Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces

Hà Quang Minh **Marco San Biagio** **Vittorio Murino**
Istituto Italiano di Tecnologia
Via Morego 30, Genova 16163, ITALY
{minh.haquang,marco.sanbiagio,vittorio.murino}@iit.it

Abstract

This paper introduces a novel mathematical and computational framework, namely *Log-Hilbert-Schmidt metric* between positive definite operators on a Hilbert space. This is a generalization of the Log-Euclidean metric on the Riemannian manifold of positive definite matrices to the infinite-dimensional setting. The general framework is applied in particular to compute distances between covariance operators on a Reproducing Kernel Hilbert Space (RKHS), for which we obtain explicit formulas via the corresponding Gram matrices. Empirically, we apply our formulation to the task of multi-category image classification, where each image is represented by an infinite-dimensional RKHS covariance operator. On several challenging datasets, our method significantly outperforms approaches based on covariance matrices computed directly on the original input features, including those using the Log-Euclidean metric, Stein and Jeffreys divergences, achieving new state of the art results.

1 Introduction and motivation

Symmetric Positive Definite (SPD) matrices, in particular covariance matrices, have been playing an increasingly important role in many areas of machine learning, statistics, and computer vision, with applications ranging from kernel learning [12], brain imaging [9], to object detection [24, 23]. One key property of SPD matrices is the following. For a fixed $n \in \mathbb{N}$, the set of all SPD matrices of size $n \times n$ is not a subspace in Euclidean space, but is a Riemannian manifold with nonpositive curvature, denoted by $\mathrm{Sym}^{++}(n)$. As a consequence of this manifold structure, computational methods for $\mathrm{Sym}^{++}(n)$ that simply rely on Euclidean metrics are generally suboptimal.

In the current literature, many methods have been proposed to exploit the non-Euclidean structure of $\mathrm{Sym}^{++}(n)$. For the purposes of the present work, we briefly describe three common approaches here, see e.g. [9] for other methods. The first approach exploits the affine-invariant metric, which is the classical Riemannian metric on $\mathrm{Sym}^{++}(n)$ [18, 16, 3, 19, 4, 24]. The main drawback of this framework is that it tends to be computationally intensive, especially for large scale applications. Overcoming this computational complexity is one of the main motivations for the recent development of the Log-Euclidean metric framework of [2], which has been exploited in many computer vision applications, see e.g. [25, 11, 17]. The third approach defines and exploits Bregman divergences on $\mathrm{Sym}^{++}(n)$, such as Stein and Jeffreys divergences, see e.g. [12, 22, 8], which are not Riemannian metrics but are fast to compute and have been shown to work well on nearest-neighbor retrieval tasks.

While each approach has its advantages and disadvantages, the Log-Euclidean metric possesses several properties which are lacking in the other two approaches. First, it is faster to compute than the affine-invariant metric. Second, unlike the Bregman divergences, it is a Riemannian metric on $\mathrm{Sym}^{++}(n)$ and thus can better capture its manifold structure. Third, in the context of kernel learning, it is straightforward to construct positive definite kernels, such as the Gaussian kernel, using this metric. This is not always the case with the other two approaches: the Gaussian kernel constructed with the Stein divergence, for instance, is only positive definite for certain choices of parameters [22], and the same is true with the affine-invariant metric, as can be numerically verified.

Our contributions: In this work, we generalize the Log-Euclidean metric to the infinite-dimensional setting, both mathematically, computationally, and empirically. Our novel metric, termed *Log-Hilbert-Schmidt metric* (or Log-HS for short), measures the distances between 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 positive definite matrices. These operators have recently been shown to form an infinite-dimensional Riemann-Hilbert manifold by [14, 1, 15], who formulated the infinite-dimensional version of the affine-invariant metric from a purely mathematical viewpoint. While our Log-Hilbert-Schmidt metric framework includes the Log-Euclidean metric as a special case, the infinite-dimensional formulation is significantly different from its corresponding finite-dimensional version, as we demonstrate throughout the paper. In particular, one *cannot* obtain the infinite-dimensional formulas from the finite-dimensional ones by letting the dimension approach infinity.

