Chapter 5
From Covariance Matrices to Covariance Operators: Data Representation from Finite to Infinite-Dimensional Settings

Hà Quang Minh and Vittorio Murino

Abstract This chapter presents some of the recent developments in the generalization of the data representation framework using finite-dimensional covariance matrices to infinite-dimensional covariance operators in Reproducing Kernel Hilbert Spaces (RKHS). We show that the proper mathematical setting for covariance operators is the infinite-dimensional Riemannian manifold of positive definite Hilbert–Schmidt operators, which are the generalization of symmetric, positive definite (SPD) matrices. We then give the closed form formulas for the affine-invariant and Log-Hilbert–Schmidt distances between RKHS covariance operators on this manifold, which generalize the affine-invariant and Log-Euclidean distances, respectively, between SPD matrices. The Log-Hilbert–Schmidt distance in particular can be used to design a two-layer kernel machine, which can be applied directly to a practical application, such as image classification. Experimental results are provided to illustrate the power of this new paradigm for data representation.

5.1 Introduction

Symmetric Positive Definite (SPD) matrices, in particular covariance matrices, play an important role in many areas of mathematics, statistics, machine learning, and their applications. In practice, the applications of SPD matrices are numerous, including brain imaging [3, 12, 34], kernel learning [19] in machine learning, object detection [39, 40] and image retrieval [11] in computer vision, and radar signal processing [5, 15].

In the field of computer vision and image processing, covariance matrices have recently been utilized as a powerful image representation approach, which is

H.Q. Minh (✉) · V. Murino
Pattern Analysis and Computer Vision (PAVIS),
Istituto Italiano di Tecnologia (IIT), Genova, 16163, Italy
e-mail: minh.haquang@iit.it

V. Murino
e-mail: vittorio.murino@iit.it

H.Q. Minh and V. Murino (eds.), *Algorithmic Advances in Riemannian Geometry and Applications*, Advances in Computer Vision and Pattern Recognition, DOI 10.1007/978-3-319-45026-1_5

commonly called *covariance descriptor*. In this approach, an image is compactly represented by a covariance matrix encoding correlations between different features extracted from that image. This representation has been demonstrated to work very well in practice and consequently, covariance descriptors have been applied with success to many computer vision tasks, including tracking [33], object detection and classification [39, 40], and image retrieval [11]. A more detailed discussion of the covariance matrix representation can be found in the chapter by Cherian and Sra in this volume.

Riemannian geometric framework for covariance matrices. Covariance matrices, properly regularized if necessary, are examples of SPD matrices. In the following, we denote by $\mathrm{Sym}^{++}(n)$ the set of all $n \times n$ SPD matrices. A key mathematical property of $\mathrm{Sym}^{++}(n)$ is that it is not a vector subspace of Euclidean space under the standard matrix addition and scalar multiplication operations. Instead, it is an open convex cone, since it is only closed under positive scalar multiplication, and at the same time admits a differentiable manifold structure. Consequently, in general, the optimal measure of similarity between covariance matrices is not the Euclidean distance, but a metric that captures the geometry of $\mathrm{Sym}^{++}(n)$. Among the most widely used metrics for $\mathrm{Sym}^{++}(n)$ is the classical *affine-invariant Riemannian metric* [6, 7, 23, 30, 31, 40], under which $\mathrm{Sym}^{++}(n)$ becomes a Riemannian manifold with nonpositive curvature. Another commonly used Riemannian metric for $\mathrm{Sym}^{++}(n)$ is the recently introduced Log-Euclidean metric [3, 4], which is *bi-invariant* and under which the manifold is flat. Compared to the affine-invariant metric, the Log-Euclidean metric is faster to compute, especially on large datasets, and can be used to define many positive definite kernels, such as the Gaussian kernel, allowing kernel methods to be applied directly on the manifold [17, 24].

Positive definite kernels and covariance operators. While they have been shown to be effective in many applications, one major limitation of covariance matrices is that they only capture *linear* correlations between input features. In order to encode *nonlinear* correlations, we generalize the covariance matrix representation framework to the infinite-dimensional setting by the use of positive definite kernels defined on the original input features. Intuitively, from the viewpoint of kernel methods in machine learning [37], each positive definite kernel, such as the Gaussian kernel, induces a feature map that nonlinearly maps each input point into a high (generally infinite) dimensional feature space. We then represent each image by an infinite-dimensional covariance operator, which can be thought as the covariance matrix of the infinite-dimensional features in the feature space. Since the high-dimensional feature maps are nonlinear, the resulting covariance operators thus encode the nonlinear correlations between the original input features. A key property of this framework, as is common for kernel methods, is that the infinite-dimensional feature maps and the corresponding covariance operators are all *implicit*, and all necessary computations are carried out via the Gram matrices associated with the given kernels.

Infinite-dimensional Riemannian manifold setting for covariance operators. Having represented each image by a covariance operator, we need to define a notion of distances between these operators. Instead of the finite-dimensional manifold setting for covariance matrices, in the infinite-dimensional setting, regularized covariance operators lie on an infinite-dimensional Hilbert manifold. This is the manifold of positive definite unitized Hilbert–Schmidt operators, which are scalar perturbations of Hilbert–Schmidt operators on a Hilbert space and which are infinite-dimensional generalizations of SPD matrices. On this manifold, the generalization of the affine-invariant Riemannian metric on $\mathrm{Sym}^{++}(n)$ was recently carried out by [1, 21, 22] from a purely mathematical viewpoint. For the case of RKHS covariance operators, the explicit formulas for the affine-invariant distance, in terms of the Gram matrices, were obtained in [26]. The generalization of the Log-Euclidean metric, called the *Log-Hilbert–Schmidt metric*, was formulated by [28], including the explicit formulas for the distances between RKHS covariance operators. As with the Log-Euclidean metric, the Log-Hilbert–Schmidt metric can be used to define many positive definite kernels, such as the Gaussian kernel, allowing kernel methods to be applied on top of the infinite-dimensional manifold and effectively creating a two-layer kernel machine.

Differences between the finite and infinite-dimensional settings. In [32], in the context of functional data analysis, the authors discussed the difficulty of generalizing the affine-invariant and Log-Euclidean metrics to the infinite-dimensional setting and proposed several other metrics instead. As we analyze in [26, 28] and below, this difficulty is due to the fundamental differences between the finite and infinite-dimensional cases. The reason is that many concepts, such as *principal matrix logarithm, determinant, and norm*, all involve infinite sums and products and therefore are well-defined only on specific classes of infinite-dimensional operators. In particular, the infinite-dimensional distance formulas are *not* the limits of the finite-dimensional ones as the dimension approaches infinity.

The aim of this chapter. In the present chapter, we first show how to generalize the data representation framework by covariance matrices to RKHS covariance operators. We then report on the recent development in mathematical theory of infinite-dimensional positive definite operators [1, 21, 22, 28], which successfully resolves the problems of extending the affine-invariant and Log-Euclidean metrics to the infinite-dimensional setting. We then show how this theory can be applied to compute distances between RKHS covariance operators. In particular, we describe in detail the two-layer kernel machine which arises from the Log-Hilbert–Schmidt distance between RKHS operators, which can be used in a practical application such as image classification.

Related work. In the literature on kernel methods in machine learning, it is well-known that RKHS covariance operators defined on nonlinear features, which are obtained by mapping the original input data into a high-dimensional feature space, can better capture input correlations than covariance matrices defined on the original input data, see e.g. KernelPCA [36]. However, the use of RKHS covariance operators for data representation is quite recent and has its origin in computer vision [14, 16, 41]. The main problem with the approaches in [14, 16, 41] is that they lack the theoretical foundation provided by the mathematical theory of infinite-dimensional operators and infinite-dimensional manifolds. As such, they are necessarily heuristic and many results obtained are only valid in the finite-dimensional setting.

Organization. We start by recalling the data representation framework using covariance matrices in Sect. 5.2. Then in Sect. 5.3, we show how this framework generalizes to RKHS covariance operators which are induced by positive definite kernels and their associated feature maps. In Sect. 5.4, we give the closed form formulas for the Hilbert–Schmidt, affine-invariant, and Log-Hilbert–Schmidt distances between RKHS covariance operators. The two-layer kernel machine resulting from the Log-Hilbert–Schmidt distance is described in Sect. 5.5, with experimental results illustrating its power in Sect. 5.6. Mathematical proofs are given in the Appendix.

5.2 Covariance Matrices for Data Representation

Before presenting covariance operators for data representation, we first recall how covariance matrices are employed as a form of image representation. For each image, at every pixel (or a subset of the pixels), we extract an image feature vector consisting of n features, for example intensity, gradient, and colors. Suppose that we perform feature extraction at m pixels, each one giving a feature vector $x_i \in \mathbb{R}^n$, $i = 1, \ldots, m$, we then obtain a data matrix of size $n \times m$, given by

$$\mathbf{x} = [x_1, \ldots, x_m], \tag{5.1}$$

with each column consisting of image features extracted at one pixel. The $n \times n$ covariance matrix

$$C_{\mathbf{x}} = \frac{1}{m}\mathbf{x}J_m\mathbf{x}^T = \frac{1}{m}\sum_{j=1}^{m}(x_j - \mu)(x_j - \mu)^T, \tag{5.2}$$

then encodes *linear correlations* between all the different extracted features and is used as the representation for the image. Here J_m is the centering matrix, defined by $J_m = I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T$, where $\mathbf{1}_m = (1, \ldots, 1)^T \in \mathbb{R}^m$ and $\mu = \frac{1}{m}\sum_{j=1}^m x_j \in \mathbb{R}^n$ denotes the mean column of $\mathbf{x}$. In general, $C_{\mathbf{x}}$ is a symmetric, positive *semi-definite* matrix.

In a practical application, such as classification, we need to have a similarity measure between images. By representing images as covariance matrices, this means that we need to compute distances between covariance matrices. Let A and B be two symmetric, positive semi-definite matrices. A straightforward distance between A and B is the Euclidean distance, given by

$$d_E(A, B) = ||A - B||_F. \tag{5.3}$$

Here $||\ ||_F$ denotes the Frobenius norm, which, for $A = (a_{ij})_{i,j=1}^n$, is defined by

$$||A||_F^2 = \mathrm{tr}(A^TA) = \sum_{i,j=1}^{n} a_{ij}^2. \tag{5.4}$$

It is clear from the definition of the Frobenius norm that the distance $||A - B||_F$ depends only on the entries of $A - B$, without taking into account any structure of A and B. Furthermore, the set of symmetric, positive semi-definite matrices is not a vector subspace of Euclidean space under the standard matrix addition and scalar multiplication operations, but a convex cone, since it is only closed under *positive* scalar multiplication. By simply vectorizing A and B, the Euclidean distance $||A - B||_F$ reflects neither the positivity of A and B nor the convex cone structure of the set of positive matrices.

We note that, *empirically*, by a simple regularization, the regularized covariance matrix $(C_{\mathbf{x}} + \gamma I)$ for any constant $\gamma > 0$ is an element of $\mathrm{Sym}^{++}(n)$, which has been studied extensively, both mathematically and computationally. Thus, we can apply to the set of regularized covariance matrices any distance on $\mathrm{Sym}^{++}(n)$ that reflects its intrinsic geometry as a set of SPD matrices. The regularization $(C_{\mathbf{x}} + \gamma I)$ is called *diagonal loading* in the literature (see [2, 13] for more general forms of regularizations). We show in Sect. 5.4 below that for infinite-dimensional covariance operators, this form of regularization is always necessary, both *theoretically* and *empirically*.

