, the section $f_P^{-1}(S) = P^{-\frac{1}{2}} S P^{-\frac{1}{2}}$ corresponds to the parallel transport $\mathrm{PT}^{\times}$ when the data is transported to the identity.*

Proof. For the Affine Invariant metric, the $\mathrm{PT}^{\times}$ is known analytically and is given by $\mathrm{PT}^{\times}_{P\to Q}(A) = (QP^{-1})^{\frac{1}{2}} A (P^{-1}Q)^{\frac{1}{2}}$ [60]. For $Q = I$ this corresponds to the formula for the section. □

Lemma 3.7.8. *For the Affine Invariant metric, the Riemannian linear features correspond to the tangent space features of the data manifold transported to the identity.*

Proof. $f_I^{-1}(X) = f_I(X) = X$ with corollary 3.6.8. □

Lemma 3.7.9. *For the Affine Invariant metric, manifold transport preserves pairwise distances on the manifold, i.e., $d_{AIR}(A,B) = d_{AIR}(\mathrm{MT}_{P\to Q}(A), \mathrm{MT}_{P\to Q}(B))$ for all $P, Q, A, B \in \mathcal{P}^n$.*

Proof.

$$\begin{aligned}
d_{AIR}(\mathrm{MT}_{P\to Q}(A), \mathrm{MT}_{P\to Q}(B))^2 &= \mathrm{tr}(\mathrm{Log}^2(\mathrm{MT}_{P\to Q}(A)^{-1}\,\mathrm{MT}_{P\to Q}(B))) \\
&= \mathrm{tr}(\mathrm{Log}^2(((QP^{-1})^{\frac{1}{2}}A(P^{-1}Q)^{\frac{1}{2}})^{-1}(QP^{-1})^{\frac{1}{2}}B(P^{-1}Q)^{\frac{1}{2}})) \\
&= \mathrm{tr}(\mathrm{Log}^2((P^{-1}Q)^{-\frac{1}{2}}A^{-1}(QP^{-1})^{-\frac{1}{2}}(QP^{-1})^{\frac{1}{2}}B(P^{-1}Q)^{\frac{1}{2}})) \\
&= \mathrm{tr}(\mathrm{Log}^2((P^{-1}Q)^{-\frac{1}{2}}A^{-1}B(P^{-1}Q)^{\frac{1}{2}})) \\
&= \mathrm{tr}((P^{-1}Q)^{-\frac{1}{2}}\,\mathrm{Log}^2(A^{-1}B)(P^{-1}Q)^{\frac{1}{2}}) \\
&= \mathrm{tr}(\mathrm{Log}^2(A^{-1}B)) \\
&= d_{AIR}(A, B)
\end{aligned}$$

where we used the similarity invariance property of the trace trace and the fact that similarity transformations can be pulled out of the matrix logarithm function. □

As the Affine Invariant metric does not allow for embedding in Hilbert space, we would like to approximate its distance geometry using a distance measure that allows for embeddability. By [42, Thm. 2.11], the Log-Euclidean metric can be seen as an approximation of the Affine Invariant metric up to their commutators $[A, B] = AB - BA$. [62, Eq. 25] states the approximation of the Affine Invariant metric to the Euclidean distance between what we have defined as Riemannian linear features at the Frechét mean. We already know

Theorem 3.7.10. *[8, Thm. 6.1.4]*
$d_{AIR}(A, B) \geq d_{LE}(A, B)$ for all $A, B \in \mathcal{P}^n$ where d_{AIR} is the geodesic distance of the Affine Invariant metric and d_{LE} is the geodesic distance of the Log-Euclidean metric.

We can formalize the relationship between Affine Invariant distance and the Euclidean distance between Riemannian linear features in the same way:

Proposition 3.7.11. *Let $A, B, P \in \mathcal{P}^n$, d_E the Euclidean distance and d_{AIR} the geodesic distance of the Affine Invariant metric. Then*

$$d_{AIR}(A, B) \geq d_E(\mathrm{RLF}_P^{AIR}(A), \mathrm{RLF}_P^{AIR}(B)).$$

Proof. We have for arbitrary $P \in \mathcal{P}^n$

$$\begin{aligned}
d_{AIR}(A, B) &= d_{AIR}(\mathrm{MT}_{P\to I}(A), \mathrm{MT}_{P\to I}(B)) && (3.7.12)\\
&= d_{AIR}(P^{\frac{1}{2}}AP^{\frac{1}{2}}, P^{\frac{1}{2}}BP^{\frac{1}{2}}) && (3.7.13)\\
&\geq d_{LE}(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}, P^{-\frac{1}{2}}BP^{-\frac{1}{2}}) && (3.7.14)\\
&= d_E(\mathrm{RLF}_P(A), \mathrm{RLF}_P(B)) && (3.7.15)
\end{aligned}$$

showing the inequality using theorem 3.7.10, where d_{LE} is the Log-Euclidean geodesic distance. This inequality holds independent of the choice of P.

□

On a last note: the metric that is usually referred to as *the* Affine Invariant metric is just one example of the one-parameter subgroup of Affine Invariant metrics on SPD matrices, cf. [60]. However, in accordance with other literature on the field, we only consider this one example and continue to refer to it as the Affine Invariant metric.

Affine Invariant	
Inner Product	$g_P(A,B) = \mathrm{tr}(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}P^{-\frac{1}{2}}BP^{-\frac{1}{2}})$
Section	$f_P^{-1}(A) = P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$
Geodesic	$\gamma_{A\to B}(t) = A^{\frac{1}{2}}(A^{-\frac{1}{2}}BA^{-\frac{1}{2}})^t A^{\frac{1}{2}}$
Exponential Map	$\mathrm{Exp}_P(S) = P^{\frac{1}{2}}\,\mathrm{Exp}\,(P^{-\frac{1}{2}}QP^{-\frac{1}{2}})P^{\frac{1}{2}}$
Logarithmic Map	$\mathrm{Log}_P(Q) = P^{\frac{1}{2}}\,\mathrm{Log}(P^{-\frac{1}{2}}QP^{-\frac{1}{2}})P^{\frac{1}{2}}$
Distance	$d(P,Q) = \|\,\mathrm{Log}(P^{-\frac{1}{2}}QP^{-\frac{1}{2}})\| = \mathrm{tr}(\mathrm{Log}^2(P^{-1}Q))^{\frac{1}{2}}$
Parallel Transport	$\mathrm{PT}^{\times}_{P\to Q}(S) = (QP^{-1})^{\frac{1}{2}}S(P^{-1}Q)^{\frac{1}{2}}$
Manifold Transport	$\mathrm{MT}_{P\to Q}(A) = (QP^{-1})^{\frac{1}{2}}A(P^{-1}Q)^{\frac{1}{2}}$
Linear Features	$RLF_P(X) = \mathrm{Log}(P^{-\frac{1}{2}}XP^{-\frac{1}{2}})$

Table 3.5: **Affine Invariant Metric Properties.**

3.7.4 Bures-Wasserstein

Until extended in [41, 9], the Bures-Wasserstein framework was only known in terms of a distance function. This distance function is given as

$$d_{BW}(A,B) = (\mathrm{tr}(A) + \mathrm{tr}(B) - 2\,\mathrm{tr}((A^{\frac{1}{2}}BA^{\frac{1}{2}})^{\frac{1}{2}}))^{\frac{1}{2}}$$

In [9], it was shown that this distance function is a geodesic distance according to a Riemannian metric. The metric was then derived explicitly and given in the paper. For $P \in \mathcal{P}^n$, let $P = V_P\Lambda_P V_P^\top$ be the eigendecomposition where $[\Lambda_P]_{ii} = [\lambda_P]_i$ and $[\lambda_P]_i$ denotes the i-th eigenvalue of P. Then the Bures-Wasserstein metric is given as

$$g_P(A,B) = \frac{1}{2}\sum_{i,j}\frac{[V_P^TAV_P]_{ij}[V_P^\top BV_P]_{ij}}{[\lambda_P]_i + [\lambda_P]_j}.$$

Defining $[L_P]_{ij} = \frac{1}{\sqrt{[\lambda_P]_i+[\lambda_P]_j}}$, this can be rewritten as

$$g_P(A,B) = \frac{1}{2}\,\mathrm{tr}((L_P \odot (V_P^TAV_P))(L_P \odot (V_P^TBV_P)))$$

where $\odot$ is the Hadamard product.

The metric can be defined on positive semi-definite matrices.

In addition to the metric, [9, Eq. 39] provides an expression for the Bures-Wasserstein geodesic.

$$\gamma_{A,B}(t) = (1-t)^2A + t^2B + t(1-t)((AB)^{\frac{1}{2}} + (BA)^{\frac{1}{2}})$$

In [41], additional properties like exponential and logarithmic maps are stated. The Bures-Wasserstein metric has non-negative curvature.

Proposition 3.7.16. *For the Bures-Wasserstein metric, the section $f_P^{-1}(A) = \frac{1}{\sqrt{2}}V_P(L_P \odot (V_P^TAV_P))V_P^T$ corresponds to the section that is needed to construct parallel transport $\mathrm{PT}^{\times}$ when a tangent vector is transported to the tangent space of the identity.*

Proof. The analytic parallel transport for commuting matrices is given in [60]. As the identity commutes with any matrix, it can directly be applied. □

Bures-Wasserstein	$P = V_P^T \Lambda_P V_P$, $[L_P]_{ij} = \frac{1}{\sqrt{[\lambda_P]_i + [\lambda_P]_j}}$, $[L_P^*]_{ij} = \sqrt{[\lambda_P]_i + [\lambda_P]_j}$
Inner Product	$g_P(A,B) = \frac{1}{2}\operatorname{tr}((L_P \odot (V_P^T A V_P))(L_P \odot (V_P^T B V_P)))$
Section	$f_P^{-1}(A) = \frac{1}{\sqrt{2}} V_P (L_P \odot (V_P^T A V_P)) V_P^T$
Geodesic	$\gamma_{A\to B}(t) = (1-t)^2 A + t^2 B + t(1-t)((AB)^{\frac{1}{2}} + (BA)^{\frac{1}{2}})$
Exponential Map	$\operatorname{Exp}_P(S) = P + S + V_P((V_P^T S V_P \odot L_P^2)\Lambda(V_P^T S V_P \odot L_P^2))V_P^T$
Logarithmic Map	$\operatorname{Log}_P(Q) = (PQ)^{\frac{1}{2}} + (QP)^{\frac{1}{2}} - 2P$
Distance	$d(A,B) = (\operatorname{tr}(A) + \operatorname{tr}(B) - 2\operatorname{tr}((A^{\frac{1}{2}} B A^{\frac{1}{2}})^{\frac{1}{2}}))^{\frac{1}{2}}$
Parallel Transport	$\mathrm{PT}_{P\to Q}(S) = (V_Q(L_Q^* \odot (V_Q^T(V_P(L_P \odot (V_P^T S V_P))V_P^T)V_Q))V_Q^T)$
Manifold Transport	$\mathrm{MT}_{P\to Q}(A) = \operatorname{Exp}_Q(\mathrm{PT}_{P\to Q}(\operatorname{Log}_P(S))$
Linear Features	$RLF_P(X) = \frac{1}{\sqrt{2}} V_P(L_P \odot (V_P^T(\operatorname{Log}_P(X))V_P))V_P^T$

Table 3.6: **Bures-Wasserstein Metric Properties.** As the parallel transport according to the Levi-Civita connection along a geodesic between arbitrary matrices is unknown for the Bures-Wasserstein metric, we give the expression for the parallel transport according to the Levi-Civita connection along the concatenated geodesic $\gamma_{P\to I} \circ \gamma_{I\to Q}$.

Contrary to the Affine Invariant case, this section for parallel transport is not the evident section from the metric. The evident candidate would have been $f_P^{-1}(A) = \frac{1}{\sqrt{2}} L_P \odot (V_P^T A V_P)$. However, the orthogonal transformation required as noted in lemma 3.4.3 is given simply as $V_P f V_P^T$.

This merits the following lemma:

Lemma 3.7.17. *The Riemannian linear features for the Bures-Wasserstein metric correspond to the identity transported tangent space features scaled by* 2*.*

Proof. With $f_P^{-1}(A) = \frac{1}{\sqrt{2}} V_P (L_P \odot (V_P^T A V_P)) V_P^T$, $V_I = I$ and $[L_I]_{ij} = \frac{1}{\sqrt{2}}$ we get $f_I^{-1}(A) = \frac{1}{2}A$ and thus $f_I(A) = 2A$. By lemma 3.6.4 we have

$$\operatorname{Log}_I(\mathrm{MT}_{P\to I}(X)) = 2\,\mathrm{RLF}_P(X). \tag{3.7.18}$$

□

Like the Euclidean metric, the Bures-Wasserstein metric is not geodesically complete [59] and suffers from the swelling effect [34, Table 1].

3.8 Discussion of Theoretical Novelties

The concept of sections plays a minor theoretical role within the field of differential geometry. Even if the orthonormal frame bundle is proven to admit global sections, this finding can often not be translated to yield any applicative value as the analytic expression of a section may not be found.

We have an evident candidate section for all known Riemannian metrics on $\mathcal{P}^n$. The candidate section can be obtained for all of the proposed metrics by the common mathematical procedure of "looking at it". In the case of the Affine Invariant metric, the section is evident from the construction of the metric. For the Bures-Wasserstein metric, one needs to reformulate the sum expression of the metric found in [9] in terms of the Frobenius inner product using the matrix trace. For pullback metrics, the section in the pullback space is

determined by the differential. The formalization of the Riemannian metrics using sections offers a concise way to describe the Riemannian metrics on SPD matrices and the impact that the Riemannian metric has on an individual tangent vector. By considering the section as an individual function rather than part of an inner product, we can use it to describe the properties of the Riemannian manifold.

Connecting a section f_P at a point $P \in \mathcal{P}^n$ with a section f_Q at a point $Q \in \mathcal{P}^n$ by $f_Q(f_P^{-1})$ is a linear isometry between tangent spaces. Parallel transport, according to the Levi-Civita connection along a geodesic, is also a linear isometry between tangent spaces, which allows us to leverage sections to find an analytical expression. It remains to find an orthogonal transformation that relates the two concepts. This orthogonal transformation, if necessary, is dependent only on the reference points P and Q at which the sections are defined. This orthogonal transformation might not be easily obtained for the tangent spaces of two general points. However, we can obtain the correct section easily for all of the proposed curved metrics. Only the Bures-Wasserstein metric section needs an orthogonal transformation determined by the orthogonal matrix obtained by the eigendecomposition of the reference point.

By leveraging that sections map to standard Euclidean space, we can concisely describe a feature map that maps from any manifold structure on $\mathcal{P}^n$ to standard Euclidean space. Riemannian linear feature space provides a tool for the description of previously unnamed features, described for the Affine Invariant metric, for example, in [62, 3]. This description reduces the complexity of many mathematical expressions and allows for simplified analysis in the following chapters of this thesis. This space does not usually require any particular section for its description.

Only when entering the realm of transfer learning does one need to more carefully consider the section used to construct the common domain on which the data is analyzed further. If we want a single simple expression to align data sets from different domains, we need to use the section that can be used to describe the Levi-Civita parallel transport. Any section can be used in tasks that use additional alignment steps (including orthogonal transformations) to align data sets from different sources projected to Riemannian linear feature space. By breaking the self-introduced restriction that the tangent space is an inner product space of symmetric matrices under the Frobenius inner product and instead considering it as a vector space under the standard Euclidean inner product, we can leverage a much larger class of orthogonal transformation that can be applied to the vectors. This allows for more flexibility in the alignment of datasets.

4 Linearization of the Manifold

When considering a linear approximation of a manifold, the first idea is commonly the tangent space which is, by definition, a linear space connected to any point on the manifold. By introducing the Riemannian linear features, we presented a method of linearizing the manifold using vectors different from tangent space vectors. The logarithmic map gives the tangent vectors for any Riemannian metric $g_P(A, B) =< f_P^{-1}(A), f_P^{-1}(B) >$. $\mathrm{Log}_P(X)$ describes the tangent vector of a point $X \in \mathcal{P}^n$ on the tangent space of the reference point $P \in \mathcal{P}^n$. In contrast, the Riemannian linear feature of a point $X \in \mathcal{P}^n$ at a reference point $P \in \mathcal{P}^n$ is defined by $\mathrm{RLF}_P(X) := f_P^{-1}(\mathrm{Log}_P(X))$. This allows for incorporation of the inner product structure of the manifold in the feature vector that is used for linearization. In this chapter, we assess how well the two linearization methods reproduce the geometric structure of the underlying manifold in terms of distance geometry.

4.1 Methods

We analyze the distance geometry of the tangent space and the Riemannian linear feature space and compare it to the distance geometry of the manifold under the geodesic distance. We view both linear spaces as standard Euclidean spaces, using the Euclidean distance to compare it to the true distance geometry of the underlying manifold. We employ two different paradigms to validate the distance geometry of linearizations.

The first paradigm is based on the AUClogRNX score, which can be used to compare the distance geometries between two spaces. As this scoring method is not heavily used in the literature, we included a short introduction to AUClogRNX in the appendix, see section 8.6. Using proposed linearization methods, we calculate the AUClogRNX score to assess the preservation of neighborhood structure as a proxy for the distance geometry when translating from the manifold to a linear space. We use all available datasets described in section 8.4. As we do not test generalization ability, the following procedure is used: Given a set of data points, all points are projected onto the linear space at the Frechét mean of all points. The AUClogRNX score is then calculated for all linearized points, with the ground truth assumed to be the neighborhood structure of the points on the manifold under the geodesic distance. For all available datasets, we repeat this procedure for each session of each subject and report the average AUClogRNX score for all proposed Riemannian metrics across all sessions for each subject in each dataset.

In a second comparison of linearization methods, we assess the linear separability of data points from two classes in a classification task by the ROC AUC score of the simplest linear classifier, the minimum distance to mean (MDM), also known as the nearest centroid classifier. We compare these separability scores to the MDM classifier's ROC AUC score on the manifold using the geodesic distance.

This paradigm incorporates the generalization ability of the linearization method. We perform a within-subject cross-session classification of a two-class motor imagery task using all datasets described in section 8.4. The cross-validation is performed in a leave-one-session-out scheme. On the manifold, points are classified according to the given geodesic distance directly on the manifold (MF), where the class-wise means

are estimated on the training data, and the test data is classified according to the geodesic distance to the means. For tangent space features, the data is projected onto the tangent space (TS) of the global mean of the training dataset. The tangent space is considered to be the standard Euclidean space, and the class-wise means are estimated according to the Euclidean distance on the tangent space. The test points are projected onto the tangent space of the mean of the training points and classified according to the closest class mean. As for the tangent space features, the Riemannian linear features (RLF) are calculated at the mean of the training dataset. Again, the Euclidean distance is used to estimate the class-wise means on Riemannian linear feature space. The test points are projected onto the Riemannian linear feature space of the mean of the training points and classified according to the closest class mean.

4.2 Results

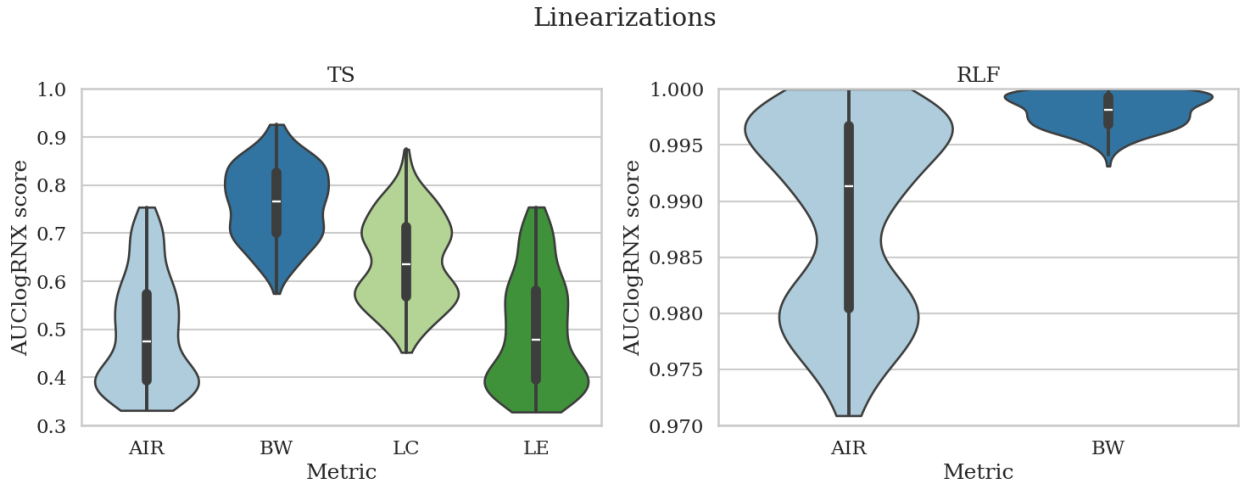

Figure 4.1: **Comparison of the distance geometry of linearization with the manifold.** AUClogRNX score of tangent space (TS) projections and Riemannian linear features (RLF) at the Frechét mean, where the ground truth is the neighborhood structure using the geodesic distance on the manifold. Metrics are Affine Invariant (AIR), Bures-Wasserstein (BW), Log-Cholesky (LC), and Log-Euclidean (LE). **Left:** AUClogRNX score of tangent space projections. **Right:** AUClogRNX score of Riemannian linear features. LE and LC metrics are omitted, as the AUClogRNX score on Riemannian linear features is always one for these metrics. Note the different limits on the y-axis.