Computationally, we apply our abstract mathematical framework to compute distances between covariance operators on an RKHS induced by a positive definite kernel. From a kernel learning perspective, this is motivated by the fact that covariance operators defined on nonlinear features, which are obtained by mapping the original data into a high-dimensional feature space, can better capture input correlations than covariance matrices defined on the original data. This is a viewpoint that goes back to KernelPCA [21]. In our setting, we obtain closed form expressions for the Log-Hilbert-Schmidt metric between covariance operators via the Gram matrices.

Empirically, we apply our framework to the task of multi-class image classification. In our approach, the original features extracted from each input image are implicitly mapped into the RKHS induced by a positive definite kernel. The covariance operator defined on the RKHS is then used as the representation for the image and the distance between two images is the Log-Hilbert-Schmidt distance between their corresponding covariance operators. On several challenging datasets, our method significantly outperforms approaches based on covariance matrices computed directly on the original input features, including those using the Log-Euclidean metric, Stein and Jeffreys divergences.

Related work: The approach most closely related to our current work is [26], which computed probabilistic distances in RKHS. This approach has recently been employed by [10] to compute Bregman divergences between RKHS covariance operators. There are two main theoretical issues with the approach in [26, 10]. The *first* issue is that it is assumed *implicitly* that the concepts of *trace* and *determinant* can be extended to any bounded linear operator on an infinite-dimensional Hilbert space $\mathcal{H}$. This is *not* true in general, as the concepts of trace and determinant are only well-defined for certain classes of operators. Many quantities involved in the computation of the Bregman divergences in [10] are in fact infinite when $\dim(\mathcal{H}) = \infty$, which is the case if $\mathcal{H}$ is the Gaussian RKHS, and only cancel each other out in special cases [1]. The *second* issue concerns the use of the Stein divergence by [10] to define the Gaussian kernel, which is not always positive definite, as discussed above. In contrast, the Log-HS metric formulation proposed in this paper is theoretically rigorous and it is straightforward to define many positive definite kernels, including the Gaussian kernel, with this metric. Furthermore, our empirical results consistently outperform those of [10].

Organization: After some background material in Section 2, we describe the manifold of positive definite operators in Section 3. Sections 4 and 5 form the core of the paper, where we develop the general framework for the Log-Hilbert-Schmidt metric together with the explicit formulas for the case of covariance operators on an RKHS. Empirical results for image classification are given in Section 6. The proofs for all mathematical results are given in the Supplementary Material.

2 Background

The Riemannian manifold of positive definite matrices: The manifold structure of $\mathrm{Sym}^{++}(n)$ has been studied extensively, both mathematically and computationally. This study goes as far

[1] We will provide a theoretically rigorous formulation for the Bregman divergences between positive definite operators in a longer version of the present work.

back as [18], for more recent treatments see e.g. [16, 3, 19, 4]. The most commonly encountered Riemannian metric on $\mathrm{Sym}^{++}(n)$ is the **affine-invariant metric**, in which the geodesic distance between two positive definite matrices A and B is given by

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

where $\log$ denotes the matrix logarithm operation and F is an Euclidean norm on the space of symmetric matrices $\mathrm{Sym}(n)$. Following the classical literature, in this work we take F to be the Frobenious norm, which is induced by the standard inner product on $\mathrm{Sym}(n)$. From a practical viewpoint, the metric (1) tends to be computationally intensive, which is one of the main motivations for the **Log-Euclidean metric** of [2], in which the geodesic distance between A and B is given by

$$d_{\mathrm{logE}}(A,B) = ||\log(A) - \log(B)||_F. \tag{2}$$

The main goal of this paper is to generalize the Log-Euclidean metric to what we term the Log-Hilbert-Schmidt metric between positive definite operators on an infinite-dimensional Hilbert space and apply this metric in particular to compute distances between covariance operators on an RKHS.