Let $\mathrm{Sym}^{++}(n)$ denote the set of SPD matrices of size $n \times n$. Let $A, B \in \mathrm{Sym}^{++}(n)$ be arbitrary. We now review three distances that exploit the geometry of $\mathrm{Sym}^{++}(n)$, namely the affine-invariant distance, Log-Euclidean distance, and Bregman divergences.

Affine-invariant metric. In the first approach, the set $\mathrm{Sym}^{++}(n)$ is equipped with a Riemannian metric, the so-called *affine-invariant metric* [6, 7, 23, 30, 31]. For each $P \in \mathrm{Sym}^{++}(n)$, the tangent space at P is $T_P(\mathrm{Sym}^{++}(n)) \cong \mathrm{Sym}(n)$, the space of symmetric matrices of size $n \times n$. The affine-invariant metric is defined by the following inner product on the tangent space at P

$$\langle A, B\rangle_P = \langle P^{-1/2}AP^{-1/2}, P^{-1/2}BP^{-1/2}\rangle_F, \quad \forall P \in \mathrm{Sym}^{++}(n), A, B \in \mathrm{Sym}(n). \tag{5.5}$$

Under the affine-invariant metric, $\mathrm{Sym}^{++}(n)$ becomes a Riemannian manifold with *nonpositive sectional curvature*. The affine-invariant geodesic distance between A and B is given by

$$d_{\mathrm{aiE}}(A, B) = ||\log(A^{-1/2}BA^{-1/2})||_F, \tag{5.6}$$

where log denotes the principal matrix logarithm.

Log-Euclidean metric. In the second approach, the set $\mathrm{Sym}^{++}(n)$ is equipped with a *bi-invariant* Riemannian metric, the so-called Log-Euclidean metric [4]. This metric arises from the following commutative Lie group multiplication on $\mathrm{Sym}^{++}(n)$

$$\begin{aligned}\odot : \mathrm{Sym}^{++}(n) \times \mathrm{Sym}^{++}(n) &\to \mathrm{Sym}^{++}(n),\\ A \odot B &= \exp(\log(A) + \log(B)).\end{aligned} \tag{5.7}$$

Under the Log-Euclidean metric, the geodesic distance between A and B is given by

$$d_{\mathrm{logE}}(A,B) = ||\log(A) - \log(B)||_F. \tag{5.8}$$

Along with the group operation $\odot$, one can also define the scalar multiplication

$$\begin{aligned} \circledast : \mathbb{R} \times \mathrm{Sym}^{++}(n) &\to \mathrm{Sym}^{++}(n), \\ \lambda \circledast A = \exp(\lambda \log(A)) &= A^{\lambda}, \quad \lambda \in \mathbb{R}. \end{aligned} \tag{5.9}$$

Endowed with the commutative group multiplication $\odot$ and the scalar multiplication $\circledast$, $(\mathrm{Sym}^{++}, \odot, \circledast)$ becomes a vector space [4]. Furthermore, we can endow this vector space with the *Log-Euclidean inner product*.

$$\langle A, B\rangle_{\mathrm{logE}} = \langle \log(A), \log(B)\rangle_F = \mathrm{tr}[\log(A)\log(B)]. \tag{5.10}$$

along with the corresponding *Log-Euclidean norm*

$$||A||^2_{\mathrm{logE}} = \langle \log(A), \log(A)\rangle_F = \mathrm{tr}[\log^2(A)], \tag{5.11}$$

giving us the inner product space

$$(\mathrm{Sym}^{++}(n), \odot, \circledast, \langle\ ,\ \rangle_{\mathrm{logE}}). \tag{5.12}$$

This inner product space structure was first discussed in [24]. The Log-Euclidean distance in Eq. (5.8) is then expressed as

$$d_{\mathrm{logE}}(A,B) = ||\log(A) - \log(B)||_F = ||A \odot B^{-1}||_{\mathrm{logE}}. \tag{5.13}$$

With this viewpoint, it follows that $\mathrm{Sym}^{++}(n)$ under the Log-Euclidean metric is flat, that is it has *zero sectional curvature*. Furthermore, the map

$$\begin{aligned} \log : (\mathrm{Sym}^{++}(n), \odot, \circledast, \langle\ ,\ \rangle_{\mathrm{logE}}) &\to (\mathrm{Sym}(n), +, \cdot, \langle\ ,\ \rangle_F) \\ A &\to \log(A). \end{aligned} \tag{5.14}$$

is an isometrical isomorphism between inner product spaces, where $(+, \cdot)$ denote the standard matrix addition and scalar multiplication operations, respectively.

Log-Euclidean versus Euclidean. The previous discussion shows that the Log-Euclidean metric essentially flattens $\mathrm{Sym}^{++}(n)$ via the map $A \to \log(A)$. However, the vector space operations $(\odot, \circledast)$ are *not* the Euclidean space operations $(+, \cdot)$ and $(\mathrm{Sym}^{++}(n), \odot, \circledast, \langle\ ,\ \rangle_{\mathrm{logE}})$ is *not* a vector subspace of Euclidean space. Furthermore, $(\mathrm{Sym}^{++}(n), ||\ ||_E)$ is an incomplete metric space, whereas, since $||\ ||_{\mathrm{logE}}$ is an inner product distance, the metric space $(\mathrm{Sym}^{++}(n), ||\ ||_{\mathrm{logE}})$ is complete, which is a desirable property when dealing with converging sequences of SPD matrices.

One can also clearly see that the SPD property of the matrices A and B is encoded by the principal matrix logarithms in the distance formula $||\log(A) - \log(B)||_F$ (if A has a negative eigenvalue, for example, its principal matrix logarithm is *not* even defined). This is in strong contrast to the Euclidean distance formula $||A - B||_F$, which depends only on the entries of $A - B$ and therefore does *not* reflect any inherent structure in A and B.

Kernel methods with the Log-Euclidean metric. For the purposes of kernel methods in machine learning and applications, since $(\mathrm{Sym}^{++}(n), \odot, \circledast, \langle\ ,\ \rangle_{\mathrm{logE}})$ is an inner product space, one can define positive definite kernels on $\mathrm{Sym}^{++}(n)$ using the inner product $\langle\ ,\ \rangle_{\mathrm{logE}}$ and the corresponding norm $||\ ||_{\mathrm{logE}}$. This enables us to apply kernel methods directly on $\mathrm{Sym}^{++}(n)$, as is done in [17, 18, 24]. In particular, we have the following result.

Proposition 1 *The following kernels* $K : \mathrm{Sym}^{++}(n) \times \mathrm{Sym}^{++}(n) \to \mathbb{R}$ *are positive definite*

$$K(A, B) = (c + \langle A, B\rangle_{\mathrm{logE}})^d = (c + \langle \log(A), \log(B)\rangle_F)^d, \quad c \geq 0, d \in \mathbb{N}. \tag{5.15}$$

$$\begin{aligned} K(A, B) &= \exp\left(-||A \odot B^{-1}||^p_{\mathrm{logE}}\right), \quad \sigma \neq 0, \quad 0 < p \leq 2, \\ &= \exp\left(-\frac{||\log(A) - \log(B)||^p_F}{\sigma^2}\right). \end{aligned} \tag{5.16}$$

Remark 1 The proofs of Proposition 1 and all subsequent propositions are given in the Appendix. The kernel K in Eq. (5.16) in particular generalizes the results in [17, 18, 24], which show that K is positive definite for $p = 2$.

Bregman divergences. In the third approach, one defines distance-like functions based on the convex cone structure of $\mathrm{Sym}^{++}(n)$. One well-known example of this approach is the Stein divergence, defined by [38]

$$d^2_{\mathrm{stein}}(A, B) = \log \frac{\det(\frac{A+B}{2})}{\sqrt{\det(A)\det(B)}}. \tag{5.17}$$

The Bregman divergences do not arise from Riemannian metrics on $\mathrm{Sym}^{++}(n)$ and, apart from special cases such as d_{stein} in Eq. (5.17), they are generally *not* metric distances. However, they can be computed efficiently and have been shown to work well in diverse applications [11, 19].

In this chapter, we show how to generalize both the affine-invariant distance in Eq. (5.6) and the Log-Euclidean distance in Eq. (5.8) to the infinite-dimensional setting, in particular to RKHS covariance operators. The generalization of the Bregman divergences to the infinite-dimensional setting will be presented in a separate work.

5.3 Infinite-Dimensional Covariance Operators

Having reviewed the data representation framework using finite-dimensional covariance matrices, we now present infinite-dimensional covariance operators in RKHS and show how they can be used as a form of data representation. This framework is grounded in the setting of positive definite kernels and their associated RKHS and feature maps, which we discuss first.

5.3.1 Positive Definite Kernels, Reproducing Kernel Hilbert Spaces, and Feature Maps

Positive definite kernels. Let $\mathscr{X}$ be an arbitrary non-empty set. A function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is said to be a positive definite kernel if it is symmetric and satisfies

$$\sum_{i,j=1}^{N} a_i a_j K(x_i, x_j) \geq 0 \tag{5.18}$$

for any set of points $\mathbf{x} = \{x_i\}_{i=1}^N$ in $\mathscr{X}$ and any set of real numbers $\{a_i\}_{i=1}^N$. In other words, the $N \times N$ matrix $K[\mathbf{x}]$ defined by $(K[\mathbf{x}])_{ij} = K(x_i, x_j)$ is symmetric, positive semi-definite.

Examples of commonly used positive definite kernels include the Gaussian kernel $K(x, y) = \exp\left(-\frac{||x-y||^2}{\sigma^2}\right)$, $\sigma \neq 0$, and polynomial kernels $K(x, y) = (\langle x, y\rangle + c)^d$, $c \geq 0$, $d \in \mathbb{N}$, for $x, y \in \mathbb{R}^n$, $n \in \mathbb{N}$.

Reproducing kernel Hilbert spaces (RKHS). Each positive definite kernel K corresponds to a unique Hilbert space of functions on $\mathscr{X}$ as follows. For each $x \in \mathscr{X}$, there corresponds a function $K_x : \mathscr{X} \to \mathbb{R}$ defined by $K_x(y) = K(x, y)$. Consider the set $\mathscr{H}_0$ of all linear combinations of functions of the form K_x, $x \in \mathscr{X}$, that is

$$\mathscr{H}_0 = \left\{ \sum_{j=1}^{N} a_j K_{x_j} \ : \ a_j \in \mathbb{R}, x_j \in \mathscr{X}, N \in \mathbb{N} \right\}. \tag{5.19}$$

On $\mathscr{H}_0$, we define the following inner product

$$\langle \sum_{i=1}^{N} a_i K_{x_i}, \sum_{j=1}^{M} b_j K_{y_j} \rangle_{\mathscr{H}_K} = \sum_{i=1}^{N} \sum_{j=1}^{M} a_i b_j K(x_i, y_j). \tag{5.20}$$

This inner product is well-defined by the assumption that K is a positive definite kernel, making $\mathscr{H}_0$ an inner product space. Let $\mathscr{H}_K$ be the Hilbert completion of $\mathscr{H}_0$, obtained by adding the limits of all the Cauchy sequences in $\mathscr{H}_0$, then $\mathscr{H}_K$

is a Hilbert space of functions on $\mathscr{X}$, called the *reproducing kernel Hilbert space* (RKHS) induced by the kernel K.