The results of the comparisons of distance geometry are shown in fig. 4.1. For the Bures-Wasserstein metric, the tangent space projection results in a median AUClogRNX score of 0.78, whereas the Affine Invariant metric tangent space projection achieves a score of 0.48. Considering flat metrics, the tangent space projection of the Log-Cholesky metric with a median score of 0.63 closer resembles the distance geometry on the manifold than that of the Log-Euclidean metric with a median score of 0.48. The Riemannian linear features perfectly reproduce the distance geometry on the manifold for the flat metrics. The results show an AUClogRNX score close to one for Riemannian linear features. The median AUClogRNX score using the Affine Invariant metric is 0.989, and for the Bures-Wasserstein metric, it is 0.998. The lowest AUClogRNX score of

the affine invariant metric is 0.971, while the lowest score for the Bures-Wasserstein metric is 0.993. For flat metrics, the score is always one.

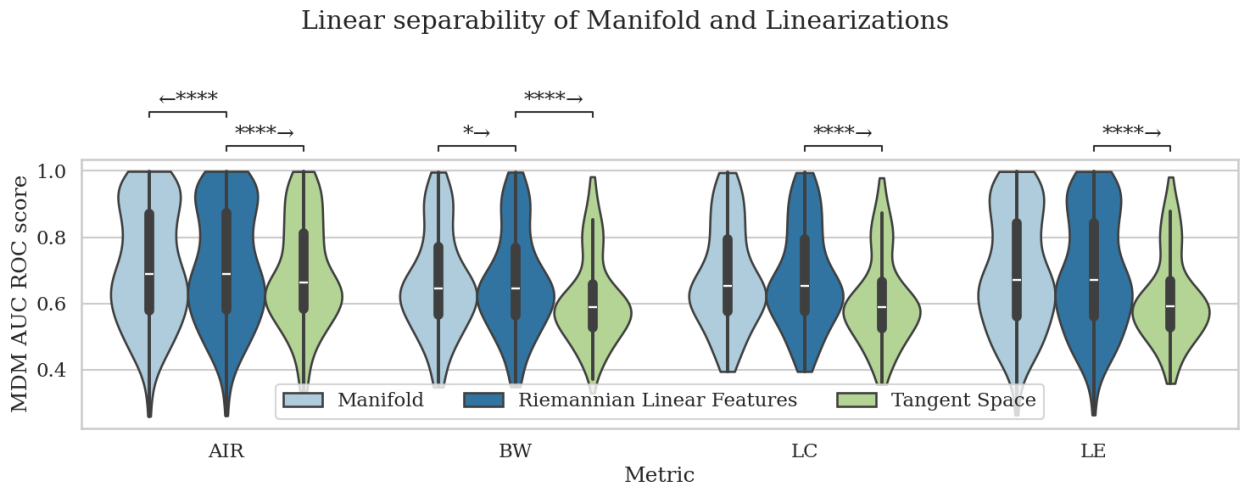

Figure 4.2: **Comparison of linear separability.** AUC ROC score of minimum distance to mean (MDM) classifier comparing manifold (MF), Riemannian linear feature space (RLF), and tangent space (TS) features. Riemannian linear features and tangent vectors are projections to the Frechét mean of the training data. The metrics are Affine Invariant (AIR), Bures-Wasserstein (BW), Log-Cholesky (LC), and Log-Euclidean (LE). Results of all datasets described in section 8.4 are combined. Significances (for details, see section 8.5) are only shown between MF and RLF and between RLF and TS. There is no significant difference between the results if no significance is shown.

Figure 4.2 shows the results of the linear separability assessment comparing the manifold, tangent space, and Riemannian linear feature space. Across all metrics, the tangent space projection performs significantly worse than using the features directly on the manifold or Riemannian linear features. For flat metrics, the Riemannian linear features and points on the manifold performed equivalently. For curved metrics, Riemannian linear features and points on the manifold performed similarly. For the Affine Invariant metric, the Riemannian linear features projection performed significantly better than the points on the manifold whereas we see the opposite for the Bures-Wasserstein metric.

For the Affine Invariant metric, the AUC ROC score for the manifold features is 0.688, the score for the Riemannian linear features is 0.69, and the score for the tangent space is 0.662, resulting in a median difference of 2.6 percentage points (p.p.) and 2.8 p.p. from tangent space to manifold and Riemannian linear feature space respectively. For the Bures-Wasserstein metric, the AUC ROC score for the manifold features is 0.645, the score for the Riemannian linear features is 0.6445, and the score for the tangent space is 0.589, resulting in a median difference of around 5.5 p.p. from tangent space to manifold and Riemannian linear feature space. Considering flat metrics, the classification using manifold and Riemannian linear features attained the same score using both metrics, with a median AUC ROC score of 0.654 for the Log-Cholesky metric and a median score of 0.67 for the Log-Euclidean metric. The tangent space features attained a score of 0.59 for both metrics.

4.3 Discussion

The results show that the linear approximation of the manifold by Riemannian linear features is significantly superior to that of the tangent space vectors.

In the comparison of distance geometries, see fig. 4.1, the distance geometry is retained almost perfectly for curved metrics, with median AUClogRNX scores of close to one for the Affine Invariant and the Bures-Wasserstein metrics across all datasets. Even with datasets of 64 channels, the Affine Invariant Riemannian linear features never score an AUClogRNX below 0.97, and the Bures-Wasserstein Riemannian linear features never score below 0.99. For flat metrics, the distance geometry is exactly retained, which coincides with the definition of a flat manifold.

The general notion of the results is reproduced when comparing linear separability in a classification task on the manifold, see fig. 4.2. While linearization by Riemannian linear features almost exactly reproduces the AUC ROC score of the MDM classifier on the manifold, linearization using the tangent space performs significantly worse across all metrics. However, the differences are not as pronounced as in the AUClogRNX score. Especially for the Affine Invariant metric, the median AUC ROC score difference between manifold or Riemannian linear features and the tangent space is only around 3 p.p.

We lose significant information about the manifold's distance geometry by projecting onto the tangent space equipped with the standard Euclidean metric rather than Riemannian linear feature space. This shows the value of including the linear structure provided by the Riemannian metric in the description of the linearization of a feature vector.

5 Kernels on Riemannian Manifolds

In this section, we will describe kernels on Riemannian manifolds and show under which circumstances they can be shown to be positive definite, enabling their unrestricted use in machine learning applications. First, we give a general introduction to kernels. We then introduce two distinct classes of kernels that can be constructed on Riemannian manifolds.

We distinguish inner product kernels and metric space kernels. Inner product kernels can be applied to general Riemannian manifolds without restriction. We construct the so-called canonical inner product kernels that can be transferred to arbitrary manifolds, not only SPD matrices. The kernels for metric spaces need the mild prerequisite of connectedness of the Riemannian manifold, which is fulfilled for SPD matrices, see section 2.2.3. The main theoretical result of this section is that for Riemannian metrics that are pullback metrics from the Euclidean metric, all results from the Euclidean distance function can be readily used. This allows the transfer of results from kernels in standard Euclidean space to all flat Riemannian metrics on SPD matrices.

We test how these two different classes of kernel methods compare in classification and dimensionality reduction tasks using the common kernel methods Support Vector Machines and kernel Principal Component Analysis.

5.1 Background: Kernels

Definition 5.1.1. *[52, Def. 2.3] Let $\mathfrak{X}$ be a non-empty set. Let $k : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ be a function. Let $(X_i)_{i=1,\dots,n}, X_i \in \mathfrak{X}$ be a dataset. A matrix $K_{ij} := k(X_i, X_j)$ is called a Gram matrix.*

5.1.1 Positive Definite Kernels

Definition 5.1.2. *[52, Def. 2.5] Let $\mathfrak{X}$ be a non-empty set. A function $k : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ which for all $m \in \mathbf{N}$ and all $(X_i)_{i=1,\dots,m}$ gives rise to a positive semi-definite Gram matrix is called a positive definite kernel.*

This definition is inconsistent as it relates positive definite kernels with positive semi-definite Gram matrices. However, in kernel theory, we are interested in positive semi-definite Gram matrices, and calling the corresponding kernels positive definite is the standard way in the literature.

Definition 5.1.3. *[52, Def. 2.20] A symmetric matrix $K \in \mathbb{R}^{m \times m}$ satisfying*

$$\sum_{i,j}^{m} c_i c_j K_{ij} \geq 0 \text{ for all } c_i \in \mathbb{R}, \text{ with } \sum_{i}^{m} c_i = 0$$

is called conditionally positive definite.

Definition 5.1.4. *A kernel k giving rise to a conditionally positive definite kernel matrix is called a conditionally positive definite kernel.*

Note that if a kernel is positive definite, it is also conditionally positive definite. Thus, the class of conditionally positive definite kernels is larger than and includes the class of positive definite kernels. Kernel applications like support vector machines (section 8.3) and kernel Principal Component Analysis (section 8.2.1) allow for conditionally positive definite kernels and are not restricted to positive definite kernels [52, p. 52].

Constructing (conditionally) positive definite kernels can be achieved in various ways. This is often done by transforming an already known positive definite kernel while retaining its positive definiteness. There are numerous possible transformations, so we will restrict listing those that we will need to prove the positive definiteness of kernels later in the thesis.

The simplest class of positive definite kernels are the inner products that are positive definite by definition:

Theorem 5.1.5. *[52] Let $\mathfrak{H}$ be a Hilbert space with inner product $< .,. >$. Then the kernel defined by $k : \mathfrak{H} \times \mathfrak{H} \to \mathbb{R}$, $k(X_i, X_j) =< X_i, X_j >$ is positive definite.*

We can use positive definite functions to construct positive definite kernels on metric spaces by reparameterization, as shown by the following proposition:

Proposition 5.1.6. *[52, cf. Prop. 2.28] Let (M, d) be a metric space that can be isometrically embedded in Hilbert space, $(X_i)_{i=1,\dots,n} \in M$, and f be a positive definite function. Then the kernel $k : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ defined by*

$$k(X_i, X_j) = f(d(X_i, X_j))$$

is positive definite.

We can describe the kernels arising from distances of metric spaces that can be isometrically embedded in Hilbert space by the following proposition:

Proposition 5.1.7. *$k(x, x') = -d(x, x')^\beta$, $0 \leq \beta \leq 2$ is conditionally positive definite if and only if the metric space (M, d) can be isometrically embedded into Hilbert space (extends [52, Eq. 2.81] by [52, Prop. 2.24]).*

This allows us to transfer the theorem connecting positive definite functions with Hilbert embeddability (see theorem 2.2.7) from metric spaces to kernels:

Proposition 5.1.8. *[52, Prop. 2.28] k is a conditionally positive definite kernel if and only if $e^{\lambda k}$ is a positive definite kernel for all $\lambda > 0$.*

The connection of (conditionally) positive definite functions and completely monotone functions can be transferred to (conditionally) positive definite kernels.

Proposition 5.1.9. *[52, p. 54] Let (M, d) be a metric space that can be embedded isometrically into Hilbert space. Let $h : \mathbb{R} \to \mathbb{R}$ be a function defined on the values of d.*

Let $k : M \times M \to \mathbb{R}$ such that $k(x_i, x_j) := h(d(x_i, x_j)^2)$. We have
$k(x_i, x_j)$ is a positive definite kernel $\iff$ $h(t)$ is completely monotonic of order 0.
$k(x_i, x_j)$ is a conditionally positive definite kernel $\iff$ $h(t)$ is completely monotonic of order 1.

Proposition 5.1.10. *[52, Prop. 2.23] Let $\mathfrak{X}$ be non-empty. Let $k : \mathfrak{X} \times \mathfrak{X} \to (-\infty, 0]$ be a conditionally positive definite kernel, $0 < \alpha < 1$ and $b \in \mathbb{R}$. Then $k + b$, $-(-k)^\alpha$, and $-ln(1 - k)$ are conditionally positive definite kernels.*

Proposition 5.1.11. *[6] Let $k_n : \mathfrak{X} \times \mathfrak{X}$ be a family of positive definite kernels such that $\lim_{n\to\infty} k_n = k$ exists. Then, k is a positive definite kernel.*

Proposition 5.1.12. *Let $k : \mathfrak{X} \times \mathfrak{X} \to (-\infty, 0]$ be a conditionally positive definite kernel. Then $(1-k)^{-\alpha}$ for $\alpha > 0$ is a positive definite kernel.*

Proof. Follows from proposition 5.1.10 and proposition 5.1.8. □

5.2 Inner Product Kernels

We distinguish two basic types of inner product kernels on Riemannian manifolds. There is an inherent way of deriving an inner product kernel for SPD matrices given a specific Riemannian metric tensor. These kernels are constructed using the logarithmic function derived from each Riemannian metric in combination with the Riemannian metric. We will call the resulting kernels canonical kernels, as they form a canonical way of describing inner products between SPD matrices, and inner products are inherently positive definite.

Other inner product kernels are other inner products that are defined on Euclidean space. As the manifold of SPD matrices forms an open subset of the symmetric matrices which are isomorphic to standard Euclidean space, we can transfer the definition of any inner product on Euclidean space to this subset, without directly adhering to a strict derivation from the underlying geometry.

5.2.1 Canonical Kernels

We denote by canonical kernels the kernels derived from the logarithmic map in combination with the Riemannian inner product at a given reference point P. The first canonical kernel that was derived for SPD matrices was the kernel for the Affine Invariant metric [3].

The paper also provides the general approach to the derivation.

For a given Riemannian manifold $(\mathcal{P}^n, g)$, a canonical kernel k is derived by combining a Riemannian metric $g_P =< ., . >_P$ on the tangent space $T_P\mathcal{P}^n$ at P with the corresponding logarithmic map $\mathrm{Log}_P(.)$:

$$k_P(A, B) =< \mathrm{Log}_P(A), \mathrm{Log}_P(B) >_P =< \mathrm{RLF}_P(A), \mathrm{RLF}_P(B) > \tag{5.2.1}$$

where the kernel depends on a reference point P.

By leveraging the notation of Riemannian linear features, we note that canonical kernels on Riemannian manifolds are linear kernels in Riemannian linear feature space.

Canonical kernels have a single hyperparameter $P \in \mathcal{P}^n$, the so-called reference point determining the tangent space to which a given dataset is projected to estimate the kernel. This point critically determines the kernel's performance and is usually chosen as the Frechét mean of the data points. If the Frechét mean is used as the reference point, Riemannian linear features are centered, see lemma 3.6.9. The reference point also enables the kernel to be used for online learning, as it can be updated in the learning process to adjust for data drifts [68].

Unless explicitly stated otherwise, we assume the reference point P to be the Frechét mean of the dataset in the construction of canonical kernels and omit it as a hyperparameter in the definition.

5.2.1.1 Affine Invariant

The canonical kernel for the Affine Invariant Riemannian metric for SPD matrices was first derived in [3].

The Riemannian linear features are given as

$$RLF_P(X) = \mathrm{Log}(P^{-\frac{1}{2}} X P^{-\frac{1}{2}})$$

and the canonical kernel is

$$k(X_i, X_j) = \mathrm{tr}(\mathrm{Log}(P^{-\frac{1}{2}} X_i P^{-\frac{1}{2}}) \mathrm{Log}(P^{-\frac{1}{2}} X_j P^{-\frac{1}{2}})).$$

5.2.1.2 Bures-Wasserstein

The Bures-Wasserstein kernel has not previously been derived. Let $P = V^T \Lambda V$ and $\tilde{X} = V^T (PX)^{\frac{1}{2}} V$, $[L]_{ij} = \frac{1}{\lambda_i + \lambda_j}$.

The Bures-Wasserstein Riemannian linear feature are given as

$$RLF_P(X) = \frac{1}{\sqrt{2}}(V(L \odot (V^T((PX)^{\frac{1}{2}} + (XP)^{\frac{1}{2}} - 2P)V))V^T)$$

.

Combining the Frobenius inner product and the Riemannian linear features, we get

$$k(X_i, X_j) = \frac{1}{2} \mathrm{tr}((L \odot (V^T((PX_i)^{\frac{1}{2}} + (X_i P)^{\frac{1}{2}} - 2P)V))(L \odot (V^T((PX_j)^{\frac{1}{2}} + (X_j P)^{\frac{1}{2}} - 2P)V)))$$

The resulting formula has a more complex expression than for the other kernels. It has roughly the same computational complexity as the Affine Invariant kernel since the matrix root has to be calculated for all data points.

5.2.1.3 Pullback Metrics

For pullback metrics using an isometry $\phi : M \to N$ with logarithmic map $\mathrm{Log}^*()$ and metric g^*, the canonical kernel takes the form

$$\begin{aligned} k(X_i, X_j) &= g^*_P(\mathrm{Log}^*_P(X_i), \mathrm{Log}^*_P(X_j)) \\ &= g_{\phi(P)}([\mathrm{D}_P \phi]([\mathrm{D}_P \phi]^{-1}(\mathrm{Log}_{\phi(P)}(\phi(X_i)))), [\mathrm{D}_P \phi]([\mathrm{D}_P \phi]^{-1}(\mathrm{Log}_{\phi(P)}(\phi(X_j))))) \\ &= g_{\phi(P)}(\mathrm{Log}_{\phi(P)}(\phi(X_i)), \mathrm{Log}_{\phi(P)}(\phi(X_j))) \end{aligned} \tag{5.2.2}$$

where g and $\mathrm{Log}()$ are the Riemannian metric and the logarithmic map from the Riemannian manifold N whose structure is pulled back to M.

In case the metric is pulled back from the Euclidean metric, we have $g_P(A, B) =< A, B >$ and $\mathrm{Log}_{\phi(P)}(\phi(X)) = \phi(X) - \phi(P)$, so the resulting canonical pullback kernel is constructed as

$$k(X_i, X_j) =< \phi(X_i) - \phi(P), \phi(X_j) - \phi(P) > .$$

Log-Euclidean. In [3], the Log-Euclidean kernel, which should be assumed to be derived from the Log-Euclidean Riemannian metric, is stated as a special case of the Affine Invariant canonical kernel. This is in contrast to a canonical derivation from the Riemannian metric. The resulting kernel lacks important features of canonical kernels.

The author first derived the canonical Log-Euclidean kernel in [7]. By eq. (5.2.2) the kernel is

$$k(X_i, X_j) =< \mathrm{Log}(X_i) - \mathrm{Log}(P), \mathrm{Log}(X_j) - \mathrm{Log}(P) > .$$

Log-Cholesky. As the Log-Cholesky metric was not derived as a simple pullback from Euclidean space in [40], its expression following the inner product and logarithmic map construction procedure is more complex than for other metrics. Following the derivation in [40] using the propeorties in table 3.6, the kernel can be constructed as

$$\begin{aligned} k(X_i, X_j) &= < \lfloor \mathrm{Log}_{\mathcal{L}(P)}(\mathcal{L}(X_i)) \rfloor, \lfloor \mathrm{Log}_{\mathcal{L}(P)}(\mathcal{L}(X_j)) \rfloor > \\ &\quad + < \mathbb{D}(\mathcal{L}(P))^{-1}\mathbb{D}(\mathcal{L}(\mathrm{Log}_{\mathcal{L}(P)}(\mathcal{L}(X_i)))), \mathbb{D}(\mathcal{L}(P))^{-1}\mathbb{D}(\mathcal{L}(\mathrm{Log}_{\mathcal{L}(P)}(\mathcal{L}(X_i)))) > \\ &= < \lfloor \mathcal{L}(X_i) \rfloor - \lfloor \mathcal{L}(P) \rfloor, \lfloor \mathcal{L}(X_j) \rfloor - \lfloor \mathcal{L}(P) \rfloor > \\ &\quad + < \log(\mathbb{D}(\mathcal{L}(P))^{-1}\mathbb{D}(\mathcal{L}(X_i)), \log(\mathbb{D}(\mathcal{L}(P))^{-1}\mathbb{D}(\mathcal{L}(X_j)) > . \end{aligned}$$

In contrast, using the isometry

$$\phi(X)_{ij} = \begin{cases} \log(\mathcal{L}(X)_{ij}) & \text{for } i = j \\ \mathcal{L}(X)_{ij} & \text{for } i \neq j \end{cases}$$

constructed in eq. (3.7.5), the kernel can be written much simpler in the form of a pullback from Euclidean space, i.e.

$$k(X_i, X_j) =< \phi(X_i) - \phi(P), \phi(X_j) - \phi(P) > .$$

Although the equations look very different, they are necessarily equivalent per definition of flat metrics.

Name	Kernel $k(X_i, X_j)$
Log-Euclidean	$\mathrm{tr}((\mathrm{Log}(X_i) - \mathrm{Log}(P))(\mathrm{Log}(X_j) - \mathrm{Log}(P)))$
Affine-invariant	$\mathrm{tr}(\mathrm{Log}(P^{-\frac{1}{2}} X_i P^{-\frac{1}{2}}) \mathrm{Log}(P^{-\frac{1}{2}} X_j P^{-\frac{1}{2}}))$
Log-Cholesky	$< \lfloor \mathcal{L}(X_i) \rfloor - \lfloor \mathcal{L}(P) \rfloor, \lfloor \mathcal{L}(X_j) \rfloor - \lfloor \mathcal{L}(P) \rfloor >$ $+ < \mathrm{Log}(\mathbb{D}(\mathcal{L}(P))^{-1}\mathbb{D}(\mathcal{L}(X_i))), \mathrm{Log}(\mathbb{D}(\mathcal{L}(P))^{-1}\mathbb{D}(\mathcal{L}(X_j))) >$
Bures-Wasserstein	$\mathrm{tr}((L \odot (\tilde{X}_i + \tilde{X}_i^T))(L \odot \tilde{X}_j) - (\tilde{X}_i + \tilde{X}_j + \Lambda))$

Table 5.1: **Canonical kernels for SPD matrices.** The Affine Invariant kernel was derived in [3]. The author derived the Log-Euclidean kernel in [7]. All other kernels are explicitly stated here for the first time. We have $P = V^T \Lambda V$ and $\tilde{X} = V^T (PX)^{\frac{1}{2}} V$, $[L]_{ij} = \frac{1}{\lambda_i + \lambda_j}$. $\mathcal{L}$ denotes the Cholesky decomposition, $\lfloor . \rfloor$ extracts the strictly lower triangular part of a matrix and $\mathbb{D}(.)$ extracts the diagonal part of a matrix.