Covariance operators: Let the input space $\mathcal{X}$ be an arbitrary non-empty set. Let $\mathbf{x} = [x_1, \ldots, x_m]$ be a data matrix sampled from $\mathcal{X}$, where $m \in \mathbb{N}$ is the number of observations. Let K be a positive definite kernel on $\mathcal{X} \times \mathcal{X}$ and $\mathcal{H}_K$ its induced reproducing kernel Hilbert space (RKHS). Let $\Phi : \mathcal{X} \to \mathcal{H}_K$ be the corresponding feature map, which gives the (potentially infinite) mapped data matrix $\Phi(\mathbf{x}) = [\Phi(x_1), \ldots, \Phi(x_m)]$ of size $\dim(\mathcal{H}_K) \times m$ in the feature space $\mathcal{H}_K$. The corresponding covariance operator for $\Phi(\mathbf{x})$ is defined to be

$$C_{\Phi(\mathbf{x})} = \frac{1}{m}\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^T : \mathcal{H}_K \to \mathcal{H}_K, \tag{3}$$

where J_m is the centering matrix, defined by $J_m = I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T$ with $\mathbf{1}_m = (1, \ldots, 1)^T \in \mathbb{R}^m$. The matrix J_m is symmetric, with $\mathrm{rank}(J_m) = m - 1$, and satisfies $J_m^2 = J_m$. The covariance operator $C_{\Phi(\mathbf{x})}$ can be viewed as a (potentially infinite) covariance matrix in the feature space $\mathcal{H}_K$, with rank at most $m - 1$. If $X = \mathbb{R}^n$ and $K(x,y) = \langle x, y\rangle_{\mathbb{R}^n}$, then $C_{\Phi(\mathbf{x})} = C_{\mathbf{x}}$, the standard $n \times n$ covariance matrix encountered in statistics. [2]

Regularization: Generally, covariance matrices may not be full-rank and thus may only be positive semi-definite. In order to apply the theory of $\mathrm{Sym}^{++}(n)$, one needs to consider the regularized version $(C_{\mathbf{x}} + \gamma I_{\mathbb{R}^n})$ for some $\gamma > 0$. In the infinite-dimensional setting, with $\dim(\mathcal{H}_K) = \infty$, $C_{\Phi(\mathbf{x})}$ is always rank-deficient and regularization is always necessary. With $\gamma > 0$, $(C_{\Phi(\mathbf{x})} + \gamma I_{\mathcal{H}_K})$ is strictly positive *and* invertible, both of which are needed to define the Log-Hilbert-Schmidt metric.

3 Positive definite unitized Hilbert-Schmidt operators

Throughout the paper, let $\mathcal{H}$ be a separable Hilbert space of arbitrary dimension. Let $\mathcal{L}(\mathcal{H})$ be the Banach space of bounded linear operators on $\mathcal{H}$ and $\mathrm{Sym}(\mathcal{H})$ be the subspace of self-adjoint operators in $\mathcal{L}(\mathcal{H})$. We first describe in this section the manifold of positive definite unitized Hilbert-Schmidt operators on which the Log-Hilbert-Schmidt metric is defined. This manifold setting is motivated by the following *two crucial differences between the finite and infinite-dimensional cases*.

(A) Positive definite: If $A \in \mathrm{Sym}(\mathcal{H})$ and $\dim(\mathcal{H}) = \infty$, in order for $\log(A)$ to be well-defined and bounded, it is *not* sufficient to require that all eigenvalues of A be strictly positive. Instead, it is necessary to require that all eigenvalues of A be *bounded below* by a positive constant (Section 3.1).

(B) Unitized Hilbert-Schmidt: The infinite-dimensional generalization of the Frobenious norm is the Hilbert-Schmidt norm. However, if $\dim(\mathcal{H}) = \infty$, the identity operator I is not Hilbert-Schmidt and would have infinite distance from any Hilbert-Schmidt operator. To have a satisfactory framework, it is necessary to enlarge the algebra of Hilbert-Schmidt operators to include I (Section 3.2).

These differences between the cases $\dim(\mathcal{H}) = \infty$ and $\dim(\mathcal{H}) < \infty$ are sharp and manifest themselves in the concrete formulas for the Log-Hilbert-Schmidt metric which we obtain in Sections 4.2 and 5. In particular, the formulas for the case $\dim(\mathcal{H}) = \infty$ are *not* obtainable from their corresponding finite-dimensional versions when $\dim(\mathcal{H}) \to \infty$.