The terminology RKHS comes from the *reproducing property*, which states that for all $f \in \mathscr{H}_K$ and all $x \in \mathscr{X}$,

$$f(x) = \langle f, K_x \rangle_{\mathscr{H}_K}. \tag{5.21}$$

Feature maps. A very useful and intuitive geometrical view of positive definite kernels is that of *feature maps*, which comes from machine learning and pattern recognition. In this viewpoint, a function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is a positive definite kernel if and only if there exists a Hilbert space $\mathscr{H}$, called *feature space*, and a map $\Phi : \mathscr{X} \to \mathscr{H}$, called *feature map*, such that

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathscr{H}} \quad \forall x, y \in \mathscr{H}. \tag{5.22}$$

As the simplest example, consider the quadratic polynomial kernel $K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ defined by $K(x, y) = \langle x, y \rangle^2 = (x_1y_1 + x_2y_2)^2$. It can be readily verified, via a simple algebraic calculation, that this kernel possesses the 3-dimensional feature map $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$, defined by

$$\Phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \in \mathbb{R}^3.$$

For a general positive definite kernel K, from the definition of RKHS above, it follows that the RKHS $\mathscr{H}_K$ induced by K is a feature space associated with K, with corresponding feature map $\Phi : \mathscr{X} \to \mathscr{H}_K$, defined by

$$\Phi(x) = K_x \quad \forall x \in \mathscr{X}, \tag{5.23}$$

which is called the *canonical feature map* [27] associated with K. If $\mathscr{X} \subset \mathbb{R}^n$ is a set with non-empty interior, then $\dim(\mathscr{H}_K) = \infty$ (see [25]), so that the feature map Φ is infinite-dimensional. We refer to [27] for a more detailed discussion of feature maps, including their equivalence to the canonical feature map, and many other examples.

The feature map viewpoint is particularly useful from an algorithmic perspective, since it allows one to transform any *linear* algorithm, which is expressed in terms of the inner product $\langle x, y \rangle$ of input examples in Euclidean space, into a *nonlinear* algorithm, simply by replacing $\langle x, y \rangle$ with $\langle \Phi(x), \Phi(y) \rangle_{\mathscr{H}_K} = K(x, y)$ for some nonlinear kernel K.

For our present purposes, feature maps enable us to generalize covariance matrices, which encode *linear correlations* between input features, to covariance operators in RKHS, which encode *nonlinear correlations* between input features.

5.3.2 Covariance Operators in RKHS and Data Representation

With the kernels and feature maps in Sect. 5.3.1, we now define RKHS covariance operators using these feature maps and show how they are employed for image representation. This framework is a generalization of the covariance matrix representation described in Sect. 5.2.

As in Sect. 5.2, for each image, let $\mathbf{x} = [x_1, \ldots, x_m]$ be the data matrix of size $n \times m$, with each column being the vector of features $x_i \in \mathbb{R}^n$ sampled at pixel i, $1 \leq i \leq m$. Now let $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a positive definite kernel, such as the Gaussian kernel, which induces implicitly a feature map $\Phi : \mathbb{R}^n \to \mathscr{H}_K$, where $\mathscr{H}_K$ is the RKHS induced by K. The map Φ gives us the matrix of features in $\mathscr{H}_K$

$$\Phi(\mathbf{x}) = [\Phi(x_1), \ldots, \Phi(x_m)], \tag{5.24}$$

which can be viewed informally as a (potentially infinite) matrix of size $\dim(\mathscr{H}_K) \times m$. Formally, it is a bounded linear operator $\Phi(\mathbf{x}) : \mathbb{R}^m \to \mathscr{H}_K$, defined by

$$\Phi(\mathbf{x})\mathbf{b} = \sum_{i=1}^{m} b_i \Phi(x_i), \tag{5.25}$$

with corresponding adjoint operator $\Phi(\mathbf{x})^* : \mathscr{H}_K \to \mathbb{R}^m$. The operator $\Phi(\mathbf{x})$ gives rise to the RKHS covariance operator

$$C_{\Phi(\mathbf{x})} = \frac{1}{m}\Phi(\mathbf{x}) J_m \Phi(\mathbf{x})^* : \mathscr{H}_K \to \mathscr{H}_K, \tag{5.26}$$

which can be viewed informally as a (potentially infinite) matrix of size $\dim(\mathscr{H}_K) \times \dim(\mathscr{H}_K)$. The covariance operator $C_{\Phi(\mathbf{x})}$ is now the *representation* for the image and encodes, for a nonlinear kernel K, *nonlinear correlations* between all the different extracted features.

Remark 2 We say that the covariance operator representation is a generalization of the covariance matrix representation, since for the linear kernel $K(x, y) = \langle x, y \rangle$ on $\mathbb{R}^n \times \mathbb{R}^n$, we have $\Phi(x) = x$ and $C_{\Phi(\mathbf{x})} = C_{\mathbf{x}}$.

A crucial feature of the RKHS covariance operator representation is that it is *implicit*, that is neither the matrix of features $\Phi(\mathbf{x})$ nor the covariance operator $C_{\Phi(\mathbf{x})}$ is ever computed. Instead, all the necessary computations involving $\Phi(\mathbf{x})$ and $C_{\Phi(\mathbf{x})}$ are done via the Gram matrices of the kernel K on the original data matrix $\mathbf{x}$. We show that this is indeed the case for the distances between the covariance operators.

5.4 Distances Between RKHS Covariance Operators

Having described the image representation framework by RKHS covariance operators, we now describe the distances between covariance operators. These distances can then be directly employed in a practical application, e.g. image classification.

Since covariance operators are Hilbert–Schmidt operators, a natural distance between them is the Hilbert–Schmidt distance, which is the infinite-dimensional generalization of the Euclidean distance given by the Frobenius norm $|| \; ||_F$. However, just like the Euclidean distance, the Hilbert–Schmidt distance does *not* capture the *positivity* of covariance operators. In order to do so, as with $\mathrm{Sym}^{++}(n)$, we need to consider the manifold setting of covariance operators.

As a generalization from the finite-dimensional setting, it can be shown [22] that regularized covariance operators lie on an *infinite-dimensional* Hilbert manifold, namely the manifold of *positive definite unitized Hilbert–Schmidt operators* on a separable Hilbert space $\mathscr{H}$. Each point on this manifold has the form $A + \gamma I > 0$, $\gamma > 0$, where A is a Hilbert–Schmidt operator on $\mathscr{H}$. As we now show, both the affine-invariant distance in Eq. (5.6) and the Log-Euclidean distance in Eq. (5.8) admit a generalization on this manifold. However, there are several *key differences* between the finite and infinite-dimensional settings:

1. In the finite-dimensional case, the regularization $(C_{\mathbf{x}} + \gamma I)$ is often necessary *empirically* since in general $C_{\mathbf{x}}$ is not guaranteed to be positive definite. In contrast, when $\dim(\mathscr{H}) = \infty$, the regularization form $(A + \gamma I)$ is *always* needed, both *theoretically* and *empirically*, even if A is strictly positive. This is because $\log(A)$ is unbounded and we must always consider $\log(A + \gamma I)$. We explain this in detail in Sect. 5.4.2.1.
2. When $\dim(\mathscr{H}) = \infty$, the identity operator I is not Hilbert–Schmidt and therefore the Hilbert–Schmidt norm of $\log(A + \gamma I)$ is generally infinite. Furthermore, the distance between any two different multiples of I would be infinite. This problem is resolved by the introduction of the *extended Hilbert–Schmidt inner product*. We explain this in detail in Sect. 5.4.2.2.

In general, the distance formulas for the finite and infinite-dimensional cases are different and the infinite-dimensional formulas are generally *not* the limits of the finite-dimensional ones as the dimension approaches infinity. For RKHS covariance operators, all three distances admit closed forms in terms of Gram matrices.

5.4.1 Hilbert–Schmidt Distance

We first consider the generalization of the Frobenius norm in Eq. (5.4) to the separable Hilbert space setting. We recall that a bounded linear operator $A : \mathscr{H} \to \mathscr{H}$ is said to be a Hilbert–Schmidt operator if

$$||A||_{\mathrm{HS}}^2 = \mathrm{tr}(A^*A) = \sum_{k=1}^{\infty} ||Ae_k||^2 < \infty, \quad (5.27)$$

for any countable orthonormal basis $\{e_k\}_{k\in\mathbb{N}}$ in $\mathscr{H}$. $||\ ||_{\mathrm{HS}}$ is called the Hilbert–Schmidt norm, which is the infinite-dimensional version of the Frobenius norm in Eq. (5.4).

Let $\mathrm{HS}(\mathscr{H})$ denote the class of all Hilbert–Schmidt operators on $\mathscr{H}$. The Hilbert–Schmidt norm corresponds to the Hilbert–Schmidt inner product on $\mathrm{HS}(\mathscr{H})$, which is defined by

$$\langle A, B\rangle_{\mathrm{HS}} = \mathrm{tr}(A^*B) = \sum_{k=1}^{\infty}\langle Ae_k, Be_k\rangle, \quad A, B \in \mathrm{HS}(\mathscr{H}). \quad (5.28)$$

For a self-adjoint operator $A \in \mathrm{HS}(\mathscr{H})$, A is compact and hence possesses a countable spectrum $\{\lambda_k\}_{k=1}^{\infty}$, with $\lim_{k\to\infty}\lambda_k = 0$, and

$$||A||_{\mathrm{HS}}^2 = \sum_{k=1}^{\infty}\lambda_k^2 < \infty. \quad (5.29)$$

It is clear then that if $\dim(\mathscr{H}) = \infty$, then the identity operator I is not Hilbert–Schmidt, since obviously

$$||I||_{\mathrm{HS}} = \infty.$$

We explain the consequence of this fact on the infinite-dimensional generalization of the affine-invariant and Log-Euclidean distances in Sect. 5.4.2.2.