5.2.2 Other Inner Product Kernels

Theoretically, any inner product defined on the tangent space $\mathcal{S}^n \simeq T_P\mathcal{P}^n$ can be transferred directly to $\mathcal{P}^n$. This is because $\mathcal{P}^n$ is an open subset of $\mathcal{S}^n \simeq T_P\mathcal{P}^n$. Thus, we can restrict any inner product on $T_P\mathcal{P}^n$ to $\mathcal{P}^n$ and always ensure we yield a valid inner product directly on $\mathcal{P}^n$. This includes, for example, all Riemannian metrics without performing tangent space projection.

In practice, this may lead to suboptimal kernels, as the inner products are defined on a given tangent space which is different for each Riemannian metric. Thus, we will only state other inner products mentioned and used to some extent in the literature.

Frobenius. The Frobenius inner product between matrices $X_i \in \mathbb{R}^{n \times m}$

$$k(X_i, X_j) =< X_i, X_j >^F = \operatorname{tr}(X_i X_j^T)$$

is not restricted to symmetric matrices and can be applied to any two matrices of the same dimension. As there exists a Euclidean Riemannian metric, one should adhere to the naming conventions of Riemannian metrics and refrain from referring to the Frobenius inner product as the Euclidean inner product, although there is a resemblance to the Euclidean inner product for vectors.

Log-Frobenius. The Log-Frobenius inner product is what has thus far been termed the Log-Euclidean kernel [3]. It is defined as the Frobenius inner product after the application of the matrix logarithm to an SPD matrix:

$$k(X_i, X_j) =< \operatorname{Log}(X_i), \operatorname{Log}(X_j) >^F .$$

It is equivalent to the Log-Euclidean and Affine Invariant canonical kernels when the reference point in both is set to the identity matrix. Naming this inner product the Log-Euclidean kernel does make sense, as the Euclidean inner product for vectorized matrices is the same as the Frobenius matrix inner product. However, as there is a Riemannian metric named Log-Euclidean metric, this name would be used twice. We suggest a change in nomenclature to avoid confusion. The kernel previously called the Log-Euclidean kernel should be referred to as the Log-Frobenius kernel, a name that directly states its construction. The kernel arising canonically from the Log-Euclidean Riemannian metric should be named as such, to have consistent naming conventions for other inner products (or kernels) arising from Riemannian metrics.

5.3 Kernels for Metric Spaces

In this section, we cover kernels defined on Riemannian manifolds as metric spaces. The kernels use only a distance function as an input and are often called distance-based or isotropic kernels or radial basis functions. In this chapter, we will list common kernel functions for metric spaces and provide proof of positive definiteness with the Euclidean metric. Afterwards, we show how the results from Euclidean space can be transferred to metric spaces arising from Riemannian manifolds.

Name	Distance	Flat	References
Euclidean	$d(A,B) = \|A - B\|$	✓	
Log-Euclidean	$d(A,B) = \|\operatorname{Log}(A) - \operatorname{Log}(B)\|$	✓	[2]
Affine Invariant	$d(A,B) = \|\operatorname{Log}(A^{-1}B)\|$	✗	[54]
Log-Cholesky	$d(A,B) =(\|\lfloor\mathcal{L}(A)\rfloor - \lfloor\mathcal{L}(B)\rfloor\|_F^2 + \|\operatorname{Log}(\mathbb{D}(\mathcal{L}(A))) - \operatorname{Log}(\mathbb{D}(\mathcal{L}(B)))\|_F^2)^{\frac{1}{2}}$	✓	[40]
Bures-Wasserstein	$d(A,B) = \sqrt{\operatorname{tr}(A + B - 2(A^{\frac{1}{2}}BA^{\frac{1}{2}})^{\frac{1}{2}})}$	✗	[1]
LogDet	$d(A,B) = \sqrt{\log\det(\frac{A+B}{2}) - \frac{1}{2}\log\det(AB)}$	✗	[58]

Table 5.2: **Distances for SPD matrices.** $\mathcal{L}$ denotes the Cholesky decomposition, $\lfloor . \rfloor$ extracts the strictly lower triangular part of a matrix and $\mathbb{D}(.)$ extracts the diagonal part of a matrix.

5.3.1 Common Euclidean Metric Kernels

In this section we will provide proofs for the positive definiteness of known metric space kernels from the literature for the Euclidean metric $d_E(A, B) = ||A - B||$. The proofs rely on the fact that the Euclidean metric d_E is conditionally negative definite, i.e., the kernel $k(A, B) = -d_E(A, B)$ is conditionally positive definite. Note that all proofs depend on kernel parameter ranges.

Gaussian

$$k(X_i, X_j) = e^{-\gamma d^2(X_i, X_j)} . \tag{5.3.1}$$

Positive definite for all positive values of γ for the Euclidean metric by virtue of proposition 5.1.8 and proposition 5.1.7.

Inverse Multiquadratic For $\beta > 0$ and $\theta \in \mathbb{R}$

$$k(X_i, X_j) = (d^2(X_i, X_j) + \theta^2)^{-\beta} \tag{5.3.2}$$

defines the inverse multiquadratic kernel.

Proposition 5.3.3. *The inverse multiquadratic kernel is positive definite for the Euclidean metric for all* $\theta \geq 1, \beta > 0$.

Proof. By proposition 5.1.7, $-d^2$ is conditionally positive definite. Proposition 5.1.10 implies that $-d^2 - \theta^2 + 1$ is conditionally positive definite. Then, proposition 5.1.12 gives that $(1 - (-d^2 - \theta^2 + 1))^\alpha = (d^2 + \theta^2)^\alpha$ is positive definite. □

The inverse multiquadratic kernel can be rewritten as

$$(d^2(X_i, X_j) + \theta^2)^{-\beta} = e^{\log((d^2(X_i, X_j)+\theta^2)^{-\beta}} = e^{-\beta \log((d^2(X_i, X_j)+\theta^2)}.$$

Rational Quadratic

$$k^{RQ}(X_i, X_j) = \left(1 + \frac{d^2(X_i, X_j)}{2\beta l^2}\right)^{-\beta} . \tag{5.3.4}$$

The Gaussian kernel is the limit of the Rational Quadratic kernel with $\beta \to \infty$.

Proposition 5.3.5. *The Rational Quadratic kernel is positive definite for the Euclidean metric for* $2\beta l^2 \geq 1$.

Proof. The Rational Quadratic kernel is a rewriting of the Inverse Multiquadratic kernel in the following sense:

$$k^{RQ}(X_i, X_j) = \left(1 + \frac{d^2(X_i, X_j)}{2\beta l^2}\right)^{-\beta} = \xi^\beta \left(\xi + d^2(X_i, X_j)\right)^{-\beta} \text{ where } \xi = 2\beta l^2.$$

Now, if $\xi \geq 1$, then the right hand side is a positive multiple of the Inverse Multiquadratic kernel, which is known to be positive definite for the Euclidean metric for $\xi \geq 1$. □

5.3.2 Metric Space Kernels on Riemannian Manifolds

The construction of metric space kernels can be transferred to Riemannian manifolds by leveraging the geodesic distance to view a Riemannian manifold as a metric space. There is one key theorem for the Gaussian kernel that is the basis for the positive definiteness of the proposed metric space kernels: It connects the Gaussian kernel positive definiteness to the embeddability of the metric space into Hilbert space (cf. e.g. [22, Thm. 1 & 2]).

Theorem 5.3.6. *Let $(\mathcal{P}^n, g)$ be a simply connected Riemannian manifold with geodesic distance d. Then the Gaussian kernel $e^{-\gamma d^2}$ is positive definite for all $\gamma > 0$ if and only if $(\mathcal{P}^n, d)$ can be isometrically embedded in Hilbert space, i.e., $\mathcal{P}^n$ is a flat manifold.*

Without further mathematical work, this theorem cannot directly be extended to all other metric space kernels. We have found empirically that no other metric space kernel is positive definite for any of the curved metrics for the same parameter ranges that result in positive definite kernels in Euclidean space.

5.3.2.1 Flat Metrics

Theorem 5.3.6 and proposition 5.1.7 show that all results on positive definiteness of metric space kernels can directly be transferred to flat Riemannian manifolds on SPD matrices. This is because these manifolds are globally isometric to Euclidean space and thus allow for the construction of the distance as $d_\phi(A, B) = ||\phi(A) - \phi(B)||$ where ϕ is the isometry to construct the pullback metric from the Euclidean metric, cf. e.g. [22, 32]. By proposition 5.1.7 we know that $d_\phi(A, B)$ is conditionally negative definite, i.e., $-d_\phi(A, B)$ is conditionally positive definite. We noted that the proofs rely on the conditionally negative definite metric; thus, all proofs can be directly transferred.

5.3.2.2 Curved Metrics

For the proposed curved metrics on $\mathcal{P}^n$, none of the metric space kernels are positive definite for the whole parameter ranges in which they are positive definite with the Euclidean metric.

However, certain parameter ranges for each kernel using each geodesic distance might produce positive definite kernels.

Only one noteworthy result specifies parameter ranges for the positive definiteness of a metric space kernel that does not stem from the geodesic distance of a flat Riemannian manifold on SPD matrices.

There is a special case of the Gaussian kernel, which has its own name in the literature. The Gaussian kernel using the LogDet distance is often called the Stein kernel. It is given as

$$k^s(P_1, P_2) = e^{-\beta S(P_1, P_2)} = 2^{c\beta} \frac{\sqrt{\det(P_1)^\beta \det(P_2)^\beta}}{\det(P_1 + P_2)^\beta}. \tag{5.3.7}$$

It is a positive definite kernel when $\beta \in \{1/2, \ldots, (d-1)/2\} \cup \{\gamma \in \mathbb{R} | \gamma > (d-1)/2\}$ [27, Thm. 1]. This is the only known result for any number of data points.

To transfer the results from Euclidean geometry to curved metrics, we can leverage the distance between Riemannian linear features rather than the (geodesic) distances on the manifold. As the Riemannian linear feature space is Euclidean space, the Euclidean distance between Riemannian linear features is used. This is a first-order approximation of the distances of curved spaces, as noted for the Affine Invariant metric in [62].

Method	Parameters	
Gaussian Kernel	$\gamma \in [10^i]_{i=-3,-2,...,3}$	
Stein Kernel	$\gamma \in [\frac{i}{2}]_{i=1,2,...,5}$	
Inverse Multiquadratic Kernel	$\beta \in [10^i]_{i=-3,-2,...,3}$	$\theta \in [10^{i/5}]_{i=1,2,...,5}$
Rational Quadratic kernel	$\beta \in [10^i]_{i=-3,-2,...,3}$	$l \in [10^{i/5}]_{i=1,2,...,5}$
SVC	$C \in [10^i]_{i=-3,-2,...,3}$	
Online Learning	$n_{window} \in [10, 20, 40, 80]$	

Table 5.3: Parameter ranges for kernel methods.

5.4 Methods

We assert the performance of canonical kernels and metric space kernels to validate the feasibility of using kernels derived from Riemannian geometry for machine learning. The underlying paradigm was the same two-class within-subject cross-session motor imagery task that was used for the linear separability assessment in the linearization chapter, see section 4.1. The cross-validation was performed in a leave-one-session-out scheme. Hyperparameters were optimized using a grid search [46] in a nested four-fold cross-validation on the training set. The significance analysis is described in table 8.2. As canonical kernels are inherently positive definite for any metric, we use all kernels from table 5.1. Metric space kernels derived from the geodesic distance are only positive definite for flat metrics. To be able to include analyses of non-flat metrics, we use the linear approximation of the distance geometry of curved manifolds by Riemannian linear features. As Riemannian linear features are only defined on Riemannian manifolds and not general metric spaces, the LogDet distance does not have a linear formulation. For a list of metric space kernels with tested parameter ranges, see table 5.3.

We use two evaluations for testing kernel methods, classification and dimensionality reduction. The classification analysis is split into offline and online classification. The offline classification analysis was conducted for all canonical kernels and three metric space kernels for all Riemannian metrics except the Euclidean metric (i.e., Affine Invariant, Log-Euclidean, Log-Cholesky, Bures-Wasserstein). The Euclidean metric is excluded as it commonly produces poor results in the proposed classification tasks [15]. The only known positive definite kernel for the LogDet distance is the Gaussian kernel with limited parameter ranges, so testing other metric space kernels was omitted for the LogDet distance.

The best-performing metric space kernel was then used as a representative in the online classification (see below) task, where it was again used with the same Riemannian metrics. In the dimensionality reduction analysis, we again chose the best-performing metric space kernel to compare it with the canonical kernel for each metric.

5.4.1 Classification

We assess the accuracy of an offline classification and a (simulated) online classification task. The performance is measured as the accuracy per session. As classes are evenly distributed across all sessions, this is equivalent to balanced accuracy. The classifier used for all kernels was the kernel C-Support-Vector-Machine [63] with the hyperparameter C selected as described in table 5.3.

Offline Classification. Offline learning is a setting in which a fixed set of hyperparameters is estimated on an entire training run of a classification pipeline and then applied without alteration on a test set.

Simulated Online Classification. Online learning (cf., e.g., [35, 10]) is a setting in which a classification pipeline is updated during training and/or validation. In EEG recordings, there are typically non-stationarities that cause drifts in the data. We can adjust for these drifts by updating hyperparameters or reference values in the pipeline. In Riemannian geometry learning methods, the reference point P can be updated during the learning process to implement online learning (cf. [68]). Regardless of the classification method, we can use parallel or manifold transport to project the points on the tangent space or on the manifold from the continuously updated starting reference point towards a common fixed target reference point according to a given Riemannian metric. We projected to the identity matrix as the target reference point, as proposed in [68]. After this online projection to the identity, we can apply any offline method to the data points. The reference point is recalculated for each point using the last n_{window} covariance samples in an unsupervised fashion, i.e., without using labels. The optimal n_{window} is estimated as a hyperparameter in the nested grid search. This way, we simulate an online setting with the existing datasets. For brevity, we will refer to the simulated online learning as online learning. We note that this implementation of online learning is unsupervised, i.e., it does not require knowledge of the classes of the training data to be performed and can be applied to the test data in the same fashion.

Figure 5.1 shows the effect of the online update of the reference point as a visualization of two principal components on Riemannian linear feature space of the Affine Invariant metric.

5.4.2 Dimensionality Reduction

We assessed the ability of kernels to perform meaningful dimensionality reduction of data. We again used the AUClogRNX (see section 8.6) to assess the preservation of distance geometry approximated by the neighborhood structure. The scores are calculated across the same datasets under the same paradigm from the classification task using kernel PCA with the canonical and Gaussian kernels for all metrics. This means that a dimensionality reduction ability trained on a training set is assessed on a test set.

We compare kernel PCA with the canonical kernels and the Gaussian metric space kernel, which performed best in the classification task for all metrics. To restrict computational complexity, we selected the number of dimensions tested to be two and the number of channels from the EEG recording of the datasets denoted by $n_{channels}$. Two is the common dimensionality for visual representation and $n_{channels}$ measures how well dimensionality reduction works for a dimensionality that only scales linearly with matrix dimensionality.

5.5 Results

5.5.1 Classification

In this section, we compare the generalization ability of kernel methods in a classification task using two different paradigms: offline and unsupervised online learning.

Figure 5.2 shows offline classification accuracy of canonical and metric space kernels. For each metric, we can observe the same trend: The canonical kernel significantly outperforms the Gaussian kernel. The Gaussian kernel significantly outperforms the inverse multiquadratic kernel. The inverse multiquadratic kernel

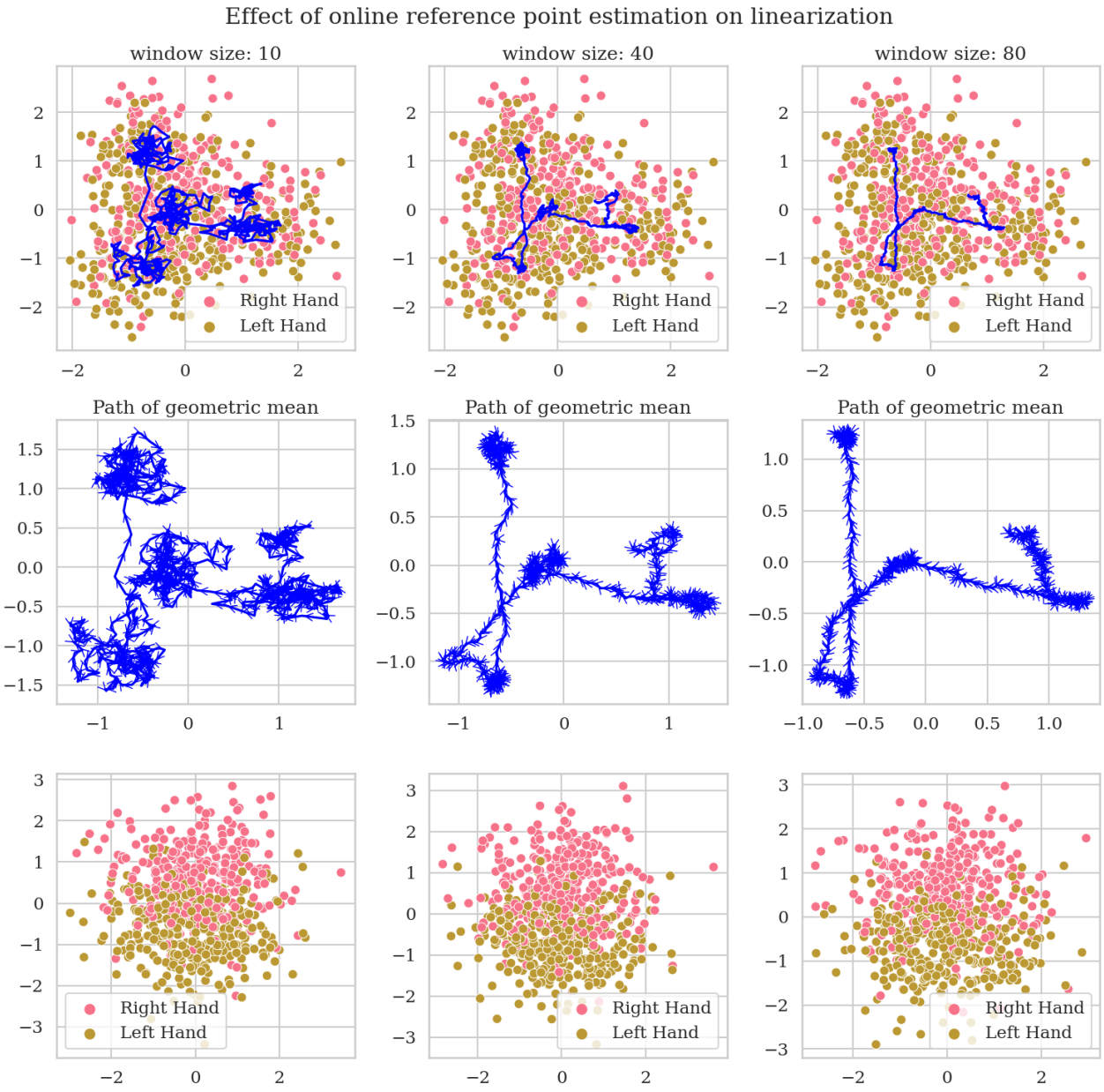

Figure 5.1: **Effect of online updates of the reference point.** Plots of data projected onto the first two principal components of Riemannian linear features for the Affine Invariant metric. **Top:** Projection onto the mean of the entire dataset. Overlay with the path of the geometric mean according to the window size given in samples across which the mean is estimated online. **Middle:** Detailed view of the path of the geometric mean as projected onto the first two principal components. **Bottom:** Riemannian linear feature projection according to the online updated geometric mean in the middle row.

significantly outperformed the rational quadratic kernel. The median accuracy score of the canonical kernel for each metric was 0.66 (Affine Invariant), 0.65 (Bures-Wasserstein), 0.64 (Log-Cholesky), and 0.66 for the Log-Euclidean metric. For the Affine Invariant (0.25 p.p.) and Log-Euclidean (1 p.p.) metrics, the difference

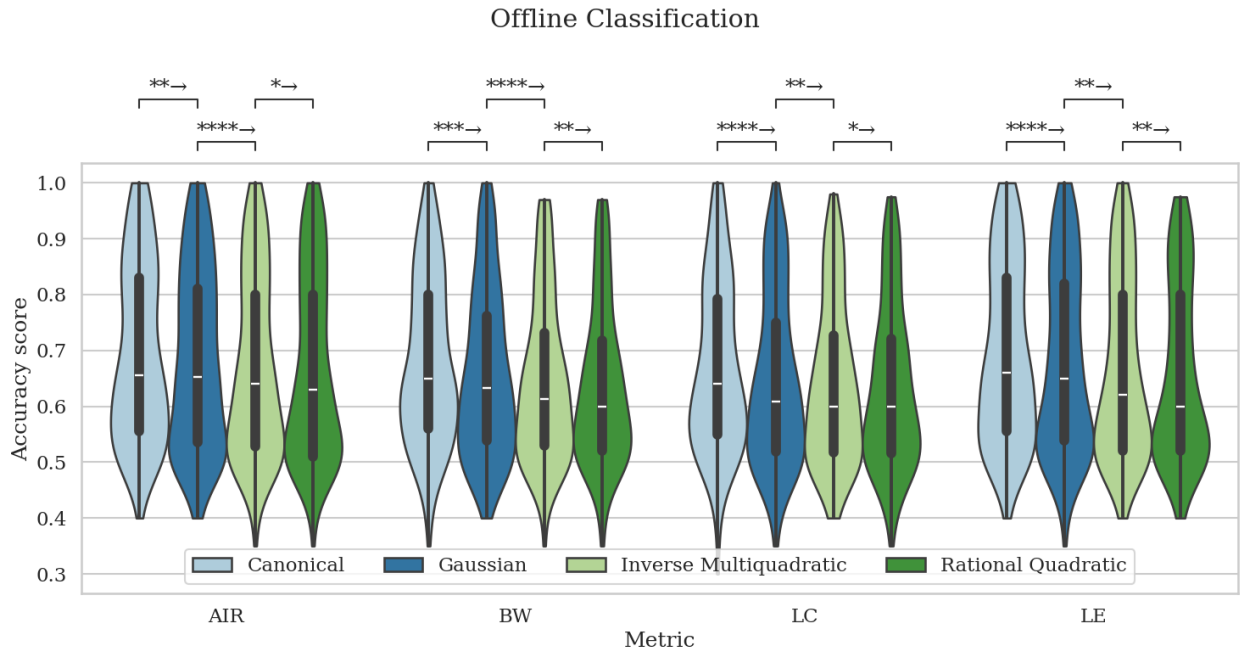

Figure 5.2: **Offline Classification performance comparison of Riemannian kernels.** Classification performance is measured in balanced accuracy with a kernel SVC using canonical and metric space kernels for each metric. Metrics are Affine Invariant (AIR), Bures-Wasserstein (BW), Log-Cholesky (LC), and Log-Euclidean (LE). Classification is performed in a leave-one-session-out cross-validation scheme for all datasets described in section 8.4. Significances (for details, see section 8.5) are given per metric for consecutive kernels.

in median accuracy between the canonical and Gaussian kernels was the lowest. For the Bures-Wasserstein and the Log-Cholesky distances, the difference in median accuracy was 1.7 p.p. and 3.2 p.p., respectively.