[2] One can also define $C_{\Phi(\mathbf{x})} = \frac{1}{m-1}\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^T$. This should not make much practical difference if m is large.

3.1 Positive definite operators

Positive and strictly positive operators: Let us discuss *the first crucial difference* between the finite and infinite-dimensional settings. Recall that an operator $A \in \mathrm{Sym}(\mathcal{H})$ is said to be *positive* if $\langle Ax, x\rangle \geq 0 \; \forall x \in \mathcal{H}$. The eigenvalues of A, if they exist, are all nonnegative. If A is positive and $\langle Ax, x\rangle = 0 \Longleftrightarrow x = 0$, then A is said to be *strictly positive*, and all its eigenvalues are positive. We denote the sets of all positive and strictly positive operators on $\mathcal{H}$, respectively, by $\mathrm{Sym}^{+}(\mathcal{H})$ and $\mathrm{Sym}^{++}(\mathcal{H})$. Let $A \in \mathrm{Sym}^{++}(\mathcal{H})$. Assume that A is compact, then A has a countable spectrum of positive eigenvalues $\{\lambda_k(A)\}_{k=1}^{\dim(\mathcal{H})}$, counting multiplicities, with $\lim_{k\to\infty}\lambda_k(A) = 0$ if $\dim(\mathcal{H}) = \infty$. Let $\{\phi_k(A)\}_{k=1}^{\dim(\mathcal{H})}$ denote the corresponding normalized eigenvectors, then

$$A = \sum_{k=1}^{\dim(\mathcal{H})} \lambda_k(A)\phi_k(A) \otimes \phi_k(A), \tag{4}$$

where $\phi_k(A) \otimes \phi_k(A) : \mathcal{H} \to \mathcal{H}$ is defined by $(\phi_k(A) \otimes \phi_k(A))w = \langle w, \phi_k(A)\rangle\phi_k(A), \quad w \in \mathcal{H}$. The logarithm of A is defined by

$$\log(A) = \sum_{k=1}^{\dim(\mathcal{H})} \log(\lambda_k(A))\phi_k(A) \otimes \phi_k(A). \tag{5}$$

Clearly, $\log(A)$ is bounded if and only if $\dim(\mathcal{H}) < \infty$, since for $\dim(\mathcal{H}) = \infty$, we have $\lim_{k\to\infty}\log(\lambda_k(A)) = -\infty$. Thus, when $\dim(\mathcal{H}) = \infty$, the condition that A be strictly positive is *not* sufficient for $\log(A)$ to be bounded. Instead, the following stronger condition is necessary.

Positive definite operators: A self-adjoint operator $A \in \mathcal{L}(\mathcal{H})$ is said to be *positive definite* (see e.g. [20]) if there exists a constant $M_A > 0$ such that

$$\langle Ax, x\rangle \geq M_A||x||^2 \quad \text{for all } \; x \in \mathcal{H}. \tag{6}$$

The eigenvalues of A, if they exist, are bounded below by M_A. This condition is equivalent to requiring that A be *strictly positive* and *invertible*, with $A^{-1} \in \mathcal{L}(\mathcal{H})$. Clearly, if $\dim(\mathcal{H}) < \infty$, then strict positivity is equivalent to positive definiteness. Let $\mathbb{P}(\mathcal{H})$ denote the open cone of self-adjoint, positive definite, bounded operators on $\mathcal{H}$, that is

$$\mathbb{P}(\mathcal{H}) = \{A \in \mathcal{L}(\mathcal{H}), A^* = A, \exists M_A > 0 \text{ s.t. } \langle Ax, x\rangle \geq M_A||x||^2 \quad \forall x \in \mathcal{H}\}. \tag{7}$$

Throughout the remainder of the paper, we use the following notation: $A > 0 \Longleftrightarrow A \in \mathbb{P}(\mathcal{H})$.