For two RKHS covariance operators $C_{\Phi(\mathbf{x})}$ and $C_{\Phi(\mathbf{y})}$, their Hilbert–Schmidt distance is expressed explicitly in terms of Gram matrices, as follows. Let $K[\mathbf{x}]$, $K[\mathbf{y}]$, $K[\mathbf{x},\mathbf{y}]$ denote the $m \times m$ matrices defined by

$$(K[\mathbf{x}])_{ij} = K(x_i, x_j), \quad (K[\mathbf{y}])_{ij} = K(y_i, y_j), \quad (K[\mathbf{x},\mathbf{y}])_{ij} = K(x_i, y_j). \quad (5.30)$$

By definition of feature maps, we have $K(x, y) = \langle\Phi(x), \Phi(y)\rangle_{\mathscr{H}_K}\ \forall (x, y) \in \mathscr{X}\times\mathscr{X}$, so that the Gram matrices and the feature maps are closely related as follows

$$K[\mathbf{x}] = \Phi(\mathbf{x})^*\Phi(\mathbf{x}), \quad K[\mathbf{y}] = \Phi(\mathbf{y})^*\Phi(\mathbf{y}), \quad K[\mathbf{x},\mathbf{y}] = \Phi(\mathbf{x})^*\Phi(\mathbf{y}). \quad (5.31)$$

Lemma 1 *The Hilbert–Schmidt distance between two RKHS covariance operators $C_{\Phi(\mathbf{x})}$ and $C_{\Phi(\mathbf{y})}$ is given by*

$$||C_{\Phi(\mathbf{x})} - C_{\Phi(\mathbf{y})}||_{\mathrm{HS}}^2 = \frac{1}{m^2}\langle J_m K[\mathbf{x}], K[\mathbf{x}]J_m\rangle_F - \frac{2}{m^2}\langle J_m K[\mathbf{x},\mathbf{y}], K[\mathbf{x},\mathbf{y}]J_m\rangle_F + \frac{1}{m^2}\langle J_m K[\mathbf{y}], K[\mathbf{y}]J_m\rangle_F. \quad (5.32)$$

Remark 3 For the linear kernel $K(x, y) = \langle x, y\rangle_{\mathbb{R}^n}$, we have $\Phi(\mathbf{x}) = \mathbf{x}$, $\Phi(\mathbf{y}) = \mathbf{y}$ and thus Eq. (5.32) gives us the Euclidean distance $||C_{\mathbf{x}} - C_{\mathbf{y}}||_F$.

5.4.2 Riemannian Distances Between Covariance Operators

We now show how the affine-invariant and Log-Euclidean distances in Eqs. (5.6) and (5.8), respectively, can be generalized to the infinite-dimensional settings, specifically to self-adjoint, positive Hilbert–Schmidt operators on a separable Hilbert space $\mathscr{H}$. These generalizations were carried out recently in [21, 22], for the affine-invariant metric, and in [28], for the Log-Euclidean metric. As a special case of these formulations, we obtain the respective distances between infinite-dimensional covariance operators on Hilbert spaces, which assume explicit forms in the RKHS setting [26, 28].

Looking at Eqs. (5.6) and (5.8), we see that generalizing them to the infinite-dimensional setting requires the following two steps

1. Generalization of the set $\mathrm{Sym}^{++}(n)$ of all $n \times n$ SPD matrices and the corresponding generalization for the principal matrix logarithm.
2. Generalization of the Frobenius norm $||\ ||_F$.

We show next that the first step leads us to the concept of *positive definite operators* and the second to the concept of *extended Hilbert–Schmidt norm.*

5.4.2.1 Positive Definite Operators

We first consider the bounded operators $A : \mathscr{H} \to \mathscr{H}$ that generalize $n \times n$ SPD matrices and the corresponding generalization for the principal matrix logarithm. To this end, we first recall the definition of the principal matrix logarithm for an SPD matrix A of size $n \times n$. Let $\{\lambda_k\}_{k=1}^n$ be the eigenvalues of A, which are all positive, with corresponding normalized eigenvectors $\{\mathbf{u}_k\}_{k=1}^n$. Then A admits the spectral decomposition

$$A = \sum_{k=1}^{n} \lambda_k \mathbf{u}_k \mathbf{u}_k^T.$$

The principal matrix logarithm of A is then given by

$$\log(A) = \sum_{k=1}^{n} \log(\lambda_k) \mathbf{u}_k \mathbf{u}_k^T. \tag{5.33}$$

To generalize this formula to the Hilbert space setting, let us assume that $A : \mathscr{H} \to \mathscr{H}$ is a self-adjoint, compact operator, so that it possesses a countable spectrum

$\{\lambda_k\}_{k=1}^{\infty}$, with corresponding normalized eigenvectors $\{\mathbf{u}_k\}_{k=1}^{\infty}$. The eigenvalues λ_k's are all real and satisfy $\lim_{k\to\infty} \lambda_k = 0$. We assume next that A is *strictly positive*, that is

$$\langle x, Ax \rangle > 0 \quad \forall x \in \mathscr{H}, x \neq 0. \tag{5.34}$$

Then the eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$ of A are all strictly positive. However, in contrast to the case $\dim(\mathscr{H}) < \infty$, when $\dim(\mathscr{H}) = \infty$, strict positivity is *not* a sufficient condition for $\log(A)$ to be well-defined. To see why, consider the following direct generalization of the principal matrix logarithm in Eq. (5.33),

$$\log(A) = \sum_{k=1}^{\infty} (\log \lambda_k) \mathbf{u}_k \otimes \mathbf{u}_k : \mathscr{H} \to \mathscr{H}, \tag{5.35}$$

where $\mathbf{u}_k \otimes \mathbf{u}_k : \mathscr{H} \to \mathscr{H}$ is a rank-one operator defined by $(\mathbf{u}_k \otimes \mathbf{u}_k)w = \langle \mathbf{u}_k, w \rangle \mathbf{u}_k$, which directly generalizes the rank-one matrix $\mathbf{u}_k \mathbf{u}_k^T$. Thus

$$\log(A)w = \sum_{k=1}^{\infty} (\log \lambda_k) \langle \mathbf{u}_k, w \rangle \mathbf{u}_k \quad \forall w \in \mathscr{H}. \tag{5.36}$$

However, since $\lim_{k\to\infty} \log \lambda_k = -\infty$, this operator is *unbounded*. Thus in particular, if A is a covariance operator, then $\log(A)$ is unbounded.

This problem can be resolved via regularization as follows. Instead of considering $\log(A)$, we consider $\log(A + \gamma I)$, $\gamma > 0$, and we see immediately that

$$\log(A + \gamma I) = \sum_{k=1}^{\infty} [\log(\lambda_k + \gamma)] \mathbf{u}_k \otimes \mathbf{u}_k : \mathscr{H} \to \mathscr{H}, \tag{5.37}$$

is a bounded operator $\forall \gamma > 0$. The operator $A + \gamma I$ is an example of a *positive definite operator*, that is it is a member of the set

$$\mathbb{P}(\mathscr{H}) = \{B \in \mathscr{L}(\mathscr{H}) : \exists M_B > 0 \text{ such that } \langle x, Bx \rangle \geq M_B ||x||^2 \;\; \forall x \in \mathscr{H}\}. \tag{5.38}$$

Clearly, if $B \in \mathbb{P}(\mathscr{H})$, then the eigenvalues of B, if they exist, are all bounded below by the constant $M_B > 0$. Thus in the infinite-dimensional setting, positive definiteness is a stronger requirement than strict positivity. In fact, it can be shown that

$$B \text{ positive definite} \Longleftrightarrow B \text{ strictly positive and invertible.} \tag{5.39}$$

Subsequently, we use the notation $B > 0$ for $B \in \mathbb{P}(\mathscr{H})$.

5.4.2.2 Extended Hilbert–Schmidt Inner Product and Norm

We now consider operators of the form $A + \gamma I > 0$, where A is a self-adjoint, compact operators, so that $\log(A + \gamma I)$, as given by Eq. (5.37) is well-defined and bounded. To generalize Eq. (5.8), we then need to have

$$|| \log(A + \gamma I)||_{\mathrm{HS}}^2 = \sum_{k=1}^{\infty} [\log(\lambda_k + \gamma)]^2 < \infty. \quad (5.40)$$

We show below that this is also sufficient for the generalization of Eq. (5.6). For $\gamma = 1$, this holds if and only if $A \in \mathrm{HS}(\mathscr{H})$, as shown by the following.

Proposition 2 *Assume that $A + I > 0$, with A being a self-adjoint, compact operator. Then $\log(A + I) \in \mathrm{HS}(\mathscr{H})$ if and only if $A \in \mathrm{HS}(\mathscr{H})$.*

However, for $\gamma \neq 1$, $\gamma > 0$, $\log(A + \gamma I)$ *cannot* be a Hilbert–Schmidt operator for any compact operator A with $A + \gamma I > 0$, as shown in the following.

Proposition 3 *Assume that $A + \gamma I > 0$, $\gamma > 0$, $\gamma \neq 1$, with A being a self-adjoint, compact operator. Then $\log(A + \gamma I) \notin \mathrm{HS}(\mathscr{H})$.*

The result given in Proposition 3 is due to the fact that $||I||_{\mathrm{HS}} = \infty$, as can be viewed via the decomposition

$$\log(A + \gamma I) = \log\left(\frac{A}{\gamma} + I\right) + \log(\gamma) I. \quad (5.41)$$

On the right handside, the first term $\log\left(\frac{A}{\gamma} + I\right)$ is Hilbert–Schmidt if and only if A is Hilbert–Schmidt, as guaranteed by Proposition 2. However, for $\gamma \neq 1$, the second term $\log(\gamma)I$ *cannot* be Hilbert–Schmidt, since $||I||_{\mathrm{HS}} = \infty$ for $\dim(\mathscr{H}) = \infty$.

Thus, taken together, Propositions 2 and 3 show that to generalize Eqs. (5.6) and (5.8), we need to consider operators of the form $A + \gamma I > 0$, where A is Hilbert–Schmidt, and at the same time, extend the definition of the Hilbert–Schmidt inner product and norm so that the norm of the identity operator I is finite.

The desired extension is called the *extended Hilbert–Schmidt inner product* $\langle\ ,\ \rangle_{\mathrm{eHS}}$ [21, 22], which is defined by

$$\langle A + \gamma I, B + \mu I \rangle_{\mathrm{eHS}} = \langle A, B \rangle_{\mathrm{HS}} + \gamma\mu. \quad (5.42)$$

Under the extended Hilbert–Schmidt inner product, the scalar operators γI are orthogonal to the Hilbert–Schmidt operators. The corresponding *extended Hilbert–Schmidt norm* is then given by

$$||A + \gamma I||_{\mathrm{eHS}}^2 = ||A||_{\mathrm{HS}}^2 + \gamma^2. \quad (5.43)$$

One can see that this is a form of compactification, which gives $||I||_{\text{eHS}} = 1$, in contrast to the infinite Hilbert–Schmidt norm $||I||_{\text{HS}} = \infty$. Thus instead of the Hilbert space of self-adjoint Hilbert–Schmidt operators, we consider the Hilbert space of self-adjoint *extended (or unitized) Hilbert–Schmidt operators*

$$\mathscr{H}_{\mathbb{R}} = \{A + \gamma I \ : \ A^* = A, \ A \in \text{HS}(\mathscr{H}), \ \gamma \in \mathbb{R}\}, \tag{5.44}$$

under the extended Hilbert–Schmidt inner product. By the decomposition given in Eq. (5.41), we immediately obtain the following.