For the Affine Invariant metric, there was the smallest difference (2.5 p.p.) between the best-performing kernel (canonical) and worst-performing kernel (Rational Quadratic). For the Bures-Wasserstein, Log-Cholesky, and Log-Euclidean metrics, the median distance between the best and worst-performing kernels was 5 p.p., 4 p.p., and 6 p.p., respectively.

Also tested, but not shown in this figure, was the Gaussian kernel with the LogDet distance or Stein kernel. This kernel did not outperform any other Gaussian kernels in median accuracy, with a score of 0.625.

Figure 5.3 shows the results for the online classification. For the Affine Invariant, Bures-Wasserstein, and Log-Cholesky metrics, the online canonical kernel significantly outperforms the online Gaussian kernel. For the Log-Euclidean metric, there is no significant difference between the online canonical kernel and the online Gaussian kernel. The median accuracy of canonical and Gaussian online kernels for each metric are 0.7475 and 0.71 for the Affine Invariant metric, 0.7 and 0.65 for the Bures-Wasserstein metric, 0.68 and 0.64 for the Log-Cholesky metric and 0.72 and 0.7 for the Log-Euclidean metric.

Across all metrics, updating the reference point significantly increases classification performance for canonical and Gaussian kernels. The largest improvement from offline to online learning for both kernels measured in median accuracy is visible for the Affine Invariant metric. The median accuracy improves by 9.25 p.p. from

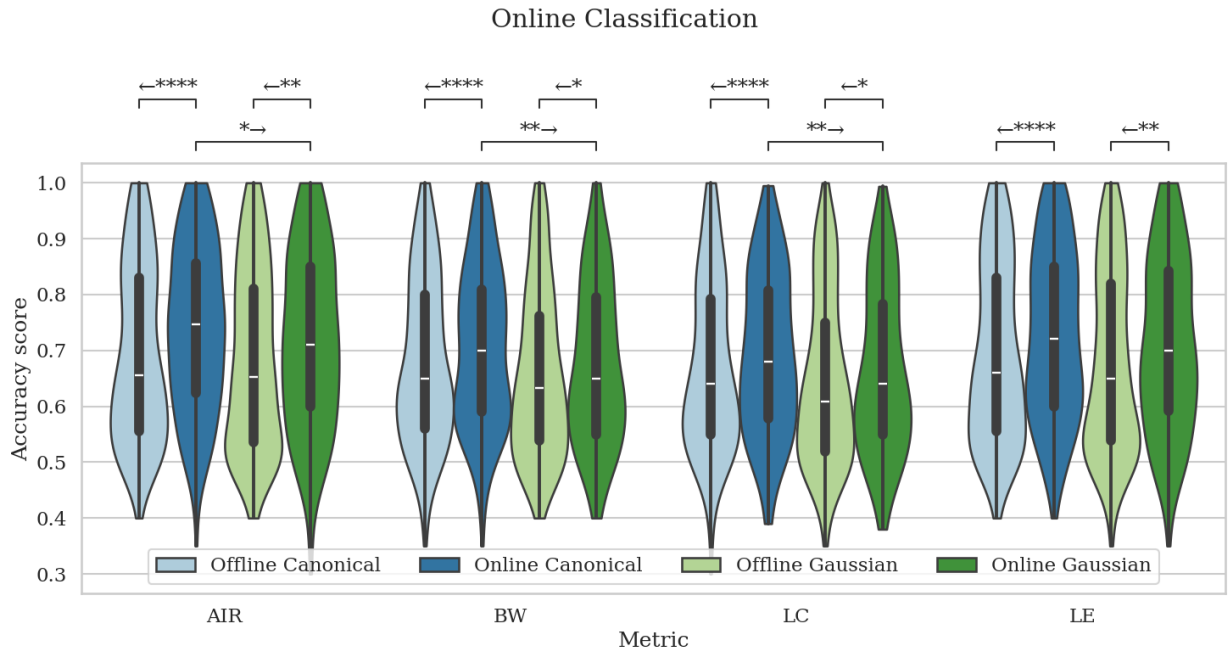

Figure 5.3: **Classification performance of online kernels.** Classification performance is measured in balanced accuracy with a kernel SVC using the canonical and Gaussian kernels for each metric. In the online paradigm, the reference point is updated for each matrix by the Frechét mean of the last n_{mean} samples, where n_{mean} is a hyperparameter optimized in the inner cross-validation loop. Offline kernels are replotted for reference. Metrics are Affine Invariant (AIR), Bures-Wasserstein (BW), Log-Cholesky (LC), and Log-Euclidean (LE). Significances (for details, see section 8.5) are given per metric between the online and offline canonical and Gaussian kernels and between the online canonical and online Gaussian kernels.

offline to online learning under the canonical kernel. The median accuracy improves by 5.75 p.p. from offline to online learning under the Gaussian kernel.

5.5.2 Dimensionality Reduction

Figure 5.4 shows the results of the dimensionality reduction. For the Affine Invariant and Log-Cholesky metrics, the two-dimensional reductions with the canonical kernels significantly outperform the dimensional reductions with the Gaussian kernel for both dimensions. In contrast, there is no significant difference for the Bures-Wasserstein and Log-Euclidean metrics. In general, the median difference between the scores between the kernels for each metric and dimensionality is always lower than 0.2 p.p. The median score for each of the metrics for the two-dimensional reduction is 0.624 for the Affine Invariant metric, 0.68 for the Bures-Wasserstein metric, 0.615 for the Log-Cholesky and 0.641 for the Log-Euclidean metric. The median score for each of the metrics for the reduction to $n_{channels}$ dimensions is 0.782 for the Affine Invariant metric, 0.905 for the Bures-Wasserstein metric, 0.879 for the Log-Cholesky and 0.799 for the Log-Euclidean metric.

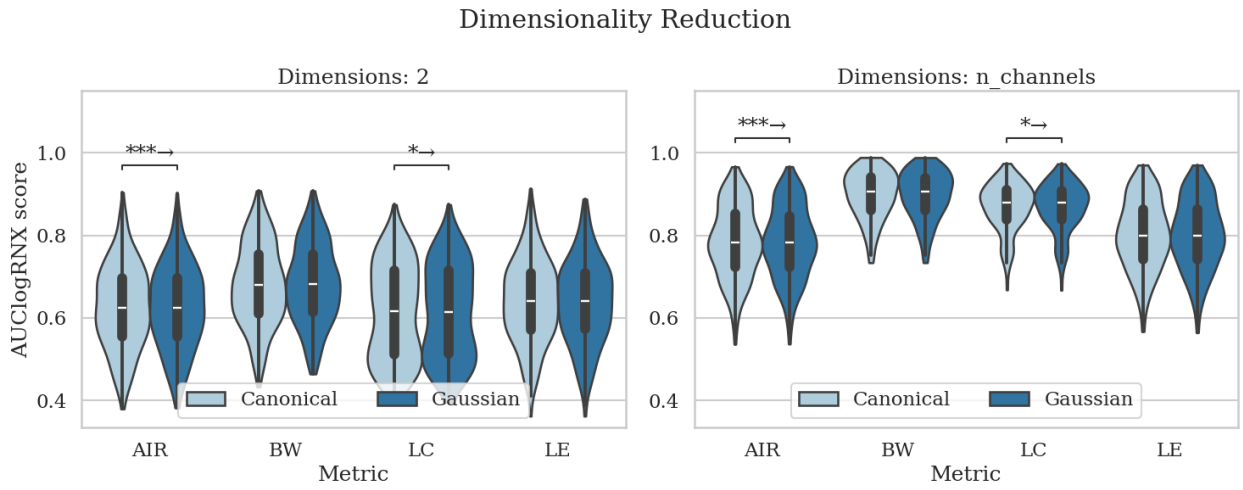

Figure 5.4: **Comparison of the distance geometry of dimensionality reductions with the manifold.** AUClogRNX score of dimensionality reduction using canonical and Gaussian kernel PCA with varying dimensions. The score is measured on a test set of the Riemannian linear features projected to the feature space at the Frechét mean of the training set. The ground truth is the neighborhood structure using the geodesic distance on the manifold. Metrics are Affine Invariant (AIR), Bures-Wasserstein (BW), Log-Cholesky (LC), and Log-Euclidean (LE). Dimensionality reduction is performed to 2 and $n_{channels}$ dimensions, where $n_{channels}$ is the number of channels in the recording. Significance analysis is performed between canonical and Gaussian kernels per metric and dimension.

5.6 Discussion

5.6.1 Positive Definiteness on Riemannian Manifolds

We stated two general classes of kernels on Riemannian manifolds for which we can assert positive definiteness. First, we introduced the notion of canonical kernels as the combination of the logarithmic map and the Riemannian metric at any point $P \in M$. This class of kernels is always positive definite, regardless of the manifold. This is ensured by the fact that the Riemannian metric, which is an inner product, is positive definite by definition.

The second class are the metric space kernels. Here, we can use any positive definiteness result from standard Euclidean space and transfer it directly to flat metrics. This is due to the theory of embeddability of metric spaces into Hilbert spaces. As noted above, there is one key theorem concerning the Gaussian kernel that is found throughout the literature that connects its positive definiteness to the embeddability of the metric space into Hilbert space (cf., e.g., [22, Thm. 1 & 2]):

Theorem 5.6.1. *Let $(\mathcal{P}^n, g)$ be a simply connected Riemannian manifold with geodesic distance d. Then the Gaussian kernel $e^{-\gamma d^2}$ is positive definite for all $\gamma > 0$ if and only if $\mathcal{P}^n$ can be isometrically embedded in Hilbert space, i.e., $\mathcal{P}^n$ is a flat manifold.*

This means that no curved Riemannian manifold can produce a positive definite Gaussian kernel for all

parameters of γ based on its geodesic distance. This theorem is a good starting point when researching the positive definiteness of metric space kernels on Riemannian manifolds, as it may allow transfer to other kernels. It appears to not always be considered in scientific literature in the field. For example, in [25, Thm. 3] we can find a (wrong) proof that the Gaussian kernel for the curved Bures-Wasserstein metric's geodesic distance is positive definite for all $\gamma > 0$.

Although there is active research on finding parameter ranges of γ for which the Gaussian kernel is positive definite for a given curved Riemannian manifold (cf., e.g., [21]), no progress has been made for any of the geodesic distances for Riemannian manifolds on $\mathcal{P}^n$ for curved metrics.

Other metric space kernels need to be considered individually to prove positive definiteness. There is strong evidence that the inverse multiquadratic and rational quadratic kernels are positive definite if and only if the metric space is isometric to Hilbert space. However, we have only been able to derive preliminary proofs that are not currently ready for publication.

5.6.2 Computational Considerations

Canonical kernels can be computed as linear kernels in Riemannian linear feature space, which makes them very efficient and allows for a choice of computational method when used with scikit-learn in Python [46]. The implementations of LinearSVC and SVC, where the kernel parameter is set to 'linear', use different underlying implementations. The SVC implementation with the linear kernel uses the libsvm library [13] and allows the implementation of custom kernels with precomputed kernel matrices. It is a general library for SVCs, written to cover many different kernels, and does not focus on possible optimizations of individual kernel methods. LinearSVC uses the liblinear library [20] optimized for linear SVCs. It scales better with a large number of samples, allows more flexibility in the choice of penalty methods and loss function, and offers additional parameter settings.

5.6.3 Classification

In offline classification with a support vector machine, the linear kernels on Riemannian linear features significantly outperform their respective metric space counterparts on the Riemannian linear features across all metrics. The Gaussian kernel performs best within the class of metric space kernels across all metrics.

A similar result is reproduced in the online classification setting. Classification using canonical kernels outperforms the Gaussian kernel for all metrics except the Log-Euclidean metric. For all metrics, using the online learning paradigm rather than the offline paradigm resulted in significantly increased performance. This should thus be the standard procedure in time-continuous tasks where a data drift is to be expected.

The Affine Invariant metric appears to provide the most robust framework for kernel application of the proposed metrics. For this metric, there is the least difference between the best-performing kernel and the worst-performing kernel in the offline classification task. Also, each kernel under the Affine Invariant metric outperforms any of its counterparts under any other proposed Riemannian metrics regarding median accuracy. Thus, although we need to use the approximation of the Affine Invariant metric for the metric space kernels, this metric produces good results in these methods.

5.6.4 Dimensionality Reduction

We used the AUClogRNX score to compare the neighborhood structure of lower dimensional feature vectors using Kernel-PCA to the neighborhood structure on the manifold using the geodesic distance of each metric. The upper bound of the AUClogRNX score of linear kernel dimensionality reduction algorithms on Riemannian linear feature space is given by the AUClogRNX score of the manifold linearization using Riemannian linear features (see fig. 4.1). As all proposed positive definite kernels can be formulated on the Riemannian linear feature space, this upper bound holds for any kernel-based dimensionality reduction.

There is no general clear significant difference in the performance of dimensionality reduction for Gaussian and canonical kernel PCA. Both methods perform similarly in terms of the median AUClogRNX score. However, there are significant differences for the Affine Invariant and Log-Cholesky metric that indicate superiority of dimensionality reduction with canonical kernels. The median difference, however, is small, and the choice of kernel method should be made based on other grounds (e.g., computational complexity).

We can assume that dimensionality reduction with kernel PCA using canonical kernels performs well in finding directions of maximal variance. It is the optimal method for flat metrics as it can be formulated as PCA on Riemannian linear feature space. For curved metrics, we still expect a near-optimal retention of explained variance of dimensionality reduction with kernel PCA due to the high approximative quality of the Riemannian linear feature space to recover the distance geometry of the manifold as shown in fig. 4.1. We would expect the retention of variance using this method to be in the realm of the results presented in [28].

6 Principal Geodesic Analysis

Principal Geodesic Analysis (PGA, [24]) is the transfer of Principal Component Analysis (PCA, see section 8.2.1) to manifolds. The idea remains the same: we want to find the direction that solves some optimization problem. Instead of vectors, these directions are now geodesics on the manifold. A prerequisite for applying PGA is that the resulting manifold is geodesically complete.

Although PGA is not a kernel method, we will show that it relates to PCA and kernel PCA.

In this chapter, we will first go through the derivation of PGA. Then, we will state under which circumstances exact solutions to PGA are possible and introduce a notion of manifold linearization using principal geodesics. We then compare the linearization using PGA to the linearization using Riemannian linear features. Last, we compare the dimensionality reduction capabilities of PGA with those of kernel PCA using the canonical kernel.

6.1 Theory

6.1.1 Derivation

Following the same structure as in the exemplary derivation of PCA in the appendix (see section 8.2.1), we start with defining principal geodesics. Assuming data with Frechét mean μ, we define a geodesic direction parametrized by an SPD matrix $M \in \mathcal{P}^n$ as

$$\gamma_{\mu \to M}(t).$$

To conform with an equivalent unit-length prerequisite as for PCA, we introduce the restriction on the direction M given by

$$|| \operatorname{Log}_\mu(M)||_\mu = < \operatorname{Log}_\mu(M), \operatorname{Log}_\mu(M) >_\mu^{\frac{1}{2}} = 1.$$

Orthogonality of components is enforced in the Riemannian sense on the tangent space of the mean μ by requiring

$$< \operatorname{Log}_\mu(M_i), \operatorname{Log}_\mu(M_j) >_\mu = 0.$$

Equivalently to the linear case, we can describe this geodesic direction as a one-dimensional subspace of $\mathcal{P}^n$ parametrized by $t \in \mathbb{R}$. This subspace is often referred to as geodesic subspace. In the following analysis, we use geodesic and geodesic subspace equivalently and only state the geodesic-defining direction M if the starting point is clear. When an argument t is provided, it means $M(t) := \gamma_{\mu \to M}(t)$ As in the linear case of PCA, the variance of a dataset is defined as $var(\{X_i\}_{i=1,\dots,n}) = \sum_i^n d^2(\mu, X_i)$ where d is the geodesic distance.

Equipped with a geodesic direction $\gamma_{\mu \to M}$, we can turn to calculate the orthogonal projection of an SPD matrix X onto $\gamma_{\mu \to M}(t)$.

Similar to PCA, this can be done by finding a t^* that minimizes the distance of some data point X to the geodesic by the optimization problem

$$t^* = \arg\min_t d(\gamma_{\mu\to M}(t), X)^2 \tag{6.1.1}$$

where d is the geodesic distance induced by the given Riemannian metric. Finding the optimal t^* is not as straightforward as in the linear case, as the optimization problem requires differentiation of the geodesic distance function.

Instead of going through the derivative calculation which would result in the same statement, we can state a second way to find t^* by directly using the orthogonality constraint in the sense of Riemannian manifolds. Thus, we are required to find a t^* such that

$$< \mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(\gamma_{\mu\to M}(t)), \mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(X) >_{\gamma_{\mu\to M}(t^*)} = 0 \tag{6.1.2}$$

for any $t \neq t^*$.

The restriction $t \neq t^*$ comes from the fact that if $t = t^*$, we have $\mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(\gamma_{\mu\to M}(t^*)) = 0$ as the logarithmic map maps the base point to zero. This would trivially fulfill the equation. Analogously to PCA we can define a projection operator $\pi_{\gamma_{\mu\to M}}(X)$ that orthogonally projects a point X onto a geodesic $\gamma_{\mu\to M}(t)$ by using the notation of t^* as

$$\pi_{\gamma_{\mu\to M}}(X) = \gamma_{\mu\to M}(t^*). \tag{6.1.3}$$

where t^* is the solution to $< \mathrm{Log}_{\gamma(t^*)}(\gamma(t)), \mathrm{Log}_{\gamma(t^*)}(X) >_{\gamma(t^*)} = 0$.

In general there is no direct way to state PGA as an iterative procedure with removing variance in each step as can be done for PCA. We need to incorporate an orthogonality constraint for all consecutive principal geodesics. To do so, we introduce $V_i = span(\{v_1, ..., v_i\})$ as the space spanned by the tangent directions $\{v_1, ..., v_i\}$ and by $V_i^\perp$ its orthogonal complement.

Again, we can formulate the optimization problem of PGA in several ways. Contrary to PCA, the approaches are not generally equivalent. We assume manifold-valued data $\{X_i\}_{i=1,...,n} \in \mathcal{P}^n$.

Maximization of Variance. This optimization problem is the 'original' PGA objective stated in [24]. The goal is to find principal geodesics parametrized by positive definite matrices M_i or their respective tangent vectors $v_i = \mathrm{Log}_\mu(M_i)$, leading to directions $\gamma_{\mu\to M_i}(t) = \mathrm{Exp}_\mu(tv_i)$ such that

$$v_i = \arg\max_{\|v\|=1, v\in V_{i-1}^\perp} \sum_{j=1}^n d\left(\mu, \pi_{\gamma_{\mu\to M_i}}(X_j)\right)^2 = \arg\max_{\|v\|=1, v\in V_{i-1}^\perp} \sum_{j=1}^n (t_j^*)^2 \tag{6.1.4}$$

where the dependency on v is hidden in the projection operator π.