3.2 The Riemann-Hilbert manifold of positive definite unitized Hilbert-Schmidt operators

Let $\mathrm{HS}(\mathcal{H})$ denote the two-sided ideal of Hilbert-Schmidt operators on $\mathcal{H}$ in $\mathcal{L}(\mathcal{H})$, which is a Banach algebra with the Hilbert-Schmidt norm, defined by

$$||A||_{\mathrm{HS}}^2 = \mathrm{tr}(A^*A) = \sum_{k=1}^{\dim(\mathcal{H})} \lambda_k(A^*A). \tag{8}$$

We now discuss *the second crucial difference* between the finite and infinite-dimensional settings. If $\dim(\mathcal{H}) = \infty$, then the identity operator I is not Hilbert-Schmidt, since $||I||_{\mathrm{HS}} = \infty$. Thus, given $\gamma \neq \mu > 0$, we have $||\log(\gamma I) - \log(\mu I)||_{\mathrm{HS}} = |\log(\gamma) - \log(\mu)| \; ||I||_{\mathrm{HS}} = \infty$, that is even the distance between two different multiples of the identity operator is infinite. This problem is resolved by considering the following *extended (or unitized) Hilbert-Schmidt algebra* [14, 1, 15]:

$$\mathcal{H}_{\mathbb{R}} = \{A + \gamma I \; : A^* = A, \; A \in \mathrm{HS}(\mathcal{H}), \gamma \in \mathbb{R}\}. \tag{9}$$

This can be endowed with the extended Hilbert-Schmidt inner product

$$\langle A + \gamma I, B + \mu I\rangle_{\mathrm{eHS}} = \mathrm{tr}(A^*B) + \gamma\mu = \langle A, B\rangle_{\mathrm{HS}} + \gamma\mu, \tag{10}$$

under which the scalar operators are orthogonal to the Hilbert-Schmidt operators. The corresponding extended Hilbert-Schmidt norm is given by

$$||(A + \gamma I)||_{\mathrm{eHS}}^2 = ||A||_{\mathrm{HS}}^2 + \gamma^2, \qquad \text{where } \; A \in \mathrm{HS}(\mathcal{H}). \tag{11}$$

If $\dim(\mathcal{H}) < \infty$, then we set $|| \; ||_{\mathrm{eHS}} = || \; ||_{\mathrm{HS}}$, with $||(A + \gamma I)||_{\mathrm{eHS}} = ||A + \gamma I||_{\mathrm{HS}}$.

Manifold of positive definite unitized Hilbert-Schmidt operators: Define

$$\Sigma(\mathcal{H}) = \mathbb{P}(\mathcal{H}) \cap \mathcal{H}_{\mathbb{R}} = \{A + \gamma I > 0 \; : A^* = A, \; A \in \mathrm{HS}(\mathcal{H}), \gamma \in \mathbb{R}\}. \tag{12}$$

If $(A + \gamma I) \in \Sigma(\mathcal{H})$, then it has a countable spectrum $\{\lambda_k(A) + \gamma\}_{k=1}^{\dim(\mathcal{H})}$ satisfying $\lambda_k + \gamma \geq M_A$ for some constant $M_A > 0$. Thus $(A + \gamma I)^{-1}$ exists and is bounded, and $\log(A + \gamma I)$ as defined by (5) is well-defined and bounded, with $\log(A + \gamma I) \in \mathcal{H}_{\mathbb{R}}$.

The main results of [15] state that when $\dim(\mathcal{H}) = \infty$, $\Sigma(\mathcal{H})$ is an *infinite-dimensional* Riemann-Hilbert manifold and the map $\log : \Sigma(\mathcal{H}) \to \mathcal{H}_{\mathbb{R}}$ and its inverse $\exp : \mathcal{H}_{\mathbb{R}} \to \Sigma(\mathcal{H})$ are diffeomorphisms. The Riemannian distance between two operators $(A+\gamma I), (B+\mu I) \in \Sigma(\mathcal{H})$ is given by

$$d[(A+\gamma I),(B+\mu I)] = ||\log[(A+\gamma I)^{-1/2}(B+\mu I)(A+\gamma I)^{-1/2}]||_{\mathrm{eHS}}. \tag{13}$$

This is the infinite-dimensional version of the affine-invariant metric (1) [3].

4 Log-Hilbert-Schmidt metric

This section defines and develops the Log-Hilbert-Schmidt metric, which is the infinite-dimensional generalization of the Log-Euclidean metric (2). The general formulation presented in this section is then applied to RKHS covariance operators in Section 5.