Proposition 4 *Assume that $A + \gamma I > 0$ where $\gamma > 0$ and A is a self-adjoint, compact operator. Then*

$$\log(A + \gamma I) \in \mathscr{H}_{\mathbb{R}} \Longleftrightarrow A + \gamma I > 0, \ A \in \text{HS}(\mathscr{H}), \ A^* = A. \tag{5.45}$$

Furthermore, when $\dim(\mathscr{H}) = \infty$*,*

$$|| \log(A + \gamma I)||^2_{\text{eHS}} = \left\|\log\left(\frac{A}{\gamma} + I\right)\right\|^2_{\text{HS}} + (\log\gamma)^2. \tag{5.46}$$

5.4.2.3 The Hilbert Manifold of Positive Definite Unitized Hilbert–Schmidt Operators

In summary, to generalize Eqs. (5.6) and (5.8) to the Hilbert space setting, we need to consider operators of the form $A + \gamma I > 0$, where A is self-adjoint, Hilbert–Schmidt, so that $\log(A + \gamma I)$ is well-defined and bounded, along with the extended Hilbert–Schmidt norm $|| \ ||_{\text{eHS}}$, so that $|| \log(A + \gamma I)||_{\text{eHS}}$ is finite. We have thus arrived at the following generalization of $\text{Sym}^{++}(n)$

$$\Sigma(\mathscr{H}) = \mathbb{P}(\mathscr{H}) \cap \mathscr{H}_{\mathbb{R}} = \{A + \gamma I > 0 \ : \ A^* = A, \ A \in \text{HS}(\mathscr{H}), \ \gamma \in \mathbb{R}\}, \tag{5.47}$$

which was first introduced by [21, 22]. This is an infinite-dimensional Hilbert manifold, with the tangent space at each point $T_P(\Sigma(\mathscr{H})) \cong \mathscr{H}_{\mathbb{R}} \ \forall P \in \Sigma(\mathscr{H})$. By Proposition 4,

$$A + \gamma I \in \Sigma(\mathscr{H}) \Longleftrightarrow \log(A + \gamma I) \in \mathscr{H}_{\mathbb{R}}. \tag{5.48}$$

5.4.3 The Affine-Invariant Distance

The affine-invariant Riemannian metric was introduced on the Hilbert manifold $\Sigma(\mathscr{H})$ by [21, 22]. Under this metric, the geodesic distance between any two positive definite operators $(A+\gamma I), (B+\mu I) \in \Sigma(\mathscr{H})$ is given by

$$d_{\mathrm{aiHS}}[(A+\gamma I), (B+\mu I)] = ||\log[(A+\gamma I)^{-1/2}(B+\mu I)(A+\gamma I)^{-1/2}]||_{\mathrm{eHS}}. \tag{5.49}$$

The following result confirms that the distance $d_{\mathrm{aiHS}}[(A+\gamma I), (B+\mu I)]$ is always finite for any pair of operators $(A+\gamma I), (B+\mu I) \in \Sigma(\mathscr{H})$.

Proposition 5 *For any two operators* $(A+\gamma I), (B+\mu I) \in \Sigma(\mathscr{H})$*, we can write* $(A+\gamma I)^{-1/2}(B+\mu I)(A+\gamma I)^{-1/2} = Z + \nu I > 0$ *for* $\nu = \frac{\mu}{\gamma}$ *and* $Z = (A+\gamma I)^{-1/2} B(A+\gamma I)^{-1/2} - \frac{\mu}{\gamma}A(A+\gamma I)^{-1}$ *satisfying* $Z = Z^*,\ Z \in \mathrm{HS}(\mathscr{H})$*. Thus the affine-invariant geodesic distance*

$$d_{\mathrm{aiHS}}[(A+\gamma I), (B+\mu I)] = ||\log(Z+\nu I)||_{\mathrm{eHS}} \tag{5.50}$$

is always finite. Furthermore, when $\dim(\mathscr{H}) = \infty$*,*

$$d^2_{\mathrm{aiHS}}[(A+\gamma I), (B+\mu I)] = \left\|\log\left(\frac{Z}{\nu}+I\right)\right\|^2_{\mathrm{HS}} + (\log \nu)^2. \tag{5.51}$$

In the RKHS setting, the affine-invariant distance between regularized RKHS covariance operators $d_{\mathrm{aiHS}}[(C_{\Phi(\mathbf{x})}+\gamma I), (C_{\Phi(\mathbf{y})}+\mu I)]$ admits a closed form, which was given by [26], as follows.

Theorem 1 ([26]) *Assume that* $\dim(\mathscr{H}_K) = \infty$*. Let* $\gamma > 0,\ \mu > 0$*. Then*

$$d^2_{\mathrm{aiHS}}[(C_{\Phi(\mathbf{x})}+\gamma I), (C_{\Phi(\mathbf{y})}+\mu I)] = \mathrm{tr}\left\{\log\left[\begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{11} & C_{12} & C_{13} \end{pmatrix} + I_{3m}\right]\right\}^2 + \left(\log\frac{\gamma}{\mu}\right)^2, \tag{5.52}$$

where the $m \times m$ *matrices* C_{ij}*,* $i, j = 1, 2, 3$*, are given by*

$$C_{11} = \frac{1}{\mu m}J_m K[\mathbf{y}]J_m,$$

$$C_{12} = -\frac{1}{\sqrt{\gamma\mu}m}J_m K[\mathbf{y},\mathbf{x}]J_m\left(I_m + \frac{1}{\gamma m}J_m K[\mathbf{x}]J_m\right)^{-1},$$

$$C_{13} = -\frac{1}{\gamma\mu m^2}J_m K[\mathbf{y},\mathbf{x}]J_m\left(I_m + \frac{1}{\gamma m}J_m K[\mathbf{x}]J_m\right)^{-1}J_m K[\mathbf{x},\mathbf{y}]J_m,$$

$$C_{21} = \frac{1}{\sqrt{\gamma\mu}m} J_m K[\mathbf{x}, \mathbf{y}] J_m,$$

$$C_{22} = -\frac{1}{\gamma m} J_m K[\mathbf{x}] J_m \left(I_m + \frac{1}{\gamma m} J_m K[\mathbf{x}] J_m\right)^{-1},$$

$$C_{23} = -\frac{1}{\gamma m} J_m K[\mathbf{x}] J_m \left(I_m + \frac{1}{\gamma m} J_m K[\mathbf{x}] J_m\right)^{-1} \frac{1}{\sqrt{\gamma\mu}m} J_m K[\mathbf{x}, \mathbf{y}] J_m.$$

Theorem 2 ([26]) *Assume that* $\dim(\mathscr{H}_K) < \infty$. *Let* $\gamma > 0$, $\mu > 0$. *Then*

$$d^2_{\mathrm{aiHS}}[(C_{\Phi(\mathbf{x})} + \gamma I), (C_{\Phi(\mathbf{y})} + \mu I)] = \mathrm{tr}\left\{\log\left[\begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{11} & C_{12} & C_{13} \end{pmatrix} + I_{3m}\right]\right\}^2$$
$$-2\left(\log\frac{\gamma}{\mu}\right)\mathrm{tr}\left\{\log\left[\begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{11} & C_{12} & C_{13} \end{pmatrix} + I_{3m}\right]\right\} + \left(\log\frac{\gamma}{\mu}\right)^2 \dim(\mathscr{H}_K), \quad (5.53)$$

where the $m \times m$ *matrices* C_{ij}*'s,* $i, j = 1, 2, 3$*, are as in Theorem 1.*

We see that the formula for the affine-invariant distance for the case $\dim(\mathscr{H}_K) = \infty$ is generally different from that for the case $\dim(\mathscr{H}_K) < \infty$, except when $\gamma = \mu$, in which case they are identical. One can see that for $m \in \mathbb{N}$ fixed, $\gamma \neq \mu$, the right hand side of Eq. (5.53) approaches infinity as $\dim(\mathscr{H}_K) \to \infty$. Thus for $\gamma \neq \mu$, one *cannot* approximate the infinite-dimensional distance in Eq. (5.52) by the finite-dimensional distance in Eq. (5.53).

5.4.3.1 Log-Hilbert–Schmidt Distance

Similar to the affine-invariance distance in Eq. (5.49), the generalization of the Log-Euclidean distance [4] is the Log-Hilbert–Schmidt distance

$$d_{\mathrm{logHS}}[(A + \gamma I), (B + \mu I)] = ||\log(A + \gamma I) - \log(B + \mu I)||_{\mathrm{eHS}}, \quad (5.54)$$

which was recently formulated by [28]. The well-definedness of this distance for any pair of operators $(A + \gamma I), (B + \mu I) \in \Sigma(\mathscr{H})$ is confirmed by the following result.

Proposition 6 *For any pair of operators* $(A + \gamma I), (B + \mu I) \in \Sigma(\mathscr{H})$*, the distance* $d_{\mathrm{logHS}}[(A + \gamma I), (B + \mu I)] = ||\log(A + \gamma I) - \log(B + \mu I)||_{\mathrm{eHS}}$ *is always finite. Furthermore, when* $\dim(\mathscr{H}) = \infty$*,*

$$||\log(A + \gamma I) - \log(B + \mu I)||^2_{\mathrm{eHS}} = \left\|\log\left(\frac{A}{\gamma} + I\right) - \log\left(\frac{B}{\mu} + I\right)\right\|^2_{\mathrm{HS}} + \left(\log\frac{\gamma}{\mu}\right)^2.$$

As in the case of the affine-invariant distance, in the case of regularized RKHS covariance operators, the Log-HS distance

$$d_{\mathrm{logHS}}[C_{\Phi(\mathbf{x})} + \gamma I, C_{\Phi(\mathbf{y})} + \mu I] = ||\log(C_{\Phi(\mathbf{x})} + \gamma I) - \log(C_{\Phi(\mathbf{y})} + \mu I)||_{\mathrm{eHS}} \tag{5.55}$$

also admits an explicit form, expressed via the Gram matrices corresponding to **x** and **y** [28]. To state this explicit form, we first define the following operators

$$A = \frac{1}{\sqrt{\gamma m}}\Phi(\mathbf{x})J_m : \mathbb{R}^m \to \mathscr{H}_K, \quad B = \frac{1}{\sqrt{\mu m}}\Phi(\mathbf{y})J_m : \mathbb{R}^m \to \mathscr{H}_K, \tag{5.56}$$

so that

$$A^*A = \frac{1}{\gamma m}J_m K[\mathbf{x}]J_m, \quad B^*B = \frac{1}{\mu m}J_m K[\mathbf{y}]J_m, \quad A^*B = \frac{1}{\sqrt{\gamma \mu}m}J_m K[\mathbf{x}, \mathbf{y}]J_m. \tag{5.57}$$

Let N_A and N_B be the numbers of nonzero eigenvalues of A^*A and B^*B, respectively. Let Σ_A and Σ_B be the diagonal matrices of size $N_A \times N_A$ and $N_B \times N_B$, and U_A and U_B be the matrices of size $m \times N_A$ and $m \times N_B$, respectively, which are obtained from the spectral decompositions

$$\frac{1}{\gamma m}J_m K[\mathbf{x}]J_m = U_A \Sigma_A U_A^T, \quad \frac{1}{\mu m}J_m K[\mathbf{y}]J_m = U_B \Sigma_B U_B^T. \tag{5.58}$$

Let $\circ$ denote the Hadamard (element-wise) matrix product and define

$$C_{AB} = \mathbf{1}_{N_A}^T \log(I_{N_A} + \Sigma_A)\Sigma_A^{-1}(U_A^T A^* B U_B \circ U_A^T A^* B U_B)\Sigma_B^{-1}\log(I_{N_B} + \Sigma_B)\mathbf{1}_{N_B}. \tag{5.59}$$

In terms of the quantities just defined, the Log-HS distance can be expressed as follows. As in the case of the affine-invariant distance, the distance formulas are different for the cases $\dim(\mathscr{H}_K) = \infty$ and $\dim(\mathscr{H}_K) < \infty$, with the finite-dimensional distance approaching infinity as $\dim(\mathscr{H}_K) \to \infty$.