Minimization of Projection Error. To minimize the projection error, we need to solve the second optimization problem [57, Eq. 2.8]

$$v_i = \arg\min_{\|v\|=1, v\in V_{i-1}^\perp} \sum_{j=1}^n d\left(X_j, \pi_{\gamma_{\mu\to M_i}}(X_j)\right)^2. \tag{6.1.5}$$

Minimization of Total Residual Variance. To minimize total residual variances, we want to find a function ψ_M parametrized by the direction M that removes variance in a given direction $\gamma_{\mu\to M}$. For this, let us consider the equation

$$< \mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(\gamma_{\mu\to M}(t)), \mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(\psi_M(X)) >_{\gamma_{\mu\to M}(t^*)} = 0. \tag{6.1.6}$$

As t^* is the point on the geodesic $\gamma_{\mu\to M}$ onto which X is projected orthogonally, removing variance in a direction $\gamma_{\mu\to M}$ in this setting would mean that the t^* that solves the equation needs to be zero.

As $\gamma(0) = \mu$, this reduces the equation to

$$< \mathrm{Log}_{\mu}(\gamma_{\mu\to M}(t)), \mathrm{Log}_{\mu}(\psi_M(X)) >_{\mu} = 0. \tag{6.1.7}$$

Without loss of generality, we can set $t = 1$ to achieve

$$< \mathrm{Log}_{\mu}(M), \mathrm{Log}_{\mu}(\psi_M(X)) >_{\mu} = 0. \tag{6.1.8}$$

Thus, the ψ that removes variance in direction $\gamma_{\mu\to M}$ is a function that transforms $X \in \mathcal{P}^n$ into a point whose tangent space projection at the mean μ is orthogonal to the tangent space projection of the geodesic-defining direction M.

In Euclidean space, this variance-removing function is unique per the Hilbert projection theorem and corresponds to the expression in eq. (8.2.5). In curved spaces, it is possible to define (at least) two functions that remove the variance in a given direction.

We state two candidate functions for ψ. The first function ψ^1 can be seen as the tangent space equivalent to the Euclidean case eq. (8.2.5) of removing variance in a given direction. It is constructed as

$$\psi^1_M(X) = \mathrm{Exp}_{\mu}(\mathrm{Log}_{\mu}(X) - < \mathrm{Log}_{\mu}(X), \mathrm{Log}_{\mu}(M) >_{\mu} \mathrm{Log}_{\mu}(M)) \tag{6.1.9}$$

and corresponds to removing the variance on the tangent space (on which orthogonality is defined) and mapping the result back to the manifold with the exponential map. It is independent of the parameter t^* for the projection of X onto M.

Proposition 6.1.10. *ψ^1 removes the variance in the direction of M.*

Proof. We want to show that the t^* for $\psi^1_M(X)$ is zero, i.e., M and $\psi^1_M(X)$ are orthogonal at the mean.

$$\begin{aligned}
&< \mathrm{Log}_{\mu}(M), \mathrm{Log}_{\mu}(\mathrm{Exp}_{\mu}(\mathrm{Log}_{\mu}(X) - < \mathrm{Log}_{\mu}(X), \mathrm{Log}_{\mu}(M) >_{\mu} \mathrm{Log}_{\mu}(M))) >_{\mu} && (6.1.11)\\
&= < \mathrm{Log}_{\mu}(M), \mathrm{Log}_{\mu}(X) - < \mathrm{Log}_{\mu}(X), \mathrm{Log}_{\mu}(M) >_{\mu} \mathrm{Log}_{\mu}(M) >_{\mu} && (6.1.12)\\
&= < \mathrm{Log}_{\mu}(M), \mathrm{Log}_{\mu}(X) >_{\mu} - < \mathrm{Log}_{\mu}(X), \mathrm{Log}_{\mu}(M) >_{\mu} \mathrm{Log}_{\mu}(M)\mu && (6.1.13)\\
&= < \mathrm{Log}_{\mu}(M), \mathrm{Log}_{\mu}(X) >_{\mu} - < \mathrm{Log}_{\mu}(M), < \mathrm{Log}_{\mu}(X), \mathrm{Log}_{\mu}(M) >_{\mu} \mathrm{Log}_{\mu}(M) >_{\mu} && (6.1.14)\\
&= < \mathrm{Log}_{\mu}(M), \mathrm{Log}_{\mu}(X) >_{\mu} - < \mathrm{Log}_{\mu}(X), \mathrm{Log}_{\mu}(M) >_{\mu} < \mathrm{Log}_{\mu}(M), \mathrm{Log}_{\mu}(M) >_{\mu} && (6.1.15)\\
&= < \mathrm{Log}_{\mu}(M), \mathrm{Log}_{\mu}(X) >_{\mu} - < \mathrm{Log}_{\mu}(X), \mathrm{Log}_{\mu}(M) >_{\mu} && (6.1.16)\\
&= 0. && (6.1.17)
\end{aligned}$$

□

A second function ψ^2 can be constructed as the manifold equivalent of variance removal using manifold transport defined in section 3.5. The underlying parallel transport $\mathrm{PT}^{\times}$ is the Levi-Civita parallel transport along a geodesic:

$$\psi^2_M(X) = \mathrm{MT}_{\gamma_{\mu\to M}(t^*)\to\mu}(X) = \mathrm{Exp}_\mu(\mathrm{PT}^{\times}_{\gamma_{\mu\to M}(t^*)\to\mu}(\mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(X)))$$

Proposition 6.1.18. *$\psi^2_M(X)$ removes the variance in direction M.*

Proof. We know that the orthogonality constraint

$$< \mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(\gamma_{\mu\to M}(t)), \mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(X) >_{\gamma_{\mu\to M}(t^*)} = 0$$

holds for all $t \in \mathbb{R}$. By leveraging the geodesic expression $\gamma_{\mu\to M}(t) = \mathrm{Exp}_\mu(t\,\mathrm{Log}_\mu(M))$ we can rephrase any tangent vector of the geodesic $\gamma_{\mu\to M}(t)$ on the tangent space of that same geodesic by

$$\mathrm{Log}_\mu(\gamma_{\mu\to M}(t)) = \mathrm{Log}_\mu(\mathrm{Exp}_\mu(t\,\mathrm{Log}_\mu(M))) = t\,\mathrm{Log}_\mu(M).$$

Thus, the tangent vector of any point of the geodesic at any other point of the geodesic is a scalar multiple of the logarithmic map at the starting point. As the Levi-Civita parallel transport along a geodesic preserves inner products, it preserves the angles between the tangent vector of the geodesic to the transported point at every tangent space along the geodesic, proving the proposition. □

As for PCA, the goal is to find a direction that minimizes the remaining variance. Thus, given a variance-removing function ψ, the optimization problem can be stated as

$$M_i = \underset{M, v_i \in V_{i-1}^{\perp}}{\arg\min} \sum_j d(\mu, \psi_M(X_j))^2$$

where $v_i := \mathrm{Log}_\mu(M_i)$.

6.1.2 Comparison of Principal Geodesic Analysis Optimization Problems

Contrary to the Euclidean case, the optimization problems of PGA do not generally yield the same solution [55]. A Problem with curved manifolds $\mathcal{M}$ is that the sum of variances along all orthogonal directions does not sum up to the total sum of variance, i.e., for an orthonormal basis $\{v_1, \ldots, v_d\}$ on $T_\mu\mathcal{M}$

$$\sum_i^n d^2(\mu, X_i) = \sum_j^d \sum_i^n d^2(\mu, \pi_{v_j}(X_i)) \tag{6.1.19}$$

does not generally hold [5, Sec. 2.3.1] for arbitrary manifolds $\mathcal{M}$. This corresponds to the applicability of the Pythagorean theorem, which does not hold on curved spaces. Thus, the first two optimization problems of variance maximizing direction and minimizing projection error are not equivalent. Additionally, for the optimization problem of minimizing total residual variance, the variance removal functions do not guarantee that only variance in the given direction is removed. For ψ^1, we can analytically solve the optimization

problem:

$$
\begin{aligned}
&\arg\min_M \sum_i d(\mu, \psi_M^1(X_i))^2 \\
=&\arg\min_M \sum_i < \mathrm{Log}_\mu(\psi_M^1(X_i)), \mathrm{Log}_\mu(\psi_M^1(X_i)) >_\mu \\
=&\arg\min_M \sum_i < \mathrm{Log}_\mu(X) - < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu \mathrm{Log}_\mu(M), \\
&\mathrm{Log}_\mu(X) - < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu \mathrm{Log}_\mu(M) >_\mu \\
=&\arg\min_M \sum_i < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(X) >_\mu \\
&-2 < \mathrm{Log}_\mu(X), < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu \mathrm{Log}_\mu(M) >_\mu \\
&+ << \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu \mathrm{Log}_\mu(M), < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu \mathrm{Log}_\mu(M) >_\mu \\
=&\arg\min_M \sum_i < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(X) >_\mu \\
&-2 < \mathrm{Log}_\mu(X), < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu \mathrm{Log}_\mu(M) >_\mu \\
&+ << \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu \mathrm{Log}_\mu(M), < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu \mathrm{Log}_\mu(M) >_\mu \\
=&\arg\min_M \sum_i < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(X) >_\mu \\
&-2 < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu \\
&+ < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu^2 < \mathrm{Log}_\mu(M), \mathrm{Log}_\mu(M) >_\mu \\
=&\arg\min_M \sum_i < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(X) >_\mu \\
&- < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu^2 \\
=&\arg\min_M - \sum_i < \mathrm{Log}_\mu(X), \mathrm{Log}_\mu(M) >_\mu^2 .
\end{aligned}
$$

Rewriting this in the notation of Riemannian linear features yields

$$\arg\min_M - \sum_i < \mathrm{RLF}_\mu(X), \mathrm{RLF}_\mu(M) >^2 . \tag{6.1.20}$$

As $\mathrm{RLF}_\mu(M)$ is the vector we are looking for and needs to be unit length, we will write $\tilde{M} = \mathrm{RLF}_\mu(M)$ and optimize for it

$$\arg\min_{\tilde{M}, ||\tilde{M}||_\mu = 1} - \sum_i < \mathrm{RLF}_\mu(X), \tilde{M} >^2 . \tag{6.1.21}$$

This corresponds to the optimization problem of PCA on Riemannian linear features. To obtain the geodesic direction, we set $M = \mathrm{RLF}_\mu^{-1}(\tilde{M})$.

For ψ_2, we can simplify

$$
\begin{aligned}
&f_\mu^{-1}(\mathrm{Log}_\mu(\mathrm{Exp}_\mu(f_\mu(f_{\gamma_{\mu\to M}(t^*)}^{-1}(\mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(X)))))) && (6.1.22)\\
=&f_{\gamma_{\mu\to M}(t^*)}^{-1}(\mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(X)) && (6.1.23)
\end{aligned}
$$

where the sections f are chosen such that $\mathrm{PT}^{\times}_{\gamma_{\mu\to M}(t^*)\to\mu}(X) = f_\mu f^{-1}_{\gamma_{\mu\to M}(t^*)}(X)$. Thus, we get the optimization objective

$$\arg\min_M \sum_j d(\mu, \psi^2_M(X_j))^2 \tag{6.1.24}$$

$$= \arg\min_M \sum_j < \mathrm{Log}_\mu(\psi^2_M(X_j)), \mathrm{Log}_\mu(\psi^2_M(X_j)) >_\mu \tag{6.1.25}$$

$$= \arg\min_M \sum_j < f^{-1}_{\gamma_{\mu\to M}(t^*)}(\mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(X_j)), f^{-1}_{\gamma_{\mu\to M}(t^*)}(\mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(X_j)) > \tag{6.1.26}$$

$$= \arg\min_M \sum_j < \mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(X_j), \mathrm{Log}_{\gamma_{\mu\to M}(t^*)}(X_j) >_{\gamma_{\mu\to M}(t^*)} \tag{6.1.27}$$

$$= \arg\min_M \sum_j d(X_j, \gamma_{\mu\to M}(t^*))^2 \tag{6.1.28}$$

$$\tag{6.1.29}$$

which corresponds to the optimization problem of minimizing projection error. We are thus left with only two new distinct optimization problems that need to be solved for PGA: finding geodesic directions that maximize variance or minimize projection error.

It is possible to define other notions of PGA that differ more substantially from PCA [5, Sec. 2.3.4], e.g. by not restricting principal geodesics to pass through the intrinsic mean of a dataset [55]. As such, care must be taken when calculating PGA, as it needs to be explicitly stated which optimization problem is addressed. Depending on the task at hand, one of the problems might prove better suited for analysis. The following section will show possible ways to obtain (approximate) solutions to PGA problems.

6.1.3 Approximate Principal Geodesic Analysis

As the solution to the PGA optimization problems may not be obtained analytically, the authors of the original paper introducing PGA [24] approximate the projection operator $\pi(x)$ using the tangent space $T_\mu\mathcal{M}$ at the Fréchet mean μ of any Riemannian manifold $(\mathcal{M}, g)$. The idea is that the geodesic is viewed as a linear vector on the tangent space of the Frechét mean μ and that the projection is performed onto the linearized geodesic rather than the true geodesic on the manifold. This approximate analytic expression for the first principal component is then given by

$$v_1 \approx \arg\max_{||v||=1} \sum_i^N < v, \mathrm{Log}_\mu(x_i) >^2 . \tag{6.1.30}$$

This is simply linear PCA on the tangent space $T_\mu\mathcal{M}$ and was long referred to as PGA, without noting that it was only an approximation. In accordance with other literature in the field, e.g., [55], we will henceforth refer to this type of PGA as *linearized* PGA, as it performs PCA on the tangent space, which is a linear approximation of the manifold.

There is uncertainty and inconsistency in the literature concerning linearized PGA. It is unclear whether the inner product in eq. (6.1.30) for the estimation of the principal components is the standard Euclidean inner product or, as would be more suited as we will see later, the Riemannian metric to essentially perform linear PCA on Riemannian linear features. Starting with the original paper [24] introducing PGA; In algorithm 2, the standard Euclidean inner product on the tangent vectors is used. This is consistent with the derivation in

the text preceding the algorithm. In a second paper by the main author of the original paper that was published three years later [23, Eq. 6], the objective function of PGA is approximated by the Riemannian metric rather than the standard Euclidean inner product. In a more recent book from 2020 co-authored by the same author, the algorithm of PGA is again stated in terms of the standard Euclidean inner product on the tangent space [5, Ch. 2, Algorithm 2.2]. For the remainder of the thesis, we will assume that the standard Euclidean inner product is used on the tangent space.

6.1.4 Exact Principal Geodesic Analysis

The approximate solution to the PGA optimization problem by linearized PGA may differ substantially from the actual solution. Many approaches have been tried to find the exact solution to the PGA optimization problem to cure the shortcoming of linearized PGA's approximative nature. [57, Sec. 5] compares linearized PGA and exact PGA, demonstrating a dependence of approximation quality on the manifold's curvature.

In general, calculating the principal geodesics of a generic manifold is a more intricate matter.

[55] provide an algorithm for performing exact PGA on various manifolds. However, this algorithm is computationally expensive and may prove unfeasible in many cases. [12] give an efficient algorithm, but their approach is limited to constant curvature manifolds.

We show that by the isometry to Euclidean space of flat manifolds, the principal geodesics coincide with the principal directions of centered data on Riemannian linear feature space. Additionally, we give an algorithmic approach to solving the projection problem of PGA using the Affine Invariant metric and derive a bound on the variance.

6.1.4.1 Flat Metrics

Flat manifolds are isometric to standard Euclidean space and thus allow for an analytic solution of PGA that coincides with PCA. We showed in lemma 3.7.2 that the pairwise distances in Riemannian linear feature space coincide with the pairwise distances on the manifold. Thus, for flat metrics, we can equivalently frame the PGA optimization problem as PCA on Riemannian linear feature space.

The assumption of centered data of PCA can either be fulfilled by direct projection onto the Riemannian linear feature space of the Frechét mean (see lemma 3.6.9) or by projecting to any Riemannian linear feature space and re-centering in Euclidean space. As by eq. (6.1.21), the variance removal objective of ψ^1 leads to PCA on Riemannian linear feature space for any metric, all optimization objectives of PGA are equivalent for flat metrics. t^* is given as

$$t^* =< \mathrm{RLF}_\mu(X), \mathrm{RLF}_\mu(M) > \tag{6.1.31}$$

and the projection operator is given as

$$\pi_M(X) = \mathrm{RLF}_\mu^{-1}(< \mathrm{RLF}_\mu(X), \mathrm{RLF}_\mu(M) > \mathrm{RLF}_\mu(M)). \tag{6.1.32}$$

In this case, exact PGA coincides with kernel PCA using the canonical kernels and PCA on Riemannian linear feature space.

We would like to point out that, despite the given equivalences for flat metrics, linearized PGA (PCA on the tangent space of the mean μ) and exact PGA (PCA on the Riemannian linear feature space) do not generally produce the same solution.

6.1.4.2 Curved Metrics

The different optimization problems of PGA do not generally have the same solution for non-flat metrics [55, Sec. 5]. As the Affine Invariant metric is the only geodesically complete non-flat metric in consideration, we will restrict the analysis to this metric.

Algorithmic approach for the Affine Invariant metric. As noted before, the solution to PGA on curved manifolds needs to be found iteratively. We will simplify the target function for the Affine Invariant metric as much as possible. We can exploit that the pairwise distances between points on the manifold does not change for manifold transported data with $\mathrm{MT}_{P\to Q}(A)$, see lemma 3.7.9. We can thus assume a Frechét mean $P = I$ without loss of generality. This ensures a simpler expression for geodesic directions M by $\gamma_{I\to M}(t) = M^t$. The first step is to find a t^* corresponding to some direction $M \in \mathcal{P}^n$ so that, given any data point $X \in \mathcal{P}^n$, M^{t^*} is the orthogonal projection of X onto the direction M^t.

The orthogonality equation is given as

$$< \mathrm{Log}_{M^{t^*}}(M^t), \mathrm{Log}_{M^{t^*}}(X) >_{M^{t^*}} = 0$$

with $t \neq t^*$.

The expression can be simplified by inserting the Affine Invariant inner product and logarithmic map. Let $M = V\Lambda V^T$ be the eigendecomposition of M. Then with $t = t^* + 1$ we get

$$\begin{aligned}
0 = &\operatorname{tr}(M^{-t^*}M^{t^*/2}\operatorname{Log}(M^{-t^*/2}M^tM^{-t^*/2})M^{t^*/2}M^{-t^*}M^{t^*/2}\operatorname{Log}(M^{-t^*/2}AM^{-t^*/2})M^{t^*/2})\\
&\operatorname{tr}(M^{-t^*/2}\operatorname{Log}(M^{-t^*/2}M^tM^{-t^*/2})\operatorname{Log}(M^{-t^*/2}AM^{-t^*/2})M^{t^*/2})\\
&\operatorname{tr}(\operatorname{Log}(M^{-t^*/2}M^tM^{-t^*/2})\operatorname{Log}(M^{-t^*/2}AM^{-t^*/2}))\\
&\operatorname{tr}(\operatorname{Log}(M^{t-t^*})\operatorname{Log}(M^{-t^*/2}AM^{-t^*/2}))\\
&\operatorname{tr}(\operatorname{Log}(M)\operatorname{Log}(M^{-t^*/2}AM^{-t^*/2}))\\
&\operatorname{tr}(\operatorname{Log}(V\Lambda V^T)\operatorname{Log}(V\Lambda^{-t^*/2}V^TAV\Lambda^{-t^*/2}V^T))\\
&\operatorname{tr}(V\operatorname{Log}(\Lambda)V^TV\operatorname{Log}(\Lambda^{-t^*/2}V^TAV\Lambda^{-t^*/2})V^T)\\
&\operatorname{tr}(\operatorname{Log}(\Lambda)\operatorname{Log}(\Lambda^{-t^*/2}V^TAV\Lambda^{-t^*/2}))
\end{aligned}$$

By this simplification, each update in an iterative optimization procedure only needs to perform a single eigendecomposition to calculate $\mathrm{Log}(\Lambda^{-t^*/2}V^TAV\Lambda^{-t^*/2})$ for updated values of t.

We can derive an upper bound for $|t^*|$ for the Affine Invariant metric by its Log-Euclidean equivalent. It always has the same sign. This upper bound can then be used as a starting point for iterative algorithms using Affine Invariant PGA.

Proposition 6.1.33. *$|t^*_{AIR}| \leq |t^*_{LE}|$ and $sign(t^*_{AIR}) = sign(t^*_{LE})$*

Proof. Let d_{AIR} and d_{LE} be the geodesic distances of the Affine Invariant and the Log-Euclidean metric. $d_{AIR}(I, A) = d_{LE}(I, A)$. Thus, for any $M \in \mathcal{P}^n$, $d_{AIR}(M^0, A) = d_{LE}(M^0, A)$. As by theorem 3.7.10 $d_{AIR}(M^t, A) \geq d_{LE}(M^t, A)$ for any $t \neq 0$ and $d_{AIR}(M^t, A)$ and $d_{LE}(M^t, A)$ are convex functions of t, we can infer that $\frac{d}{dt}d_{AIR}(M^t, A) = \frac{d}{dt}d_{LE}(M^t, A)$ for $t = 0$, $\frac{d}{dt}d_{AIR}(M^t, A) > \frac{d}{dt}d_{LE}(M^t, A)$ for $t > 0$ and $\frac{d}{dt}d_{AIR}(M^t, A) < \frac{d}{dt}d_{LE}(M^t, A)$ for $t < 0$. $t = t^*$ is the point where $\frac{d}{dt}d(M^t, A) = 0$. We can conclude that $\frac{d}{dt}d_{AIR}(M^t, A) = 0$ for a t that is closer to 0 than that for which $\frac{d}{dt}d_{LE}(M^t, A) = 0$ for any $M, A \in \mathcal{P}^n$. □

The optimization problem to find t^* for the Affine Invariant metric is convex and can be solved by standard numerical solvers.