4.1 The general setting

Consider the following operations on $\Sigma(\mathcal{H})$:

$$(A+\gamma I) \odot (B+\mu I) = \exp(\log(A+\gamma I) + \log(B+\mu I)), \tag{14}$$

$$\lambda \circledast (A+\gamma I) = \exp(\lambda \log(A+\gamma I)) = (A+\gamma I)^{\lambda}, \quad \lambda \in \mathbb{R}. \tag{15}$$

Vector space structure on $\Sigma(\mathcal{H})$: The key property of the operation $\odot$ is that, unlike the usual operator product, it is commutative, making $(\Sigma(\mathcal{H}), \odot)$ an abelian group and $(\Sigma(\mathcal{H}), \odot, \circledast)$ a vector space, which is isomorphic to the vector space $(\mathcal{H}_{\mathbb{R}}, +, \cdot)$, as shown by the following.

Theorem 1. *Under the two operations $\odot$ and $\circledast$, $(\Sigma(\mathcal{H}), \odot, \circledast)$ becomes a vector space, with $\odot$ acting as vector addition and $\circledast$ acting as scalar multiplication. The zero element in $(\Sigma(\mathcal{H}), \odot, \circledast)$ is the identity operator I and the inverse of $(A+\gamma I)$ is $(A+\gamma I)^{-1}$. Furthermore, the map*

$$\psi : (\Sigma(\mathcal{H}), \odot, \circledast) \to (\mathcal{H}_{\mathbb{R}}, +, \cdot) \quad \textit{defined by} \quad \psi(A+\gamma I) = \log(A+\gamma I), \tag{16}$$

is a vector space isomorphism, so that for all $(A+\gamma I), (B+\mu I) \in \Sigma(\mathcal{H})$ and $\lambda \in \mathbb{R}$,

$$\begin{aligned}\psi((A+\gamma I) \odot (B+\mu I)) &= \log(A+\gamma I) + \log(B+\mu I),\\ \psi(\lambda \circledast (A+\gamma I)) &= \lambda \log(A+\gamma I),\end{aligned} \tag{17}$$

where $+$ and $\cdot$ denote the usual operator addition and multiplication operations, respectively.

Metric space structure on $\Sigma(\mathcal{H})$: Motivated by the vector space isomorphism between $(\Sigma(\mathcal{H}), \odot, \circledast)$ and $(\mathcal{H}_{\mathbb{R}}, +, \cdot)$ via the mapping ψ, the following is our generalization of the Log-Euclidean metric to the infinite-dimensional setting.

Definition 1. *The* **Log-Hilbert-Schmidt distance** *between two operators $(A+\gamma I) \in \Sigma(\mathcal{H})$, $(B+\mu I) \in \Sigma(\mathcal{H})$ is defined to be*

$$d_{\mathrm{logHS}}[(A+\gamma I),(B+\mu I)] = \left\|\log[(A+\gamma I) \odot (B+\mu I)^{-1}]\right\|_{\mathrm{eHS}}. \tag{18}$$

Remark 1. *For our purposes in the current work, we focus on the Log-HS metric as defined above based on the one-to-one correspondence between the algebraic structures of $(\Sigma(\mathcal{H}), \odot, \circledast)$ and $(\mathcal{H}_{\mathbb{R}}, +, \cdot)$. An in-depth treatment of the Log-HS metric in connection with the manifold structure of $\Sigma(\mathcal{H})$ will be provided in a longer version of the paper.*

The following theorem shows that the Log-Hilbert-Schmidt distance satisfies all the axioms of a metric, making $(\Sigma(\mathcal{H}), d_{\mathrm{logHS}})$ a metric space. Furthermore, the square Log-Hilbert-Schmidt distance decomposes uniquely into a sum of a square Hilbert-Schmidt norm plus a scalar term.