Theorem 3 ([28]) *Assume that* $\dim(\mathscr{H}_K) = \infty$. *Let* $\gamma > 0$, $\mu > 0$. *Then*

$$\begin{aligned} d^2_{\mathrm{logHS}}[(C_{\Phi(\mathbf{x})} + \gamma I), (C_{\Phi(\mathbf{y})} + \mu I)] &= \mathrm{tr}[\log(I_{N_A} + \Sigma_A)]^2 + \mathrm{tr}[\log(I_{N_B} + \Sigma_B)]^2 \\ &\quad - 2C_{AB} + (\log\gamma - \log\mu)^2. \end{aligned} \tag{5.60}$$

Theorem 4 ([28]) *Assume that* $\dim(\mathscr{H}_K) < \infty$. *Let* $\gamma > 0$, $\mu > 0$. *Then*

$$\begin{aligned} d^2_{\mathrm{logHS}}[(C_{\Phi(\mathbf{x})} + \gamma I), (C_{\Phi(\mathbf{y})} + \mu I)] &= \mathrm{tr}[\log(I_{N_A} + \Sigma_A)]^2 + \mathrm{tr}[\log(I_{N_B} + \Sigma_B)]^2 - 2C_{AB} \\ &\quad + 2\left(\log\frac{\gamma}{\mu}\right)(\mathrm{tr}[\log(I_{N_A} + \Sigma_A)] - \mathrm{tr}[\log(I_{N_B} + \Sigma_B)]) \\ &\quad + (\log\gamma - \log\mu)^2 \dim(\mathscr{H}_K). \end{aligned} \tag{5.61}$$

Remark 4 In the case of the linear kernel $K(x, y) = \langle x, y\rangle$, $x, y \in \mathbb{R}^n$, Theorem 4 gives the Log-Euclidean distance $||\log(C_{\mathbf{x}} + \gamma I) - \log(C_{\mathbf{y}} + \mu I)||$.

Remark 5 We showed in [28] that the two operations $\odot$ and $\circledast$ on $\mathrm{Sym}^{++}(n)$ as defined in Sect. 5.2 can both be generalized to the Hilbert manifold $\Sigma(\mathscr{H})$, so that $(\Sigma(\mathscr{H}), \odot, \circledast)$ is a vector space. This vector space can be endowed with the *Log-Hilbert–Schmidt inner product*, defined by

$$\langle A + \gamma I, B + \mu I\rangle_{\mathrm{logHS}} = \langle \log(A + \gamma I), \log(B + \mu I)\rangle_{\mathrm{eHS}}. \tag{5.62}$$

With this inner product, the space $(\Sigma(\mathscr{H}), \odot, \circledast, \langle\ ,\ \rangle_{\mathrm{logHS}})$ is a Hilbert space and the distance in this Hilbert space is precisely the Log-Hilbert–Schmidt distance, see [28] for detail.

5.5 Two-Layer Kernel Machines with RKHS Covariance Operators

Having presented the explicit formulas for the affine-invariant and Log-Hilbert–Schmidt distances between RKHS covariance operators, we now show how the Log-Hilbert–Schmidt distance in particular can be used to design a two-layer kernel machine for machine learning, with an application in image classification.

5.5.1 The Interplay Between Positive Definite Kernels and Riemannian Manifolds

The geometric framework for RKHS covariance operators that we have just described reveals a close link between positive definite kernels and Riemannian manifolds, as follows.

Kernels giving rise to Manifolds. Let $\mathscr{X}$ be any non-empty set. Each positive definite kernel defined on $\mathscr{X} \times \mathscr{X}$ gives rise to a set of RKHS covariance operators, each of the form $C_{\Phi(\mathbf{x})}$, where $\mathbf{x}$ is a data matrix sampled from $\mathscr{X}$ according to a probability distribution. The corresponding set of regularized RKHS covariance operators $(C_{\Phi(\mathbf{x})} + \gamma I)$, $\gamma > 0$, forms a subset of the Hilbert manifold of positive definite Hilbert–Schmidt operators.

For the case of the Log-Hilbert–Schmidt distance, we have the link in the other direction as well.

Distances on Manifolds giving rise to Kernels. Since the Log-Hilbert–Schmidt distance is a Hilbert space distance, it can be used to define many positive definite kernels on $\Sigma(\mathscr{H}) \times \Sigma(\mathscr{H})$. The following result naturally generalizes Proposition 1 to the infinite-dimensional setting.

Proposition 7 ([28]) *The following kernels* $K : \Sigma(\mathscr{H}) \times \Sigma(\mathscr{H}) \to \mathbb{R}$ *are positive definite*

$$K[(A+\gamma I), (B+\mu I)] = (c + \langle \log(A+\gamma I), \log(B+\mu I)\rangle_{\mathrm{eHS}})^d, \quad c \geq 0, \; d \in \mathbb{N}, \tag{5.63}$$

$$K[(A+\gamma I), (B+\mu I)] = \exp\left(-\frac{||\log(A+\gamma I) - \log(B+\mu I)||^p_{\mathrm{eHS}}}{\sigma^2}\right), \tag{5.64}$$

for $0 < p \leq 2$.

5.5.2 Two-Layer Kernel Machines

The interplay between positive definite kernels and Riemannian manifolds as we described above allows us to design a two-layer kernel machine by utilizing the Log-Hilbert–Schmidt distance as follows.

1. In the first layer, a positive definite kernel, such as the Gaussian kernel, is applied to the original features extracted from each image, giving an *implicit* covariance operator representing that image. Using the Log-Hilbert–Schmidt distance, we then compute the pairwise distances between all the images.
2. In the second layer, using the pairwise Log-Hilbert–Schmidt distances obtained in the first layer, we define a new positive definite kernel, such as another Gaussian kernel. We can then apply any kernel method, such as SVM, using this kernel.

Remark 6 The approach in [17, 24], which applies kernel methods on top of the Log-Euclidean distance, is a special case our our framework, where the kernel in the first layer is linear (which is equivalent to *not* having the first layer).

5.6 Experiments in Image Classification

In this section, we report empirical results on the task of image classification using the two-layer kernel machine with the Log-Hilbert–Schmidt distance, as described in Sect. 5.5. The results obtained are substantially better than those obtained using the corresponding one-layer kernel machine with the Log-Euclidean distance. These thus clearly demonstrate the superior power and effectiveness of the covariance operator representation framework compared to the covariance matrix representation. The results presented here were first reported in [28].

We recall from our previous discussion that each image is represented by a covariance operator as follows. At every pixel (or a subset of pixels) of the image, we extract n low-level features, such as intensity and colors, giving us a low-level feature vector in $\mathbb{R}^n$. Sampling at m pixels in the image gives us a data matrix $\mathbf{x}$ of size $n \times m$. By applying a positive definite kernel, defined on $\mathbb{R}^n \times \mathbb{R}^n$, to the low-level feature vectors, we obtain *implicitly* a matrix of features $\Phi(\mathbf{x})$, as defined in Eq. (5.24), and the corresponding covariance operator $C_{\Phi(\mathbf{x})}$, as defined in Eq. (5.26). The image is then represented by the covariance operator $C_{\Phi(\mathbf{x})}$. In the current experiments, we used the Gaussian kernel and the resulting covariance operator is called *Gaussian-COV*. The distance between two images is the distance between the corresponding covariance operators, which in this case is the Log-Hilbert–Schmidt distance, given by Eq. (5.60) when $\dim(\mathscr{H}_K) = \infty$, e.g. for the Gaussian kernel, and Eq. (5.61) when $\dim(\mathscr{H}_K) < \infty$, e.g. for the polynomial kernels.

Given a set of images, we then have a corresponding set of covariance operators and a matrix of pairwise Log-Hilbert–Schmidt distances between these operators. In the following experiments, the task of image classification was carried out by applying Gaussian Support Vector Machine (SVM) on top of this distance matrix, using LIBSVM [10]. Thus this corresponds to a two-layer kernel machine *Gaussian-Gaussian* involving two Gaussian kernels, with the first Gaussian kernel defined on the low-level features and the second Gaussian kernel defined using the Log-Hilbert–Schmidt distances between the covariance operators of those features. For the sake of comparison, we also evaluated the kernel machine *Gaussian-Laplacian*, with the second kernel being the Laplacian kernel, which corresponds to $p = 1$ in Eq. (5.64).

In comparison, with the covariance matrix representation, one represents the image by the covariance matrix $C_{\mathbf{x}}$ defined directly on the data matrix $\mathbf{x}$ of low-level features, which we call *linearCOV*, since it is precisely the covariance operator obtained with the linear kernel. Given a set of images, we then obtain a set of corresponding covariance matrices and a matrix of pairwise Log-Euclidean distances between these covariance matrices. One can then carry out the task of image classification by applying Gaussian SVM on top of this distance matrix. Thus this corresponds to the two-layer kernel machine *linear-Gaussian*, which is equivalent to the one-layer kernel machine *Gaussian* on top of the Log-Euclidean distances, since the first layer in this case, being linear, has no effect.

Texture classification. The first dataset used is the Kylberg texture dataset [20], which contains 28 texture classes of different natural and man-made surfaces, with each class consisting of 160 images. The experimental protocols are the same as those in [16] and are as follows. Each image is resized to a dimension of 128 × 128, with $m = 1024$ observations computed on a coarse grid (i.e., every 4 pixels in the horizontal and vertical direction). At each pixel, 5 low-level features are extracted: $\mathbf{F}(x, y) = \left[I_{x,y}, |I_x|, \left|I_y\right|, |I_{xx}|, \left|I_{yy}\right|\right]$, where I, I_x, I_y, I_{xx} and I_{yy}, are the intensity, first- and second-order derivatives of the texture image. We randomly selected 5 images in each class for training and used the remaining ones as testing data, repeating the entire procedure 10 times.

Material classification. The second dataset used is the KTH-TIPS2b dataset [9], which contains images of 11 materials captured under 4 different illuminations, in 3 poses, and at 9 scales. The total number of images per class is 108. The same experimental protocols as used for the previous dataset [16] are employed, where at each pixel 23 low-level dense features are extracted: $\mathbf{F}(x, y) =$

Table 5.1 Classification accuracies over the 3 datasets. The accuracies shown are the mean accuracy across all the different splits for each dataset, along with the standard deviation. Here *Log-HS* denotes SVM with the Gaussian kernel on top of the Log-Hilbert–Schmidt distances, *Log-HS*$_\Delta$ denotes SVM with the Laplacian kernel on top of the Log-Hilbert–Schmidt distances, and *Log-E* denotes SVM with the Gaussian kernel on top of the Log-Euclidean distances

	Methods	Kylberg texture	KTH-TIPS2b	KTH-TIPS2b (RGB)	Fish
Gaussian COV	*Log-HS*	**92.58** %(±**1.23**)	**81.91** %(±**3.3**)	**79.94** %(±**4.6**)	**56.74** %(±**2.87**)
	Log-HS$_\Delta$	92.56 %(±1.26)	81.50 %(±3.90)	77.53 %(±5.2)	56.43 %(±3.02)
linear COV	*Log-E*	87.49 %(±1.54)	74.11 %(±7.41)	74.13 %(±6.1)	42.70 %(±3.45)

$\left[R_{x,y}, G_{x,y}, B_{x,y}, \left|G^{0,0}_{x,y}\right|, \ldots, \left|G^{4,5}_{x,y}\right|\right]$, where $R_{x,y}, G_{x,y}, B_{x,y}$ are the color intensities and $\left|G^{o,s}_{x,y}\right|$ are the 20 Gabor filters at 4 orientations and 5 scales. The experiment is repeated across 4 splits of the dataset.