6.1.5 Manifold Linearization with Principal Geodesic Analysis

In addition to the tangent space and Riemannian linear feature space, we can define a third way of manifold linearization by leveraging principal geodesics.

Given n principal geodesics $(M_i)_{i=1,\dots,n}$, the linearization of a data point $A \in \mathcal{P}^n$ is given as

$$\tilde{A} = \left[t_1^*, t_2^*, \dots, t_n^*\right] \tag{6.1.34}$$

where t_i^* is the solution to $t_i^* = \arg\min_t d(\gamma_{\mu \to M_i}(t), A)^2$ where d is the geodesic distance for a given Riemannian metric on $\mathcal{P}^n$. $\tilde{A} \in \mathbb{R}^{\frac{n(n+1)}{2}}$ is now a $\frac{n(n+1)}{2}$-dimensional vector and can be evaluated in standard Euclidean space under the standard Euclidean inner product.

6.1.6 Relationship Between PCA, Kernel PCA and PGA

Canonical kernels can be described as linear kernels on the Riemannian linear feature space definition 3.6.1. For all metrics, the formulations of kernel PCA with the canonical kernel and linear PCA on Riemannian linear feature space at the Frechét mean are equivalent. As the resulting kernel matrix for canonical kernels is already centered by lemma 3.6.9, the re-centering step of kernel PCA can be omitted. For flat metrics, PGA in all its formulations is equivalent to PCA on Riemannian linear features.

PGA produces different results for non-flat metrics for maximizing variance and minimizing projection error [55, Sec. 5]. Only total residual variance minimization by the tangent space variance removal function ψ^1 is equivalent to PCA on Riemannian linear features. The other objectives lack a simple analytical solution and must be solved iteratively.

6.2 Methods

We found that there are two distinct optimization objectives that produce different results for PGA using curved metrics. The Affine Invariant metric is the only proposed geodesically complete curved metric. Thus, all analysis in this section is limited to the Affine Invariant metric. We compare the performance of the proposed linearization using PGA to the linearization using Riemannian linear features. Additionally, we test the dimensionality reduction using PGA and compare it to dimensionality reduction with kernel PCA using the canonical kernel. The analyses are performed similar to the linearization analysis in section 4.1 and the dimensionality reduction analysis in section 5.4.2.

As solving the PGA optimization problem for the Affine Invariant metric is computationally expensive, we compare the two different PGA optimization methods with only one dataset. For computational reasons, we chose the dataset with the lowest number of channels, the BNCI2014-004 dataset (see table 8.1).

For linearization and dimensionality reduction testing, we compare the distance geometry from the PGA linearization and dimensionality reduction under the Euclidean metric to the true distance geometry of the manifold under the geodesic distance of the Affine Invariant metric with the AUClogRNX score (see section 8.6). We compare the linearization with PGA to Riemannian linear features and the dimensionality reduction with principal geodesic analysis to the reduction with kernel PCA using the canonical kernels.

6.3 Results

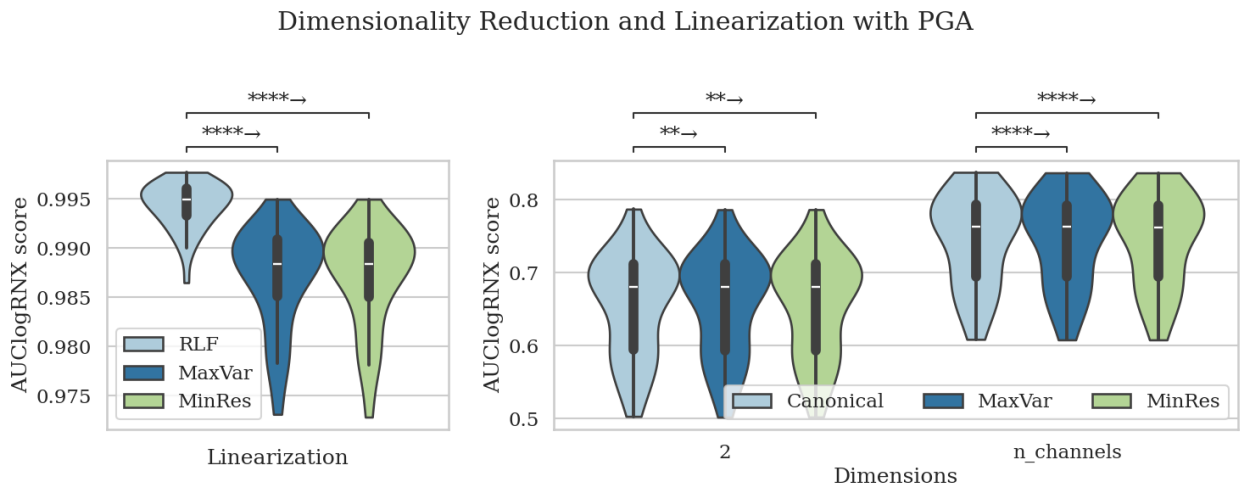

Figure 6.1: **Comparison of the distance geometry of PGA linearization and dimensionality reduction with the manifold.** MaxVar denotes the optimization by maximum variance whereas MinRes denotes the optimization by minimum residual error. **Left:** AUClogRNX score of linearization using PGA with the Affine Invariant metric compared to linearization with Riemannian linear features (RLF). Significance analysis was performed between all methods. **Right:** AUClogRNX score of dimensionality reduction using PGA with the Affine Invariant metric. Dimensionality reduction is performed to 2 and $n_{channels}$ dimensions, where $n_{channels}$ is the number of channels in the recording. The ground truth is the neighborhood structure using the geodesic distance on the manifold. The PGA dimensionality reduction results are compared to dimensionality reduction by canonical kernel PCA. A significance analysis (for details, see section 8.5) was performed between all methods within each dimension.

Figure 6.1 shows no significant difference in the two PGA optimization methods regarding dimensionality reduction and linearization quality measured by AUClogRNX. The median difference between the two PGA optimization objectives is 0.0005 p.p. for the linearization, 0.04 p.p. for reduction to two dimensions and 0.12 for reduction to $n_{channels}$ dimensions.

Linearization and dimensionality reduction using Riemannian linear features significantly outperforms PGA linearization for both optimization objectives. The difference in median AUClogRNX score is 0.6 p.p. for the linearization, 0.04 p.p. for the dimensionality reduction to two dimensions and 0.08 for the dimensionality reduction to $n_{channels}$ dimensions.

6.4 Discussion

6.4.1 Theoretical Results

The uncertainty concerning which inner product should be used in the approximate solutions to PGA can be clarified by considering that for flat metrics, using the Riemannian inner product rather than the tangent space

projection, recovers exact PGA. This can be transferred to curved metrics considering that the approximation of PGA coincides with PCA on Riemannian linear feature space. Riemannian linear features space better approximates the distance geometry of the manifold that the tangent space viewed as standard Euclidean space, as we saw in chapter 4.

We introduced a fast update rule for the iterative algorithm for the projection operator of PGA under the Affine Invariant metric. This speeds up the iterative process compared to implementing general methods, as e.g. described in [55].

It has not previously been noted in the literature that PGA can also be used as a linearization method on the manifold. Although this linearization did not show any superiority to linearization using Riemannian linear features, it may present a helpful tool for general assessment of the quality of linearization possibilities.

6.4.2 Linearization and Dimensionality Reduction

Following our analysis, the results of PGA in terms of linearization and dimensionality reduction do not indicate any superiority of PGA-based methods in the application to positive definite matrices. However, we need to point out that the meaningfulness of the analysis is limited. Due to PGA's computational complexity, we tested the method only on the dataset with the lowest number of channels. This dataset consisted of only nine subjects, making up a small subset of the available datasets used to compare other methods. This has several implications. The spread of the data on the manifold tends to be less in lower dimensions than in covariance data from a high number of channels. This hints at a good approximation ability of the manifold linearization using Riemannian linear features. Thus, we expect the principal geodesics and principal components to point in similar directions. In addition, we expect the solutions to the different optimization problems of PGA to be very close. With more channels, PGA could potentially outperform the linearization and dimensionality reduction on Riemannian linear features. The iterative nature of PGA estimation may introduce numerical fluctuations that can further reduce the exactness of the PGA solution, making it non-optimal.

In general, we would expect the minimization of projection error to produce better results for linearization and dimensionality reduction, as this objective directly optimizes for the projection quality of the principal geodesics.

7 Conclusion

In this chapter, we state this thesis's limitations, describe the various future research possibilities that the presented derivations offer, and conclude the thesis with closing words.

7.1 Limitations

This thesis was designed to be of a theoretical and descriptive nature to formalize kernel methods for SPD matrices. The analyses were constructed as self-contained comparisons of a certain class of kernels that could be described using this formalization. Thus, it contains only a limited selection of kernel methods in a limited number of tasks. This means that the proposed methods were not compared to other methods in the literature. Although this comparison with other methods would have been favorable in asserting the feasibility of proposed methods for given tasks, it would have diverted the focus of this thesis from the clear formalization of kernels and more basic concepts of Riemannian manifolds.

The applied analyses of this thesis were limited to the task of motor imagery on EEG data. In this task, we found that linear kernels on Riemannian linear features outperform metric space kernels while being computationally more efficient to implement and having fewer parameters. However, this cannot be considered a general fact, as other tasks may produce different results and favor metric space kernels.

There is no sound conclusion for manifold linearization and dimensionality reduction with principal geodesic analysis. Only one low-dimensional dataset was used, making this more speculative. The analysis's meaningfulness is limited due to PGA's high computational complexity. Higher-dimensional datasets may offer new insights into PGA results, but they require more computational power.

In theory, there are infinite possibilities to combine kernels and construct new positive definite kernels. As we found testing a larger number of kernel combinations with potentially limited additional value unfeasible, we restricted ourselves to easy-to-use and simply transferrable kernel methods for Riemannian manifolds. It could be that a specific kernel combination proves to be a good kernel of choice for a given task on a Riemannian manifold of SPD matrices.

7.2 Outlook

The formalization introduced in this thesis offers a wide range of theoretical and applied new research directions that merit closer inspection.

7.2.1 Simplified Introduction to Riemannian Geometry

We attempted to derive the theory of Riemannian geometry more simply than is done in many mathematical textbooks covering the subject (cf., e.g., [11, 37, 36]). We introduced the theory including curvature and the

Levi-Civita parallel transport without touching on advanced topics like affine connections or the Riemannian curvature tensor but rather focused on the necessary features to describe curvature as has become relevant in the following derivations. Rather than fully describing parallel transport theory, we just specified one single function of interest and described it axiomatically. However, in the derivation of a differentiable manifold, we still cover advanced and abstract topics. When restricting oneself to the set of SPD matrices, this introduction could be simplified even further. This would greatly help make this highly mathematical topic more accessible to a wider, less mathematically trained audience, allowing for a broader application of methods while still building on a sound mathematical basis.

7.2.2 Sections

Considering the manifold of SPD matrices, we are luckily provided with evident candidate sections. We can work with these sections and try to achieve new formalizations of Riemannian metrics that allow to more directly connect sections and parallel transport.

For example, we found the following form that can be achieved for all proposed Riemannian metrics on SPD. Let $P = V\Lambda V^T$ be the eigendecomposition of P. Any metric on $\mathcal{P}^n$ can be written with a section as

$$f_P^{-1}(X) = V(L \odot (V^T X V))V^T$$

with some L that depends only the eigenvalues of P. This form of sections can be seen as an orthogonal transformation towards the reference point P, Hadamard multiplication with a matrix constructed from the eigenvalues of the reference point and the inverse rotation. This could hint at a general structure to obtain the section corresponding to the Levi-Civita parallel transport to the identity. For the Bures-Wasserstein metric, the transformation is obvious by $f_P^{-1}(A) = \frac{1}{\sqrt{2}} V_P(L_P \odot (V_P^T A V_P))V_P^T$ where $[L_P]_{ij} = \frac{1}{\sqrt{[\lambda_P]_i + [\lambda_P]_j}}$.

Rewriting the Affine Invariant metric with $[D^*]_{ij} = \frac{1}{\sqrt{\lambda_i \lambda_j}}$

$$\begin{aligned} f_P^{-1}(A) &= P^{-\frac{1}{2}} A P^{-\frac{1}{2}} && (7.2.1)\\ &= V\Lambda_P^{-\frac{1}{2}}(V^T A V)\Lambda_P^{-\frac{1}{2}} V^T && (7.2.2)\\ &= V(D^* \odot (V^T A V))V^T && (7.2.3) \end{aligned}$$

shows the same underlying structure.

All pullback metrics follow the same structure, as is evident by the definition of the directional derivative in eq. (8.1.38), which defines the Riemannian metric tensor of pullback metrics. Interestingly, this notation corresponds to the sections that can be used to construct the Levi-Civita parallel transport along a geodesic.

Another possible result from this form could be the indication of the best-approximating flat metric for curved metrics. For the Affine Invariant metric, it has been shown that the Log-Euclidean metric is a good approximation. Considering the diagonal of $[D^*]_{ij} = \frac{1}{\sqrt{\lambda_i \lambda_j}}$ we get $[D^*]_{ii} = \frac{1}{\lambda_i}$. This corresponds to the diagonal of the representation of the Log-Euclidean metric in terms of the first divided difference. For the Bures-Wasserstein metric, the diagonal of $[L_P]_{ij}$ is given as $[L_P]_{ii} = \frac{1}{\sqrt{2\lambda_i}}$. The corresponding pullback metric with the same diagonal of the first divided difference is the Power-Euclidean metric (cf. [17] with exponent $\frac{1}{2}$). In [9] this relationship is indicated, but it is not explicitly stated that the Power-Euclidean metric is used as an approximation of the Bures-Wasserstein metric. We have performed tests that show that the

exponent $\frac{1}{2}$ for the Power-Euclidean metric best approximates the Bures-Wasserstein metric. However, these tests have not been extensive and this merits further research.

This form of describing a Riemannian metric on SPD could potentially be extended to general Riemannian metrics and would then allow the search for flat metrics that best approximate any given curved metric.

7.2.3 Riemannian Linear Features

The concept of Riemannian linear features is not limited to describing the Riemannian manifold for SPD matrices. We showed that this concept potentially provides a much better approximation of the manifold than the tangent space under the standard Euclidean inner product chapter 4. Thus, this concept may benefit other Riemannian manifolds, given that an evident section can be obtained.

Further research could include finding the orthogonal transformation that describes an equivalent concept to the Levi-Civita parallel transport along a geodesic in terms of the Riemannian linear feature space by finding an orthogonal transformation applied after calculating the features. This transformation can solely depend on the starting point, as this is the only parametrization of the Riemannian linear features. For the Bures-Wasserstein metric (as well as the Affine Invariant metric using the new notation of sections introduced earlier), this transformation is simply a rotation by the orthogonal matrix obtained by eigendecomposition.

Additionally, the Riemannian linear feature space allows for comparing the structure of different Riemannian metrics in a common space, the standard Euclidean space. This could be used to assess similarities and differences between Riemannian metrics and may facilitate the search for approximations of curved metrics.

7.2.4 Canonical Kernels

The concept of canonical kernels as linear kernels on Riemannian linear feature space has produced good results, allows for a concise mathematical description and shows good computational efficiency. This concept should be evaluated in other tasks, with different data, and potentially on other Riemannian manifolds.

Concerning dimensionality reduction using canonical kernel PCA, it would be interesting to compare this method with other dimensionality reduction methods, e.g. in [28, 26].

7.2.5 Riemannian Manifolds as Metric Spaces

Further research in this direction could include finding positive definite metric space kernels for geodesic distances of curved metrics. This would enable leveraging the geodesic distance on the manifold rather than relying on its linearization to ensure positive definite kernels. This could be achieved by searching for metric transforming function f that allows for embedding a given Riemannian metric in Hilbert space or specifying parameter ranges for which a given metric space kernel is positive definite on a given Riemannian manifold, as was achieved for the Gaussian kernel using the LogDet distance. This is an active field of research (see, e.g., [22]), and it may well be that it is impossible to find these transformations or parameter ranges, but rather, it could be proven that these cannot exist.

7.3 Closing Words

We presented an introduction and formalization of kernel methods on the Riemannian manifold of symmetric positive definite matrices. While working towards this goal, we created a more extensive formalization of the general properties of the Riemannian manifold. The basic tool that allowed this formalization is the existence of evident sections of the orthonormal frame bundle for any of the proposed Riemannian metrics. By introducing the formalization of Riemannian linear features, we were able to substantially reduce the page load of this thesis by simplifying proofs and finding connections between kernel methods. We also created a tool that allows for a more geometry-aware linearization of curved manifolds than simple tangent space projection. As a side effect, it also helps in understanding the notion of a flat manifold as being a Euclidean space using a global isometry ϕ. There should always be a clear distinction between the Riemannian linear feature space and the tangent space to avoid confusion about which feature is used in any given context. The Riemannian linear feature space should be the linear space of choice when using machine learning as it more closely resembles the distance geometry on the manifold.

Concerning the implemented kernel methods, the canonical kernels, shown to be linear kernels in Riemannian linear feature space, perform well when compared to (linearized) metric space kernels. These kernels can be calculated very efficiently by commonly used machine learning libraries. We suggest this method as a first starting point for analysis when a low-complexity method is needed that can capture more complex data dependencies than the minimum distance to mean classifier. Also, for dimensionality reduction, the close resemblance to PGA suggests that kernel PCA under canonical kernels provides a good low-dimensional approximation of the manifold. If a metric space kernel is needed, the usually promoted choice using the Gaussian kernel seems to also be a good choice for application on the Riemannian manifold of positive definite matrices.

In order to facilitate comparing kernel methods on Riemannian manifolds, we have contributed over 5,000 lines of efficient functional code to the public python library pyriemann [4] and have prepared the implementation of all new functional contributions of this thesis to be included after publication.

To conclude, we were able to create a general formalization of kernel methods on the Riemannian manifold of symmetric positive definite matrices that produces mathematically founded, geometry-aware kernels.

Bibliography

[1] Martial Agueh and Guillaume Carlier. “Barycenters in the Wasserstein Space”. en. In: *SIAM Journal on Mathematical Analysis* 43.2 (Jan. 2011), pp. 904–924. DOI: 10.1137/100805741.

[2] Vincent Arsigny et al. “Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices”. In: *SIAM Journal on Matrix Analysis and Applications* 29.1 (Jan. 2007), pp. 328–347. DOI: 10.1137/050637996.

[3] Alexandre Barachant et al. “Classification of covariance matrices using a Riemannian-based kernel for BCI applications”. en. In: *Neurocomputing*. Advances in artificial neural networks, machine learning, and computational intelligence 112 (July 2013), pp. 172–178. DOI: 10.1016/j.neucom.2012.12.039.

[4] Alexandre Barachant et al. *pyRiemann/pyRiemann: v0.8*. Feb. 2025. DOI: 10.5281/zenodo.14859219.

[5] Martin Bauer et al. *Riemannian Geometric Statistics in Medical Image Analysis*. Ed. by Xavier Pennec, Stefan Sommer, and Tom Fletcher. Academic Press, Jan. 2020. DOI: 10.1016/B978-0-12-814725-2.00005-4.

[6] C. van den Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. en. Springer Science & Business Media, Dec. 2012.

[7] Gabriel Leander Wagner vom Berg et al. “A New Canonical Log-Euclidean Kernel for Symmetric Positive Definite Matrices for EEG Analysis (Oct 2024)”. In: *IEEE Transactions on Biomedical Engineering* 72.3 (Mar. 2025), pp. 1000–1007. DOI: 10.1109/TBME.2024.3483936.

[8] Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

[9] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. “On the Bures–Wasserstein distance between positive definite matrices”. In: *Expositiones Mathematicae* 37.2 (June 2019), pp. 165–191. DOI: 10.1016/j.exmath.2018.01.002.

[10] A. Buttfield, P.W. Ferrez, and Jd.R. Millan. “Towards a robust BCI: error potentials and online learning”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14.2 (June 2006), pp. 164–168. DOI: 10.1109/TNSRE.2006.875555.

[11] Manfredo Perdigão do Carmo. *Riemannian geometry*. eng. Mathematics. Theory & applications. Boston: Birkhäuser, 1992.

[12] Rudrasis Chakraborty, Dohyung Seo, and Baba C. Vemuri. “An Efficient Exact-PGA Algorithm for Constant Curvature Manifolds”. en. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 3976–3984. DOI: 10.1109/CVPR.2016.431.

[13] Chih-Chung Chang and Chih-Jen Lin. “LIBSVM: A library for support vector machines”. In: *ACM Trans. Intell. Syst. Technol.* 2.3 (2011), 27:1–27:27. DOI: 10.1145/1961189.1961199.

[14] Sylvain Chevallier et al. *The largest EEG-based BCI reproducibility study for open science: the MOABB benchmark*. en. Apr. 2024. DOI: 10.48550/arXiv.2404.15319.

[15] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. “Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review”. In: *Brain-Computer Interfaces* 4.3 (2017), pp. 155–174. DOI: 10.1080/2326263X.2017.1297192.

[16] Paul Corazza. “Introduction to Metric-Preserving Functions”. In: *The American Mathematical Monthly* 106.4 (1999), pp. 309–323. DOI: 10.2307/2589554.

[17] Ian L. Dryden, Xavier Pennec, and Jean-Marc Peyrat. *Power Euclidean metrics for covariance matrices with application to diffusion tensor imaging*. Sept. 2010. DOI: 10.48550/arXiv.1009.3045.

[18] Manuel Eder, Jiachen Xu, and Moritz Grosse-Wentrup. “Benchmarking brain–computer interface algorithms: Riemannian approaches vs convolutional neural networks”. en. In: *Journal of Neural Engineering* 21.4 (Aug. 2024), p. 044002. DOI: 10.1088/1741-2552/ad6793.