Fish recognition. The third dataset used is the Fish Recognition dataset [8], which consists of 27,370 fish images belonging to 23 different classes. The number of images per class ranges from 21 to 12,112, with a medium resolution of roughly 150 × 120 pixels. The same experimental protocols are employed, where at each pixel the 3 color intensities are extracted: $\mathbf{F}(x, y) = \left[R_{x,y}, G_{x,y}, B_{x,y}\right]$. We randomly selected 5 images from each class for training and 15 for testing, repeating the entire procedure 10 times.

Results and discussion. Table 5.1 shows the classification accuracies obtained on the three tested datasets. As can be seen, the Log-Hilbert–Schmidt distance with the GaussianCOV displays significant improvements over the Log-Euclidean distance with the linearCOV. This strong improvement in performance is as expected, since, as we have discussed previously, covariance operators, by capturing nonlinear correlations between input features, offer a more general, more powerful, and more expressive data representation than covariance matrices.

5.7 Discussion, Conclusion, and Future Work

In this chapter, we have reviewed some of the recent progress in the generalization of the data representation framework using finite-dimensional covariance matrices to infinite-dimensional RKHS covariance operators, which are induced by positive definite kernels on the original data. In particular, we treated covariance operators in the setting of the infinite-dimensional manifold of positive definite operators, which is the generalization of the Riemannian manifold setting for SPD matrices. We presented the affine-invariant and Log-Hilbert–Schmidt distances on this manifold, which are generalizations of the affine-invariant and Log-Euclidean distances, respectively, between SPD matrices. For RKHS covariance operators, these distances admit closed form expressions via the corresponding Gram matrices and thus can be employed directly in a practical algorithm, such as image classification.

The Log-Hilbert–Schmidt distance, in particular, can be used to define new positive definite kernels, giving rise to a two-layer kernel machine. Experiments on the task of image classification have demonstrated that results obtained using the infinite-dimensional covariance operator representation significantly outperform those obtained using the finite-dimensional covariance matrix representation.

There are several ongoing and potential future research directions for the data representation framework using covariance operators. *On the methodological side*, one challenge faced by the framework is that both the affine-invariant and Log-Hilbert–Schmidt distances between covariance operators are computationally intensive on large scale datasets. One way to tackle this computational complexity for large scale applications is by approximating the infinite-dimensional covariance operators using approximate kernel feature maps. This has recently been carried out by [14], which effectively computed an approximate version of the affine-invariant distance, and [29], which computed approximate Log-Hilbert–Schmidt distances, both using Fourier feature maps. It would be interesting and fruitful to explore other approximations and computation schemes as well. *On the application side*, given the numerous applications of covariance matrices in diverse domains, ranging from statistics to machine learning to brain imaging, we expect that the covariance operator framework will find many more applications beyond those that we have presented or surveyed in this chapter.

Appendix

Proofs of Mathematical Results

Proof (**of Proposition** 1) For the first kernel, we have the property that the sum and product of positive definite kernels are also positive definite. Thus from the positivity of the inner product $\langle A, B\rangle_F$, it follows that $K(A, B) = (c + \langle A, B\rangle_{\text{logE}})^d$ is positive definite, as in the Euclidean setting.

For the second kernel, since $(\text{Sym}^{++}(n), \odot, \circledast, \langle\ ,\ \rangle_{\text{logE}})$ is an inner product space, it follows that the kernel

$$K(A, B) = \exp(-d^p_{\text{logE}}(A, B)/\sigma^2) = \exp(-||\log(A) - \log(B)||^p_F/\sigma^2)$$

is positive definite for $0 < p \leq 2$ by a classical result due to Schoenberg on positive definite functions and the imbeddability of metric spaces into Hilbert spaces (see [35], Theorem 1 and Corollary 1).

Proof (**of Lemma** 1) Recall that we have $\Phi(\mathbf{x})^*\Phi(\mathbf{x}) = K[\mathbf{x}]$, $\Phi(\mathbf{y})^*\Phi(\mathbf{y}) = K[\mathbf{y}]$, $\Phi(\mathbf{x})^*\Phi(\mathbf{y}) = K[\mathbf{x}, \mathbf{y}]$. By definition of the Hilbert–Schmidt norm and property of the trace operation, we have

$$
\begin{aligned}
&||C_{\Phi(\mathbf{x})} - C_{\Phi(\mathbf{y})}||^2_{\mathrm{HS}} = \left\| \frac{1}{m}\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^* - \frac{1}{m}\Phi(\mathbf{y})J_m\Phi(\mathbf{y})^* \right\|^2_{\mathrm{HS}} \\
&= \frac{1}{m^2}||\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^*||^2_{\mathrm{HS}} - \frac{2}{m^2}\langle \Phi(\mathbf{x})J_m\Phi(\mathbf{x})^*, \Phi(\mathbf{y})J_m\Phi(\mathbf{y})^*\rangle_{\mathrm{HS}} \\
&+ \frac{1}{m^2}||\Phi(\mathbf{y})J_m\Phi(\mathbf{y})^*||^2_{\mathrm{HS}} \\
&= \frac{1}{m^2}\mathrm{tr}[\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^*\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^*] - \frac{2}{m^2}\mathrm{tr}[\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^*\Phi(\mathbf{y})J_m\Phi(\mathbf{y})^*] \\
&+ \frac{1}{m^2}\mathrm{tr}[\Phi(\mathbf{y})J_m\Phi(\mathbf{y})^*\Phi(\mathbf{y})J_m\Phi(\mathbf{y})^*] \\
&= \frac{1}{m^2}\mathrm{tr}[(K[\mathbf{x}]J_m)^2 - 2K[\mathbf{y},\mathbf{x}]J_mK[\mathbf{x},\mathbf{y}]J_m + (K[\mathbf{y}]J_m)^2] \\
&= \frac{1}{m^2}[\langle J_mK[\mathbf{x}], K[\mathbf{x}]J_m\rangle_F - 2\langle J_mK[\mathbf{x},\mathbf{y}], K[\mathbf{x},\mathbf{y}]J_m\rangle_F + \langle J_mK[\mathbf{y}], K[\mathbf{y}]J_m\rangle_F].
\end{aligned}
$$

This completes the proof of the lemma. □

Lemma 2 *Let B be a constant with $0 < B < 1$. Then for all $|x| \leq B$,*

$$|\log(1+x)| \leq \frac{1}{1-B}|x|. \tag{5.65}$$

Proof For $x \geq 0$, we have the well-known inequality $0 \leq \log(1+x) \leq x$, so clearly $0 \leq \log(1+x) < \frac{1}{1-B}x$. Consider now the case $-B \leq x \leq 0$. Let

$$f(x) = \log(1+x) - \frac{1}{1-B}x.$$

We have

$$f^{'}(x) = \frac{1}{1+x} - \frac{1}{1-B} \leq 0,$$

with $f^{'}(-B) = 0$. Thus the function f is decreasing on $[-B, 0]$ and reaches its minimum at $x = 0$, which is $f(0) = 0$. Hence we have for all $-1 < -B \leq x \leq 0$

$$0 \geq \log(1+x) \geq \frac{1}{1-B}x \Rightarrow |\log(1+x)| \leq \frac{1}{1-B}|x|,$$

as we claimed. □

Proof (**of Propositions 2 and 3**) We first show that for an operator $A + \gamma I > 0$, where A is self-adjoint, compact, the operator $\log(A + \gamma I) \notin \mathrm{HS}(\mathscr{H})$ if $\gamma \neq 1$. Since A is compact, it has a countable spectrum $\{\lambda_k\}_{k\in\mathbb{N}}$, with $\lim_{k\to\infty}\lambda_k = 0$, so that $\lim_{k\to\infty}\log(\lambda_k + \gamma) = \log(\gamma)$. Thus if $\gamma \neq 1$, so that $\log\gamma \neq 0$, we have

$$||\log(A+\gamma I)||^2_{\mathrm{HS}} = \sum_{k=1}^{\infty}[\log(\lambda_k+\gamma)]^2 = \infty.$$

Hence $\log(A+\gamma I) \notin \mathrm{HS}(\mathscr{H})$ if $\gamma \neq 1$.

Assume now that $\gamma = 1$. We show that $\log(A+I) \in \mathrm{HS}(\mathscr{H})$ if and only if $A \in \mathrm{HS}(\mathscr{H})$. For the first direction, assume that $B = \log(A+I) \in \mathrm{HS}(\mathscr{H})$. By definition, we have $A+I = \exp(B) \Longleftrightarrow A = \exp(B) - I = \sum_{k=1}^{\infty}\frac{B^k}{k!}$, with

$$||A||_{\mathrm{HS}} = \left\|\sum_{k=1}^{\infty}\frac{B^k}{k!}\right\|_{\mathrm{HS}} \leq \sum_{k=1}^{\infty}\frac{||B||^k_{\mathrm{HS}}}{k!} = \exp(||B||_{\mathrm{HS}}) - 1 < \infty.$$

This shows that $A \in \mathrm{HS}$. Conversely, assume $A \in \mathrm{HS}(\mathscr{H})$, so that

$$||A||^2_{\mathrm{HS}} = \sum_{k=1}^{\infty}\lambda_k^2 < \infty,$$

and that $A+I > 0$, so that $\log(A+I)$ is well-defined and bounded, with eigenvalues $\{\log(\lambda_k+1)\}_{k=1}^{\infty}$. Since $\lim_{k\to\infty}\lambda_k = 0$, for any constant $0 < \varepsilon < 1$, there exists $N = N(\varepsilon)$ such that $|\lambda_k| < \varepsilon \; \forall k \geq N$. By Lemma 2, we have

$$\begin{aligned}||\log(A+I)||^2_{\mathrm{HS}} &= \sum_{k=1}^{\infty}[\log(\lambda_k+1)]^2 = \sum_{k=1}^{N-1}[\log(\lambda_k+1)]^2 + \sum_{k=N}^{\infty}[\log(\lambda_k+1)]^2 \\ &\leq \sum_{k=1}^{N-1}[\log(\lambda_k+1)]^2 + \frac{1}{1-\varepsilon}\sum_{k=N}^{\infty}\lambda_k^2 < \infty.\end{aligned}$$

This shows that $\log(A+I) \in \mathrm{HS}(\mathscr{H})$, which completes the proof. □

Proof (**Proof of Proposition** 4) Since the identity operator I commutes with any operator A, we have the decomposition

$$\log(A+\gamma I) = \log\left(\frac{A}{\gamma}+I\right) + (\log\gamma)I.$$

We first note that the operator $\log\left(\frac{A}{\gamma}+I\right)$ is compact, since it possesses a countable set of eigenvalues $\{\log(\frac{\lambda_k}{\gamma}+1)\}_{k\in\mathbb{N}}$ satisfying $\lim_{k\to\infty}\log(\frac{\lambda_k}{\gamma}+1) = 0$.