[19] Josef Faller et al. “Autocalibration and Recurrent Adaptation: Towards a Plug and Play Online ERD-BCI”. In: *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society* 20 (Apr. 2012), pp. 313–9. DOI: 10.1109/TNSRE.2012.2189584.

[20] Rong-En Fan et al. “LIBLINEAR: A Library for Large Linear Classification”. In: *J. Mach. Learn. Res.* 9 (June 2008), pp. 1871–1874.

[21] Aasa Feragen and Søren Hauberg. “Open Problem: Kernel methods on manifolds and metric spaces. What is the probability of a positive definite geodesic exponential kernel?” en. In: *Conference on Learning Theory*. PMLR, June 2016, pp. 1647–1650.

[22] Aasa Feragen, François Lauze, and Søren Hauberg. “Geodesic exponential kernels: When curvature and linearity conflict”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, pp. 3032–3042. DOI: 10.1109/CVPR.2015.7298922.

[23] P. Thomas Fletcher and Sarang Joshi. “Riemannian geometry for the statistical analysis of diffusion tensor data”. In: *Signal Processing*. Tensor Signal Processing 87.2 (Feb. 2007), pp. 250–262. DOI: 10.1016/j.sigpro.2005.12.018.

[24] P.T. Fletcher et al. “Principal geodesic analysis for the study of nonlinear statistics of shape”. In: *IEEE Transactions on Medical Imaging* 23.8 (Aug. 2004), pp. 995–1005. DOI: 10.1109/TMI.2004.831793.

[25] Andi Han et al. “On Riemannian Optimization over Positive Definite Matrices with the Bures-Wasserstein Geometry”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 8940–8953.

[26] Mehrtash T. Harandi, Mathieu Salzmann, and Richard Hartley. “From Manifold to Manifold: Geometry-Aware Dimensionality Reduction for SPD Matrices”. en. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Vol. 8690. Cham: Springer International Publishing, 2014, pp. 17–32. DOI: 10.1007/978-3-319-10605-2_2.

[27] Mehrtash T. Harandi et al. “Sparse Coding and Dictionary Learning for Symmetric Positive Definite Matrices: A Kernel Approach”. en. In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012, pp. 216–229. DOI: 10.1007/978-3-642-33709-3_16.

[28] Inbal Horev, Florian Yger, and Masashi Sugiyama. “Geometry-aware principal component analysis for symmetric positive definite matrices”. en. In: *Machine Learning* 106.4 (Apr. 2017), pp. 493–522. DOI: 10.1007/s10994-016-5605-5.

[29] H. Hotelling. “Analysis of a complex of statistical variables into principal components.” en. In: *Journal of Educational Psychology* 24.6 (Sept. 1933), pp. 417–441. DOI: 10.1037/h0071325.

[30] Joonas Ilmavirta. *Geometry of geodesics*. en. July 2020. DOI: 10.48550/arXiv.2008.00073.

[31] Vinay Jayaram and Alexandre Barachant. “MOABB: trustworthy algorithm benchmarking for BCIs”. en. In: *Journal of Neural Engineering* 15.6 (Sept. 2018), p. 066011. DOI: 10.1088/1741-2552/aadea0.

[32] Sadeep Jayasumana et al. “Kernel Methods on Riemannian Manifolds with Gaussian RBF Kernels”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.12 (Dec. 2015), pp. 2464–2477. DOI: 10.1109/TPAMI.2015.2414422.

[33] Sadeep Jayasumana et al. “Kernel Methods on the Riemannian Manifold of Symmetric Positive Definite Matrices”. en. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, June 2013, pp. 73–80. DOI: 10.1109/CVPR.2013.17.

[34] Xuan Kan et al. “R-Mixup: Riemannian Mixup for Biological Networks”. en. In: *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining* 2023 (Aug. 2023), p. 1073. DOI: 10.1145/3580305.3599483.

[35] J. Kivinen, A.J. Smola, and R.C. Williamson. “Online Learning with Kernels”. en. In: *IEEE Transactions on Signal Processing* 52.8 (Aug. 2004), pp. 2165–2176. DOI: 10.1109/TSP.2004.830991.

[36] John M. Lee. *Introduction to Riemannian Manifolds*. Vol. 176. Graduate Texts in Mathematics. Cham: Springer International Publishing, 2018. DOI: 10.1007/978-3-319-91755-9.

[37] John M. Lee. *Introduction to Smooth Manifolds*. en. Vol. 218. Graduate Texts in Mathematics. New York, NY: Springer, 2012. DOI: 10.1007/978-1-4419-9982-5.

[38] Min-Ho Lee et al. “EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy”. In: *GigaScience* 8.5 (May 2019), giz002. DOI: 10.1093/gigascience/giz002.

[39] Robert Leeb et al. “Brain-computer communication: motivation, aim, and impact of exploring a virtual apartment”. eng. In: *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society* 15.4 (Dec. 2007), pp. 473–482. DOI: 10.1109/TNSRE.2007.906956.

[40] Zhenhua Lin. “Riemannian Geometry of Symmetric Positive Definite Matrices via Cholesky Decomposition”. In: *SIAM Journal on Matrix Analysis and Applications* 40.4 (Jan. 2019), pp. 1353–1370. DOI: 10.1137/18M1221084.

[41] Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. “Wasserstein Riemannian geometry of Gaussian densities”. en. In: *Information Geometry* 1.2 (Dec. 2018), pp. 137–179. DOI: 10.1007/s41884-018-0014-4.

[42] Hà Quang Minh and Vittorio Murino. *Covariances in Computer Vision and Machine Learning*. en. Synthesis Lectures on Computer Vision. Cham: Springer International Publishing, 2018. DOI: 10.1007/978-3-031-01820-6.

[43] Maher Moakher and Mourad Zerai. “The Riemannian Geometry of the Space of Positive-Definite Matrices and Its Application to the Regularization of Positive-Definite Matrix-Valued Data”. In: *Journal of Mathematical Imaging and Vision* 40 (June 2011), pp. 171–187. DOI: 10.1007/s10851-010-0255-x.

[44] Knut Pawel and Helmut Reckziegel. “Affine submanifolds and the Theorem of Cartan-Ambrose-Hicks”. In: *Kodai Mathematical Journal* 25.3 (Oct. 2002), pp. 341–356. DOI: 10.2996/kmj/1071674466.

[45] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (Nov. 1901), pp. 559–572. DOI: 10.1080/14786440109462720.

[46] Fabian Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830.

[47] Pedro Luiz Coelho Rodrigues, Christian Jutten, and Marco Congedo. “Riemannian Procrustes Analysis: Transfer Learning for Brain–Computer Interfaces”. en. In: *IEEE Transactions on Biomedical Engineering* 66.8 (Aug. 2019), pp. 2390–2401. DOI: 10.1109/TBME.2018.2889705.

[48] I. J. Schoenberg. “Metric Spaces and Completely Monotone Functions”. In: *Annals of Mathematics* 39.4 (1938), pp. 811–841. DOI: 10.2307/1968466.

[49] I. J. Schoenberg. “Metric Spaces and Positive Definite Functions”. In: *Transactions of the American Mathematical Society* 44.3 (1938), pp. 522–536. DOI: 10.2307/1989894.

[50] I. J. Schoenberg. “On Certain Metric Spaces Arising From Euclidean Spaces by a Change of Metric and Their Imbedding in Hilbert Space”. In: *Annals of Mathematics* 38.4 (1937), pp. 787–793. DOI: 10.2307/1968835.

[51] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. “Kernel principal component analysis”. en. In: *Artificial Neural Networks — ICANN’97*. Ed. by Wulfram Gerstner et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 1997, pp. 583–588. DOI: 10.1007/BFb0020217.

[52] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. en. June 2018. DOI: 10.7551/mitpress/4175.001.0001.

[53] Jaeyoung Shin et al. “Open Access Dataset for EEG+NIRS Single-Trial Classification”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10 (Oct. 2017), pp. 1735–1745. DOI: 10.1109/TNSRE.2016.2628057.

[54] Lene Theil Skovgaard. “A Riemannian Geometry of the Multivariate Normal Model”. In: *Scandinavian Journal of Statistics* 11.4 (1984), pp. 211–223.

[55] S. Sommer, F. Lauze, and M. Nielsen. “Optimization over geodesics for exact principal geodesic analysis”. en. In: *Advances in Computational Mathematics* 40.2 (Apr. 2014), pp. 283–313. DOI: 10.1007/s10444-013-9308-1.

[56] Stefan Sommer, Tom Fletcher, and Xavier Pennec. “Introduction to differential and Riemannian geometry”. en. In: *Riemannian Geometric Statistics in Medical Image Analysis*. Elsevier, 2020, pp. 3–37. DOI: 10.1016/B978-0-12-814725-2.00008-X.

[57] Stefan Sommer, Francois Lauze, and Mads Nielsen. “The Differential of the Exponential Map, Jacobi Fields and Exact Principal Geodesic Analysis”. In: *Computing Research Repository - CORR* (Jan. 2010).

[58] Suvrit Sra. “A new metric on the manifold of kernel matrices with application to matrix geometric means”. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012.

[59] Yann Thanwerdas and Xavier Pennec. “Bures–Wasserstein Minimizing Geodesics between Covariance Matrices of Different Ranks”. en. In: *SIAM Journal on Matrix Analysis and Applications* 44.3 (Sept. 2023), pp. 1447–1476. DOI: 10.1137/22M149168X.

[60] Yann Thanwerdas and Xavier Pennec. “O(n)-invariant Riemannian metrics on SPD matrices”. en. In: *Linear Algebra and its Applications* 661 (Mar. 2023), pp. 163–201. DOI: 10.1016/j.laa.2022.12.009.

[61] Yann Thanwerdas and Xavier Pennec. “The geometry of mixed-Euclidean metrics on symmetric positive definite matrices”. In: *Differential Geometry and its Applications* 81 (Apr. 2022), p. 101867. DOI: 10.1016/j.difgeo.2022.101867.

[62] Oncel Tuzel, Fatih Porikli, and Peter Meer. “Pedestrian Detection via Classification on Riemannian Manifolds”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.10 (Oct. 2008), pp. 1713–1727. DOI: 10.1109/TPAMI.2008.75.

[63] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. New York, NY: Springer, 2000. DOI: 10.1007/978-1-4757-3264-1.

[64] Frank Wilcoxon. “Individual Comparisons by Ranking Methods”. In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83. DOI: 10.2307/3001968.

[65] W. A. Wilson. “A Relation Between Metric and Euclidean Spaces”. In: *American Journal of Mathematics* 54.3 (1932), pp. 505–517. DOI: 10.2307/2370894.

[66] Or Yair, Mirela Ben-Chen, and Ronen Talmon. “Parallel Transport on the Cone Manifold of SPD Matrices for Domain Adaptation”. en. In: *IEEE Transactions on Signal Processing* 67.7 (Apr. 2019), pp. 1797–1811. DOI: 10.1109/TSP.2019.2894801.

[67] Florian Yger. “A review of kernels on covariance matrices for BCI applications”. en. In: *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. Southampton, United Kingdom: IEEE, Sept. 2013, pp. 1–6. DOI: 10.1109/MLSP.2013.6661972.

[68] Paolo Zanini et al. “Transfer Learning: A Riemannian Geometry Framework With Applications to Brain–Computer Interfaces”. In: *IEEE Transactions on Biomedical Engineering* 65.5 (May 2018), pp. 1107–1116. DOI: 10.1109/TBME.2017.2742541.

List of Figures

List of Tables

8 Appendix

8.1 On Metrics, Inner Products and Matrices

This section contains the description of all mathematical concepts used throughout the thesis.

Definition 8.1.1. *Let (M_1, d_1) and (M_2, d_2) be metric spaces. Let $\phi : M_1 \to M_2$ be a function such that*

$$d_2(\phi(A), \phi(B)) = d_1(A, B).$$

Then ϕ is called an isometry. If ϕ is a linear function, it is called a linear isometry.

Definition 8.1.2. *A Hilbert space $(\mathfrak{H}, < ., . >)$ is a real or complex inner product space that is also a complete metric space with respect to the distance function induced by the inner product.*

Remark 8.1.3. *An inner product $< ., . >$ induces a norm $||A|| = \sqrt{< A, A >}$ which in turn induces a metric $d(A, B) = ||A - B|| = \sqrt{< A - B, A - B >}$*

Theorem 8.1.4. *All Hilbert spaces of dimension $n \in \mathbf{N}$ are isometric.*

Definition 8.1.5. *We call the inner product space $(\mathbb{R}^n, < ., . >)$ the standard Euclidean space of dimension n. $< a, b >= a^T b$ is referred to as the standard Euclidean inner product.*

Remark 8.1.6. *By theorem 8.1.4 all n-dimensional Hilbert spaces are isometric to $(\mathbb{R}^n, < ., . >)$.*

Definition 8.1.7. *[56, Sec. 1.7.3] An orthogonal transformation is a linear transformation $T : V \to V$ on a real inner product space V, that preserves the inner product. For each pair u, v of elements of V, we have*

$$< u, v >=< T(u), T(v) > \ .$$

Remark 8.1.8. *Orthogonal transformations depend on the inner product.*

The following proposition is found in many textbooks:

Proposition 8.1.9. *Linear isometries between two vector spaces are unique up to orthogonal transformation.*

This means that if we have two linear isometries ϕ_1 and ϕ_2 mapping from one vector space to another, we can find an orthogonal transformation T such that $\phi_1 = T(\phi_2)$

Metric spaces allow for the definition of variance

Definition 8.1.10. *Let (M, d) be a metric space, and let $x_1, x_2, \ldots, x_n$ be a finite set of points in M. For any point $p \in M$, the Frechét variance is given as*

$$\sum_i^n d(p, x_i)^2. \tag{8.1.11}$$

Whenever we refer to variance in this thesis, we refer to the Frechét variance. The notion of variance allows for a definition of the mean of a dataset.

Definition 8.1.12. *Let (M, d) be a metric space, and let $x_1, x_2, \ldots, x_n$ be a finite set of points in M. Points $\bar{x} \in M$ that minimize the Frechét variance to the given points*

$$\bar{x} = \arg\min_{x \in M} \sum_{i=1}^{n} d(x, x_i)^2 \tag{8.1.13}$$

are called the Karcher means. If the Karcher mean is unique, it is called the Frechét mean.

Definition 8.1.14. *Let $A \in \mathbb{R}^{n \times n}$. Then the operator that stacks all elements of A in a column vector is defined as*

$$\operatorname{vect}(A) = \begin{bmatrix} a_{11} & a_{21} & \ldots & a_{n1} & a_{12} & \ldots & a_{n2} & \ldots & a_{1n} & \ldots & a_{nn} \end{bmatrix}^{\top} \in \mathbb{R}^{n^2}$$

Definition 8.1.15. *Let $A, B \in \mathbb{R}^{n \times m}$ be two matrices. The operator $\odot$ defined by*

$$[A \odot B]_{ij} = a_{ij} b_{ij}$$

is called the Hadamard (or Schur) product of two matrices of the same shape.

Definition 8.1.16. *Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. The linear function*

$$\operatorname{tr}(A) = \sum_{i}^{n} A_{ii}$$

is called the trace of A.

Definition 8.1.17. *Let $A, B \in \mathbb{R}^{n \times m}$ be two matrices. The inner product defined by*

$$< A, B >^F = \operatorname{tr}(AB^T)$$

is called the Frobenius inner product.

Remark 8.1.18. *The Frobenius inner product is equivalent to the standard Euclidean inner product for vectorized matrices: With $[AB^T]_{ij} = \sum_{i,k} A_{ij} B_{kj}$, we have*

$$\operatorname{tr}(AB^T) = \sum_{i} (AB^T)_{ii} \tag{8.1.19}$$

$$= \sum_{i,j} A_{ij} B_{ij} \tag{8.1.20}$$

$$\tag{8.1.21}$$

which is the same as the standard Euclidean inner product for vectorized matrices

$$< \operatorname{vect}(A), \operatorname{vect}(B) >_{=} \operatorname{vect}(B)^T \operatorname{vect}(A) = \sum_{k} \operatorname{vect}(B)_k \operatorname{vect}(A)_k = \sum_{i,j} A_{ij} B_{ij}.$$

The norm induced by the Frobenius inner product

$$||A||_F = \sqrt{\operatorname{tr}(AA^T)}$$

is called the Frobenius norm. The distance induced by the Frobenius norm

$$d_F(A, B) = ||A - B||_F$$

is called the Frobenius distance.

Definition 8.1.22. *Let $A \in \mathbb{R}^{n\times n}$. If $A = A^T$ it is called symmetric. We write*

$$\mathcal{S}^n = \{A \in \mathbb{R}^{n\times n} | A = A^T\}$$

Definition 8.1.23. *For symmetric matrices $A \in \mathbb{R}^{n\times n}$, the* vech *operator is defined as:*

$$\operatorname{vech}(A) = \begin{bmatrix} a_{11} \\ \sqrt{2}\, a_{12} \\ \sqrt{2}\, a_{13} \\ \vdots \\ \sqrt{2}\, a_{1n} \\ a_{22} \\ \sqrt{2}\, a_{23} \\ \vdots \\ a_{(n-1)(n-1)} \\ a_{(n-1)n} \\ a_{nn} \end{bmatrix} \in \mathbb{R}^{\frac{n(n+1)}{2}}$$

The vech operator extracts upper triangular elements and applies $\sqrt{2}$ scaling to off-diagonal elements.

Remark 8.1.24. *The Hilbert space of symmetric matrices under the Frobenius inner product given by $(\mathcal{S}^n, <, >^F)$ is isometric to vectorized matrices and thus vectors under the standard Euclidean distance $(\mathbb{R}^{\frac{n(n+1)}{2}}, <, >)$ where the isometry is the function* vech *in definition 8.1.23.*

Definition 8.1.25. *A real square matrix U is called an orthogonal matrix if $UU^T = U^TU = I$.*

Remark 8.1.26. *Orthogonal transformations correspond to orthogonal matrices U.*

Definition 8.1.27. *Let $< ., . >_F$ be the Frobenius inner product.*
For a symmetric matrix S, we define the symmetric orthogonal Frobenius transformation T^{sym} as $T^{sym}(S) = USU^T$ where U is an orthogonal matrix.

Proposition 8.1.28. *The set of symmetric orthogonal Frobenius transformations T^{sym} on $\mathcal{S}^n$ is a subset of the set of orthogonal transformations on $\mathbb{R}^{\frac{n(n+1)}{2}}$.*

Proof. Let $w_1 = \operatorname{vech}(S)$ be a vector of length $\frac{n(n+1)}{2}$ from a symmetric matrix. Let $U \in \mathbb{R}^{n\times n}$ be an orthogonal matrix. Then $w_2 = \operatorname{vech}(USU^T)$ is also a vector of length $\frac{n(n+1)}{2}$ from a symmetric matrix. As $||S||_F = ||USU^T||_F$, we know by construction of vech that $||\operatorname{vech}(USU^T)|| = ||\operatorname{vech}(S)||$. Thus, w_1 and w_2 have the same norm. This mean there exists an orthogonal $V \in \mathbb{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}$ such that $Vw_1 = w_2$. V can be constructed as the Householder matrix of the vector $u := w_1 - w_2$. □

However, the opposite is not true: there does not necessarily exist a U such that for every V we can write $V \operatorname{vech}(S) = \operatorname{vech}(USU^T)$.

Corollary 8.1.29. *Let $S \in \mathcal{S}^n$ be a symmetric matrix and $V \in \mathbb{R}^{\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}}$ orthogonal. Then $\operatorname{vech}^{-1}(V \operatorname{vech}(S))$ is a symmetric matrix and $|| \operatorname{vech}^{-1}(V \operatorname{vech}(S))||_F = ||S||_F$*

Theorem 8.1.30. *Let $S \in \mathcal{S}^n$. Then S can be decomposed as*

$$S = U \Lambda_S U^T$$

where U is an orthogonal matrix where the columns are the eigenvectors of S and Λ_S is a diagonal matrix with the eigenvalues of S on the diagonal. This decomposition is called the eigendecomposition of S.

Definition 8.1.31. *Let $A \in \mathcal{S}^n$. Then A is called positive (negative) semi-definite if for all $x \in \mathbb{R}^n \backslash \{0\}$ we have*

$$x^T A x \geq 0 \ (\leq 0)$$

If the inequality is strict, the matrix is called positive (negative) definite. We denote by $\mathcal{P}^n$ the set of symmetric positive definite matrices.

Definition 8.1.32. *Let $S \in \mathcal{S}^n$ with eigendecomposition $S = U \Lambda_S U^T$ and f a scalar function defined on the eigenvalues of S. Then a function F acting on S*

$$F(S) : \mathcal{S}^n \to \mathcal{S}^n$$

is called a matrix function corresponding to the scalar function f with

$$F(S) : S \to F(S) := U f(\Lambda_S) U^T$$

where f is acting on the individual entries of Λ_S.

Notable examples of matrix functions are the logarithm, the exponential function and power functions.

Definition 8.1.33. *Let $S = U \Lambda U^T$ be a symmetric matrix and F a matrix function with corresponding differentiable scalar function f defined on the eigenvalues of S. Then the first divided difference $F^{[1]} \in \mathcal{S}^n$ is defined by*

$$[F^{[1]}(\Lambda_S)]_{ij} = \begin{cases} \frac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j}, & \lambda_i \neq \lambda_j \\ f'(\lambda_i), & \lambda_i = \lambda_j \end{cases} \tag{8.1.34}$$

where $\lambda_i, i = 1, ..., n$ are the entries of Λ_S.

Definition 8.1.35. *Let $S \in \mathcal{S}^n$ and F be a matrix function defined on S. Then the directional derivative at S in the direction of a matrix V is given as*

$$\partial_V F(S) = \lim_{h \to 0} \frac{F(S + hV) - F(S)}{h}. \tag{8.1.36}$$

The directional derivative for matrix functions is often referred to as the Gateaux derivative.

Theorem 8.1.37. *The Gateaux derivative of a matrix function $F(S)$ in the direction of a matrix V can be computed according to the Daleckii-Krein theorem as:*

$$\partial_V F(S) = U[F^{[1]}(\Lambda_S) \odot (U^T V U)]U^T \tag{8.1.38}$$

where $F^{[1]}$ is the first divided difference.

Definition 8.1.39. *Let $F : \mathcal{S}^n \to \mathcal{S}^n$ be a matrix function with differentiable corresponding scalar function f. Then the differential at point S, $[\mathrm{D}_S F]$ is defined by*

$$\begin{aligned} [\mathrm{D}_S F] : \mathbb{R}^{n\times n} &\to \mathbb{R}^{n\times n} \\ V &\mapsto \frac{d}{dt} F(S + tV)|_{t=0} \end{aligned} \tag{8.1.40}$$

Remark 8.1.41. *The Gateaux derivative and the differential of a function F are connected by*

$$\partial_V F(S) = [\mathrm{D}_S F](V). \tag{8.1.42}$$

8.2 (Kernel) Principal Component Analysis

8.2.1 Principal Component Analysis

Principal Component Analysis (PCA, [29, 45]) aims at finding meaningful directions, called principal components, of a given dataset $\{x_i\}_{i=1,\dots,p}$, $x_i \in \mathbb{R}^n$. These directions are denoted by $(v_i)_{i=1,\dots,n}$. The directions need to fulfill the requirements $||v_i|| = 1$ for all i and $< v_i, v_j >= 0$ for all $i \neq j$. Thus, the goal of PCA is to construct an orthonormal basis of $\mathbb{R}^n$. Different optimization problems can be stated to obtain a set of directions that attribute different meanings to the directions. We state three different optimization problems attributing a different meaning of these principal components of PCA and show that they all coincide and produce the same optimization result.

The first optimization objective is based on finding orthogonal directions, or one-dimensional subspaces, that maximize the variance in the sense that the sum of the variances of all data points projected onto these subspaces is maximized. The second procedure minimizes projection error. It finds the direction such that the distance of orthogonally projected points onto this direction is minimized. The third optimization is based on residual variances. The goal is to find directions that, if they are eliminated from all data points, leave the least residual variance on the data.

All optimization formulations are based on the notion of Frechét variance, which can be stated in general for metric spaces as

$$var(\{x_i\}_{i=1,\dots,p}) = \sum_i^p d^2(\mu, x_i) \tag{8.2.1}$$

where μ is the Frechét mean of the dataset.

If we want to consider the variance along a given direction v, we are interested in the variance after each data point is projected onto the given direction. The variance along a direction is given as

$$var(\{x_i\}_{i=1,\dots,p})_v = \sum_i^p d^2(\mu, \pi_v(x_i)). \tag{8.2.2}$$

where $\pi_v(x)$ is a projection operator that orthogonally projects a data point x onto v.

Now we need an operator that orthogonally projects any given data point $x \in \mathbb{R}^n$ onto the one-dimensional subspace parametrized by v. The goal is to find a $t^* \in \mathbb{R}$ by

$$t^* = \arg\min_t ||tv - x||^2 = \arg\min_t < tv - x, tv - x >$$

Solving this optimization problem results in the expression

$$t^* = \frac{< x, v >}{< v, v >} =< x, v > .$$

We thus define the projection operator π.

Definition 8.2.3.

$$\pi_v(x) =< x, v > v.$$

is the projection operator that orthogonally projects x onto v.

Finding directions can be done iteratively. After finding a direction v_i according to some optimization objective, we subtract the variance in the direction v_i by a function ψ and find the next direction v_{i+1} following the same objective on the data $(\psi(x_i))_{i=1,\dots,p}$.

Removing variance in a given direction v means finding an operator ψ_v such that

$$var(\psi_v(x_i))_v = d^2(\mu, \pi_v(\psi_v(x_i))) = 0. \tag{8.2.4}$$

Inserting the projection operator π_v from definition 8.2.3 and assuming zero mean we get

$$d^2(\mu, \pi_v(\psi_v(x_i))) =<< \psi_v(x_i), v > v, < \psi_v(x_i), v > v >= 0$$

which means

$$< \psi_v(x_i), v >= 0.$$

The solution to this is subtracting the projection of x onto w from x, i.e.,

$$\psi_v(x) = x - < x, v > v \tag{8.2.5}$$

which can be shown by

$$\begin{aligned}
< \psi_v(x), v > &=< x - < x, v > v, v > && (8.2.6)\\
&=< v, x > - << x, v > v, v > && (8.2.7)\\
&=< v, x > - < v, x >< v, v > && (8.2.8)\\
&=< v, x > - < v, x > && (8.2.9)\\
&= 0. && (8.2.10)
\end{aligned}$$

Assuming the same iterative procedure with variance subtraction in each step for all optimization objectives, we limit the following analysis to finding the first principal component.

Maximization of Variance. Given a dataset $\{x_i\}_{i=1,\dots,p}$, we want to find a vector v_1 that fulfills

$$v_1 = \arg\max_{||v||=1} \sum_{i}^{p} d^2(\mu, \pi_v(x_i)) \tag{8.2.11}$$

Assuming centered data, i.e., $\mu = 0$, and applying definition 8.2.3 this reduces to

$$v_1 = \arg\max_{||v||=1} \sum_{i}^{p} (t^*)^2 = \arg\max_{||v||=1} \sum_{i}^{p} < x_i, v >^2 \tag{8.2.12}$$

Minimization of Projection Error. The second optimization formulation for PCA is framed in terms of reconstruction quality. For this, the linear approximation of the dataset that results in the lowest reconstruction error is found. Given a direction x and a projection operator π_w, the problem can be formulated as

$$v_1 = \arg\min_{v} \sum_{i}^{p} d^2(x_i, \pi_v(x_i)) = \arg\min_{v} \sum_{i}^{p} < x_i - \pi_v(x_i), x_i - \pi_v(x_i) > \tag{8.2.13}$$

Given the linear projection operator $\pi_v(x) =< x, v > v$, each term can be written as

$$< x_i - < x_i, v > v, x_i - < x_i, v > v > \tag{8.2.14}$$
$$= < x_i, x_i > -2 < x_i, < x_i, v > v > + << x_i, v > v, < x_i, v > v > \tag{8.2.15}$$
$$= < x_i, x_i > -2 < x_i, v >< x_i, v > + < x_i, v >^2 < v, v > \tag{8.2.16}$$
$$= < x_i, x_i > - < x_i, v >^2 \tag{8.2.17}$$

this results in the optimization

$$v_1 = \arg\min_{v} - \sum_{i}^{p} < x_i, v >^2$$

as $< x_i, x_i >$ is independent of v.

Minimization of Total Residual Variance. Using the operator ψ we can formulate a third optimization problem for PCA, which directly optimizes for total residual variance after applying ψ. Assuming $\mu = 0$, the minimizing total residual variance problem can be stated as

$$v_1 = \arg\min_{v} \sum_{i}^{p} d^2(0, \psi_v(x_i))$$

which results in

$$\arg\min_v \sum_i^p < x_i - < x_i, v > v, x_i - < x_i, v > v > \tag{8.2.18}$$

$$= \arg\min_v \sum_i^p < x_i, x_i > -2 << x_i, v > v, x_i > + << x_i, v > v, < x_i, v > v > \tag{8.2.19}$$

$$= \arg\min_v \sum_i^p -2 < x_i, v >^2 + < x_i, v >^2 \tag{8.2.20}$$

$$= \arg\min_v - \sum_i^p < x_i, v >^2 \tag{8.2.21}$$

$$= \arg\max_v \sum_i^p < x_i, v >^2 . \tag{8.2.22}$$

Analytic solution to the optimization. As

$$v_1 = \arg\min_{||v||=1} \sum_i^p - < x_i, v >^2 = \arg\max_{||v||=1} \sum_i < x_i, v >^2$$

the optimization problems are equivalent.

The resulting vector v_1 is the first principal component and is the direction of highest variance that is in the data, as well as the direction that allows for the lowest reconstruction error and least residual variance after subtraction from the data. Given the first component, we extract the second component by removing the variance in direction v_1 and performing the same optimization again

$$v_2 = \arg\max_{||v||=1} \sum_i < x_i', v >^2 \tag{8.2.23}$$

where $x' = x - < x_i, v_1 >^2 v_1$. The remaining principal components are extracted in the same way.

To make the optimization problem easier to address, we can rewrite

$$\sum_i < x_i, v >^2 = \sum_i (v^T x_i)(x_i^T v) = \sum_i v^T (x_i x_i^T) v = v^T (\sum_i (x_i x_i^T)) v = v^T C v$$

where $C = \sum_i (x_i x_i^T)$.

The constrained optimization problem can be solved using Langrange multipliers. Maximize

$$L(v) = v^T C v - \lambda (v^T v - 1)$$

where λ is a Lagrange multiplier. We have

$$\frac{\partial}{\partial v} L = Cv - \lambda v.$$

Setting $\frac{\partial}{\partial v} L = 0$ yields

$$Cv = \lambda v$$

which is the eigenvalue equation. The v that results in the highest λ is exactly the eigenvector that corresponds to the largest eigenvalue of C.

The resulting $(v_i)_{i=1,\ldots,n}$ of the optimization problem are simply the eigenvectors corresponding to the eigenvalues of C sorted in descending order and the eigenvalues correspond to the variance in their associated direction.

8.2.2 Kernel Principal Component Analysis

Kernel Principal Component Analysis (kernel PCA, [51]) is the non-linear extension of PCA. Let $\{x_i\}_{i=1,\dots,p} \in \mathbb{R}^n$ be a dataset. Instead of solving PCA in $\mathbb{R}^n$, the idea is to map the data to some (unknown) Hilbert space $\mathfrak{H}$ with a non-linear map $\phi : \mathbb{R}^n \to \mathfrak{H}$ to be able to capture non-linear structures in the data. This leverages positive definite kernels as is shown in following derivation, found in [51].

Let k be a positive definite kernel. By Mercer's theorem, there exists some Hilbert space $\mathfrak{H}$ and a (non-linear, possibly unknown) map $\phi : \mathbb{R}^n \to \mathfrak{H}$ such that $k(x, y) =< \phi(x), \phi(y) >$. We will for now assume that the data is centered in this feature space, i.e., $\sum_i \phi(x_i) = 0$. To solve the problem of PCA in $\mathfrak{H}$, we would need to calculate the matrix

$$\tilde{C} = \sum \phi(x_j)\phi(x_j)^T$$

which is not possible in general as the function ϕ is not always known. The next step would be to find eigenvalues and eigenvectors of $\tilde{C}$ by solving

$$\lambda v = \tilde{C}v.$$

As v lies in the span of $\phi(x_1), \dots, \phi(x_M)$, we can instead consider the equivalent system of k equations

$$\lambda(\phi(x_k)v) = \tilde{C}(\phi(x_k)v)$$

for all k. Additionally, we can infer that there exist coefficients α_i such that

$$v = \sum \alpha_i \phi(x_i)$$

Combining this results in

$$\lambda \sum \alpha_i \phi(x_i)\phi(x_k)^T = \sum_i \alpha_i (\phi(x_i) \sum_j \phi(x_j)^T)\phi(x_i)\phi(x_j)^T$$

Defining $K_{ij} = \phi(x_i)^T\phi(x_j)$, above equation translates to

$$m\lambda K\alpha = K^2\alpha$$

which is equivalent to solving

$$m\lambda\alpha = K\alpha,$$

so solving for non-zero eigenvalues of K gives the solution for kernel PCA. As we for now assumed centered data in feature space, which may be violated as ϕ is commonly unknown, there is one additional step that needs to be performed in kernel PCA. Without knowing the feature space itself, we can center the kernel matrix $K_{ij} = k(x_i, x_j)$ by calculating the new kernel matrix K' which we use for further processing

$$K' = K - \mathbf{1}_n K - K\mathbf{1}_n + \mathbf{1}_n K \mathbf{1}_n.$$

Thus, all we need for a non-linear expansion of PCA is a kernel matrix $K_{ij} = k(x_i, x_j)$ and then to compute its non-zero eigenvalues and the corresponding eigenvectors. To show that kernel PCA is a generalization of PCA, one can show that kernel PCA with the linear kernel $k(x_i, x_j) =< x_i, x_j >$ is equivalent to PCA.

8.3 Support Vector Machines

Support vector machines are the resulting classifier derived from the theoretical framework of statistical learning theory [63]. The goal of these learning algorithms is to construct a classifier that controls the generalization ability for small sample sizes.

The C-support-vector-machine (SVC, [52]) for data points $X_i, i = 1, ..., n$ with labels $y_i, i = 1, ..., n$ can be extended from it's standard linear form

$$\begin{gathered}\max_{\alpha} L_{dual} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left(\mathbf{x_i^T x_j} + \frac{1}{2C} \delta_{ij} \right) \\ \text{s.t } : \alpha_i \geq 0, \forall i \in D, \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0\end{gathered} \tag{8.3.1}$$

to solve non linear problems by mapping the data points x_i to another space in which the problem is solvable linearly using a function ϕ. As the space in which the data points are linearly separable is usually unknown, the mapping is performed implicitly by replacing the scalar product $x_i^T x_j$ in the optimization by a kernel $k(x_i, x_j)$. Per Mercer's kernel theorem, every positive definite kernel corresponds to a scalar product in some (possibly unknown and infinitely dimensional) space.
Thus, we can construct a support vector machine acting on the space of positive definite matrices by replacing the scalar product in the optimization by a Riemannian kernel.

8.4 Data and Features

8.4.1 Datasets

	Subjects	Channels	Sampling Frequency	Classes	Sessions	Trials per Class	Trial Length
BNCI2014-004 [39]	9	3	250Hz	2	5	360	4.5s
BNCI2015-001 [19]	12	13	512Hz	2	3	200	5s
Lee2019-MI [38]	54	62	1000Hz	2	2	100	4s
Shin2017A [53]	29	30	200Hz	2	3	30	10s

Table 8.1: Dataset List.

All datasets were retrieved using the Mother-of-all-BCI-Benchmarks (MOABB, [31]) library. We selected the datasets to be all available datasets with a motor imagery task with two classes and multiple sessions that did not produce an error while downloading using MOABB. The MOABB library provides standard parameters for data preprocessing which we used without alteration. The data was recorded by EEG electrodes and is band-pass filtered from $7 - 35$ Hz. The windows for feature extraction begin at the task cue and

end after the trial length specified in each dataset respectively. The features provided by MOABB to be used in the following pipelines are trial-based windows of the multivariate EEG data with dimensionality $N_{trials} \times N_{channels} \times N_{samples}$ where $N_{samples}$ = trial length $\times$ sampling frequency.

BNCI2014-004 [39] consists of five sessions for nine subjects with two classes: left hand, and right hand. BNCI2015-001 [19] consists of three sessions for twelve subjects with two classes: right hand, and feet.

Lee2019-MI [38] consists of two sessions for 55 subjects with two classes, left hand and right hand. Shin2017A [53] consists of two sessions for 29 subjects with two classes: left and right hand.

8.4.2 Covariance as Riemannian Features

Any feature that can be represented as a positive definite matrix can be used for data analysis with the proposed Riemannian methods.

One of the most common examples are the covariances from multivariate data. Given n random variables $X_i, i = 1, ..., n$, let $X = (X_1, ..., X_n)^T$, we can calculate the covariance matrix

$$cov(X) = \mathbb{E}[(X - \mu_X)(X - \mu_X)^T] = \mathbb{E}[XX^T] - \mu_X \mu_X^T$$

where $\mathbb{E}[Y]$ is the expected value of random variable Y and $\mu_Y := \mathbb{E}[Y]$.

The covariance matrix $cov(X)$ is guaranteed to be positive definite for linearly independent random variables $X_i, i = 1, ..., n$, thus covariance matrices of independent random variables lie on the Riemannian manifold of positive definite matrices. If any (sub-)set of the random variables is linearly dependent, the covariance reduces to a positive semi-definite matrix. The set of positive semi-definite matrices, however, does not form a Riemannian manifold.

To show that for every positive definite matrix C, there is a n-dimensional random variable $X = (X_1, ..., X_n)^T$ that produces $C \in \mathbb{P}_n$ as a covariance matrix, let us consider the diagonalization of C as

$$C = QDQ^T$$

where Q is a n-dimensional real orthogonal matrix and D is a n-dimensional positive diagonal matrix. With Z as the standard n-dimensional multivariate normal distribution, i.e., $Z \sim \mathcal{N}(0, I_n)$, we can define

$$X = QD^{1/2}Z$$

for which we can show that

$$\begin{aligned} cov(X) &= \mathbb{E}[(X - \mu_X)(X - \mu_X)^T] \\ &= \mathbb{E}[XX^T] \\ &= \mathbb{E}[QD^{1/2}Z(QD^{1/2}Z)^T] \\ &= \mathbb{E}[QD^{1/2}ZZ^T(D^{1/2})^TQ^T] \\ &= \mathbb{E}[QDQ^T] \\ &= QDQ^T \\ &= C, \end{aligned}$$

thus C is the covariance matrix of X and the set of positive definite covariance matrices can be identified with the set of positive definite matrices, allowing the application of Riemannian learning methods to covariance matrices.

As sample covariance matrices may not always be positive definite, they need to be regularized to be able to use Riemannian methods. Common regularization methods are the Ledoit-Wolf (LWF) and Oracle Approximation shrinkage (OAS) algorithms. Both algorithms have its specific use cases for which they can be considered optimal. As the LWF algorithm is faster and comparison of shrinkage algorithms is not a goal of this thesis, we opted for this regularization throughout the analysis.

8.4.3 Manifold, Tangent Space and Riemannian Linear Features

Assuming data on the manifold of positive definite matrices, we are confronted with the choice of feature space in which machine learning methods are to be applied within the realm of Riemannian geometry. The first option is to directly use covariances as features in kernel methods. However, only the Stein kernel (see eq. (5.3.7)) is known to be positive definite on the manifold of SPD matrices for certain parameters.

The second option is to use the Riemannian linear feature space. Canonical kernels have equivalent formulations as linear kernels in Riemannian linear feature space. In distance-based classification algorithms, directly using the manifold of SPD matrices puts us into the realm of metric space kernels. As there are no general positive definite metric space kernels for curved manifolds, there is no kernel method that can be applied directly on the curved manifold and always ensure positive definite kernels for all parameter ranges. For flat manifolds, the pairwise distances are the same as for the Riemannian linear features. We can thus describe proposed positive definite methods except the Stein kernel in Riemannian linear feature space using the standard Euclidean inner product and distance.

The third possibility that is commonly described in the literature is to use the tangent space as if it were the standard Euclidean space and apply machine learning methods there. We can describe linear kernels on the tangent space as the corresponding concept to canonical kernels being described as linear kernels in Riemannian linear feature space.

For metric space kernels in flat manifolds, simple tangent space projection distorts the distance geometry whereas Riemannian linear features exactly retain the distance geometry from the manifold. For curved metrics, the precise mathematical description is not as simple, as both Riemannian linear features and the tangent space distort the distance geometry on the manifold by giving a linear approximation of the non-linear distance geometry.

8.5 Significance Analysis

To assess significance of performance differences between methods, we selected the Wilcoxon matched-pairs signed-rank test [64]. It is a non-parametric paired difference test. We report significance given by the P value in five different steps marked in the following way: Significances are shown with directionality. *first method* $***\rightarrow$ *second method* means that *first method* significantly outperformed *second method* in the following sense:

Let x and y be two sets of measurements. If the distribution underlying $d = x - y$ is stochastically greater than a distribution around zero, the method giving rise to the measurement x outperforms the method giving

Representation	Meaning
ns	not significant, $P > 0.05$
*	$P \leq 0.05$
**	$P \leq 0.01$
***	$P \leq 0.001$
****	$P \leq 0.0001$

Table 8.2: Significance Report.

rise to the measurement y.

8.6 AUClogRNX scorer

AUClogRNX estimates how well an algorithm can preserve neighborhood structure in lower dimensions, both locally and globally. For a dataset $(x_i)_{i=1}^N$ let v_i^K denote the K-nearest-neighbors of data point x_i in the original space and n_i^K the K-nearest-neighbors of the (lower dimensional) projection of data point x_i. AUClogRNX is then defined as

$$\begin{aligned} Q_{NX}(K) &= \frac{1}{KN} \sum_{i=1}^{N} \left| v_i^K \cap n_i^K \right| \\ R_{NX}(K) &= \frac{(N-1)Q_{NX}(K) - K}{N-1-K} \\ AUC_{lnK}\left(R_{NX}(K)\right) &= \frac{\sum_{K=1}^{N-2} R_{NX}(K)/K}{\sum_{K=1}^{N-2} \frac{1}{K}}. \end{aligned} \tag{8.6.1}$$

It can be seen as a measure of retaining distance geometry between the original dataset and a (possibly lower dimensional) projection. As AUClogRNX assesses the neighborhood structure rather than pairwise distances, it is robust to feature scaling. However, a perfect AUClogRNX score does not indicate a perfect retention of distance geometry between the original space and the projection. Given a sufficient number of data points, AUClogRNX provides a good measure of retaining distance geometry.