If A is Hilbert–Schmidt, then by Proposition 2, we have $\log\left(\frac{A}{\gamma}+I\right) \in \mathrm{HS}(\mathscr{H})$, and thus $\log(A+\gamma I) \in \mathscr{H}_{\mathbb{R}}$. By definition of the extended Hilbert–Schmidt norm,

$$||\log(A+\gamma I)||^2_{\mathrm{eHS}} = \left\|\log\left(\frac{A}{\gamma}+I\right)\right\|^2_{\mathrm{HS}} + (\log\gamma)^2 < \infty.$$

Conversely, if $\log(A+\gamma I) \in \mathscr{H}_{\mathbb{R}}$, then together with the fact that $\log\left(\frac{A}{\gamma}+I\right)$ is compact, the above decomposition shows that we must have $\log\left(\frac{A}{\gamma}+I\right) \in \mathrm{HS}(\mathscr{H})$ and hence $A \in \mathrm{HS}(\mathscr{H})$ by Proposition 2. □

Proof (**of Proposition 5**) Since $(A+\gamma I) > 0$, $(B+\mu I) > 0$, it is straightforward to see that $(A+\gamma I)^{-1/2}(B+\mu I)(A+\gamma I)^{-1/2} > 0$. Using the identity

$$(A+\gamma I)^{-1} = \frac{1}{\gamma}I - \frac{A}{\gamma}(A+\gamma I)^{-1},$$

we obtain

$$\begin{aligned}&(A+\gamma I)^{-1/2}(B+\mu I)(A+\gamma I)^{-1/2}\\ &= \frac{\mu}{\gamma}I + (A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2} - \frac{\mu}{\gamma}A(A+\gamma I)^{-1} = Z + \nu I,\end{aligned}$$

where $\nu = \frac{\mu}{\gamma}$ and $Z = (A+\gamma I)^{-1/2}B(A+\gamma I)^{-1/2} - \frac{\mu}{\gamma}A(A+\gamma I)^{-1}$. It is clear that $Z = Z^*$ and that $Z \in \mathrm{HS}(\mathscr{H})$, since $\mathrm{HS}(\mathscr{H})$ is a two-sided ideal in $\mathscr{L}(\mathscr{H})$. It follows that $\log(Z+\gamma I) \in \mathscr{H}_{\mathbb{R}}$ by Proposition 4. Thus the geodesic distance

$$\begin{aligned}d_{\mathrm{aiHS}}[(A+\gamma I),(B+\mu I)] &= ||\log[(A+\gamma I)^{-1/2}(B+\mu I)(A+\gamma I)^{-1/2}]||_{\mathrm{eHS}}\\ &= ||\log(Z+\nu I)||_{\mathrm{eHS}}\end{aligned}$$

is always finite. Furthermore, by Proposition 2, $\log(\frac{Z}{\nu}+I) \in \mathrm{HS}(\mathscr{H})$ and thus by definition of the extended Hilbert–Schmidt norm, when $\dim(\mathscr{H}) = \infty$,

$$d^2_{\mathrm{aiHS}}[(A+\gamma I),(B+\mu I)] = ||\log(Z+\nu I)||^2_{\mathrm{eHS}} = ||\log\left(\frac{Z}{\nu}+I\right)||^2_{\mathrm{HS}} + (\log\nu)^2.$$

This completes the proof. □

Proof (**of Proposition 6**) By Proposition 4, $(A+\gamma I), (B+\mu I) \in \Sigma(\mathscr{H}) \Longleftrightarrow \log(A+\gamma I), \log(B+\mu I) \in \mathscr{H}_{\mathbb{R}}$. It follows that $[\log(A+\gamma I) - \log(B+\mu I)] \in \mathscr{H}_{\mathbb{R}}$, so that $||\log(A+\gamma I) - \log(B+\mu I)||_{\mathrm{eHS}}$ is always finite.

Furthermore, by Proposition 2, $\log\left(\frac{A}{\gamma}+I\right)$, $\log\left(\frac{B}{\mu}+I\right) \in \mathrm{HS}(\mathscr{H})$ and by definition of the extended Hilbert–Schmidt norm, when $\dim(\mathscr{H}) = \infty$,

$$\begin{aligned}&||\log(A+\gamma I) - \log(B+\mu I)||^2_{\mathrm{eHS}} = \left\|\log\left(\frac{A}{\gamma}+I\right) - \log\left(\frac{B}{\mu}+I\right) + \left(\log\frac{\gamma}{\mu}\right)I\right\|^2_{\mathrm{eHS}}\\ &= \left\|\log\left(\frac{A}{\gamma}+I\right) - \log\left(\frac{B}{\mu}+I\right)\right\|^2_{\mathrm{HS}} + \left(\log\frac{\gamma}{\mu}\right)^2.\end{aligned}$$

This completes the proof. □

References

1. E. Andruchow, A. Varela, Non positively curved metric in the space of positive definite infinite matrices. Revista de la Union Matematica Argentina **48**(1), 7–15 (2007)
2. V.I. Arsenin, A.N. Tikhonov, *Solutions of Ill-Posed Problems* (Winston, Washington, 1977)
3. V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Fast and simple calculus on tensors in the Log-Euclidean framework, *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005* (Springer, New York, 2005), pp. 115–122
4. V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J. Matrix Anal. Appl. **29**(1), 328–347 (2007)
5. F. Barbaresco, Information geometry of covariance matrix: Cartan-Siegel homogeneous bounded domains, Mostow/Berger fibration and Frechet median, *Matrix Information Geometry* (Springer, New York, 2013), pp. 199–255
6. R. Bhatia, *Positive Definite Matrices* (Princeton University Press, Princeton, 2007)
7. D.A. Bini, B. Iannazzo, Computing the Karcher mean of symmetric positive definite matrices. Linear Algebra Appl. **438**(4), 1700–1710 (2013)
8. B.J. Boom, J. He, S. Palazzo, P.X. Huang, C. Beyan, H.-M. Chou, F.-P. Lin, C. Spampinato, R.B. Fisher, A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. Ecol. Inf. **23**, 83–97 (2014)
9. B. Caputo, E. Hayman, P. Mallikarjuna, Class-specific material categorisation, in *ICCV* (2005), pp. 1597–1604
10. C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27:1–27:27 (2011)
11. A. Cherian, S. Sra, A. Banerjee, N. Papanikolopoulos, Jensen-Bregman LogDet divergence with application to efficient similarity search for covariance matrices. TPAMI **35**(9), 2161–2174 (2013)
12. I.L. Dryden, A. Koloydenko, D. Zhou, Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann. Appl. Stat. **3**, 1102–1123 (2009)
13. H.W. Engl, M. Hanke, A. Neubauer, *Regularization of Inverse Problems*, vol. 375, Mathematics and Its Applications (Springer, New York, 1996)
14. M. Faraki, M. Harandi, F. Porikli, Approximate infinite-dimensional region covariance descriptors for image classification, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015)
15. P. Formont, J.-P. Ovarlez, F. Pascal, On the use of matrix information geometry for polarimetric SAR image classification, *Matrix Information Geometry* (Springer, New York, 2013), pp. 257–276
16. M. Harandi, M. Salzmann, F. Porikli, Bregman divergences for infinite dimensional covariance matrices, in *CVPR* (2014)
17. S. Jayasumana, R. Hartley, M. Salzmann, H. Li, M. Harandi, Kernel methods on the Riemannian manifold of symmetric positive definite matrices, in *CVPR* (2013)
18. S. Jayasumana, R. Hartley, M. Salzmann, H. Li, M. Harandi, Kernel methods on Riemannian manifolds with Gaussian RBF kernels. IEEE Trans. Pattern Anal. Mach. Intell. **37**(12), 2464–2477 (2015)
19. B. Kulis, M.A. Sustik, I.S. Dhillon, Low-rank kernel learning with Bregman matrix divergences. J. Mach. Learn. Res. **10**, 341–376 (2009)
20. G. Kylberg, The Kylberg texture dataset v. 1.0. External report (Blue series) 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University (2011)
21. G. Larotonda, *Geodesic Convexity, Symmetric Spaces and Hilbert-Schmidt Operators*. Ph.D. thesis, Universidad Nacional de General Sarmiento, Buenos Aires, Argentina (2005)
22. G. Larotonda, Nonpositive curvature: a geometrical approach to Hilbert-Schmidt operators. Differ. Geom. Appl. **25**, 679–700 (2007)
23. J.D. Lawson, Y. Lim, The geometric mean, matrices, metrics, and more. Am. Math. Monthly **108**(9), 797–812 (2001)

24. P. Li, Q. Wang, W. Zuo, L. Zhang, Log-Euclidean kernels for sparse representation and dictionary learning, in *ICCV* (2013)
25. H.Q. Minh, Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. Constr. Approx. **32**, 307–338 (2010)
26. H.Q. Minh, Affine-invariant Riemannian distance between infinite-dimensional covariance operators, in *Geometric Science of Information*, vol. 9389, Lecture Notes in Computer Science, ed. by F. Nielsen, F. Barbaresco (Springer International Publishing, Switzerland, 2015), pp. 30–38
27. H.Q. Minh, P. Niyogi, Y. Yao, Mercer's theorem, feature maps, and smoothing, in *Proceedings of 19th Annual Conference on Learning Theory* (Springer, Pittsburg, 2006)
28. H.Q. Minh, M. San Biagio, V. Murino, Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces, in *Advances in Neural Information Processing Systems 27 (NIPS 2014)* (2014), pp. 388–396
29. H.Q. Minh, M. San Biagio, L. Bazzani, V. Murino, Approximate Log-Hilbert-Schmidt distances between covariance operators for image classification, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016)
30. G.D. Mostow, Some new decomposition theorems for semi-simple groups. Mem. Am. Math. Soc. **14**, 31–54 (1955)
31. X. Pennec, P. Fillard, N. Ayache, A Riemannian framework for tensor computing. Int. J. Comput. Vis. **66**(1), 41–66 (2006)
32. D. Pigoli, J. Aston, I.L. Dryden, P. Secchi, Distances and inference for covariance operators. Biometrika **101**(2), 409–422 (2014)
33. F. Porikli, O. Tuzel, P. Meer, Covariance tracking using model update based on Lie algebra, in *CVPR*, vol. 1 (IEEE, 2006), pp. 728–735
34. A. Qiu, A. Lee, M. Tan, M.K. Chung, Manifold learning on brain functional networks in aging. Med. Image Anal. **20**(1), 52–60 (2015)
35. I.J. Schoenberg, Metric spaces and positive definite functions. Trans. Am. Math. Soc. **44**, 522–536 (1938)
36. B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. **10**(5), 1299 (1998)
37. J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis* (Cambridge University Press, Cambridge, 2004)
38. S. Sra, A new metric on the manifold of kernel matrices with application to matrix geometric means. Adv. Neural Inf. Process. Syst. **1**, 144–152 (2012)
39. D. Tosato, M. Spera, M. Cristani, V. Murino, Characterizing humans on Riemannian manifolds. TPAMI **35**(8), 1972–1984 (2013)
40. O. Tuzel, F. Porikli, P. Meer, Pedestrian detection via classification on Riemannian manifolds. TPAMI **30**(10), 1713–1727 (2008)
41. S.K. Zhou, R. Chellappa, From sample similarity to ensemble similarity: probabilistic distance measures in reproducing kernel Hilbert space. TPAMI **28**(6), 917–929 (2006)