Theorem 2. *The Log-Hilbert-Schmidt distance as defined in (18) is a metric, making $(\Sigma(\mathcal{H}), d_{\mathrm{logHS}})$ a metric space. Let $(A+\gamma I) \in \Sigma(\mathcal{H})$, $(B+\mu I) \in \Sigma(\mathcal{H})$. If $\dim(\mathcal{H}) = \infty$, then there exist unique operators $A_1, B_1 \in \mathrm{HS}(\mathcal{H}) \cap \mathrm{Sym}(\mathcal{H})$ and scalars $\gamma_1, \mu_1 \in \mathbb{R}$ such that*

$$A+\gamma I = \exp(A_1 + \gamma_1 I), \quad B+\mu I = \exp(B_1 + \mu_1 I), \tag{19}$$

and

$$d^2_{\mathrm{logHS}}[(A+\gamma I),(B+\mu I)] = \|A_1 - B_1\|^2_{\mathrm{HS}} + (\gamma_1 - \mu_1)^2. \tag{20}$$

If $\dim(\mathcal{H}) < \infty$, then (19) and (20) hold with $A_1 = \log(A+\gamma I)$, $B_1 = \log(B+\mu I)$, $\gamma_1 = \mu_1 = 0$.

[3] We give a more detailed discussion of Eqs. (12) and (13) in the Supplementary Material.

Log-Euclidean metric: Theorem 2 states that when $\dim(\mathcal{H}) < \infty$, we have $d_{\mathrm{logHS}}[(A+\gamma I), (B+\mu I)] = d_{\mathrm{logE}}[(A+\gamma I), (B+\mu I)]$. We have thus recovered the Log-Euclidean metric as a special case of our framework.

Hilbert space structure on $(\Sigma(\mathcal{H}), \odot, \circledast)$: Motivated by formula (20), whose right hand side is a square extended Hilbert-Schmidt distance, we now show that $(\Sigma(\mathcal{H}), \odot, \circledast)$ can be endowed with an inner product, under which it becomes a Hilbert space.

Definition 2. *Let $(A+\gamma I), (B+\mu I) \in \Sigma(\mathcal{H})$. Let $A_1, B_1 \in \mathrm{HS}(\mathcal{H}) \cap \mathrm{Sym}(\mathcal{H})$ and $\gamma_1, \mu_1 \in \mathbb{R}$ be the unique operators and scalars, respectively, such that $A + \gamma I = \exp(A_1 + \gamma_1 I)$ and $B + \mu I = \exp(B_1 + \mu_1 I)$, as in Theorem 2. The* **Log-Hilbert-Schmidt inner product** *between $(A+\gamma I)$ and $(B+\mu I)$ is defined by*

$$\langle A+\gamma I, B+\mu I\rangle_{\mathrm{logHS}} = \langle \log(A+\gamma I), \log(B+\mu I)\rangle_{\mathrm{eHS}} = \langle A_1, B_1\rangle_{\mathrm{HS}} + \gamma_1\mu_1. \tag{21}$$

Theorem 3. *The inner product $\langle\ ,\ \rangle_{\mathrm{logHS}}$ as given in (21) is well-defined on $(\Sigma(\mathcal{H}), \odot, \circledast)$. Endowed with this inner product, $(\Sigma(\mathcal{H}), \odot, \circledast, \langle\ ,\ \rangle_{\mathrm{logHS}})$ becomes a Hilbert space. The corresponding* **Log-Hilbert-Schmidt norm** *is given by*

$$||A+\gamma I||^2_{\mathrm{logHS}} = ||\log(A+\gamma I)||^2_{\mathrm{eHS}} = ||A_1||^2_{\mathrm{HS}} + \gamma_1^2. \tag{22}$$

In terms of this norm, the Log-Hilbert-Schmidt distance is given by

$$d_{\mathrm{logHS}}[(A+\gamma I), (B+\mu I)] = \left\|(A+\gamma I)\odot(B+\mu I)^{-1}\right\|_{\mathrm{logHS}}. \tag{23}$$

Positive definite kernels defined with the Log-Hilbert-Schmidt metric: An important consequence of the Hilbert space structure of $(\Sigma(\mathcal{H}), \odot, \circledast, \langle\ ,\ \rangle_{\mathrm{logHS}})$ is that it is straightforward to generalize many positive definite kernels on Euclidean space to $\Sigma(\mathcal{H}) \times \Sigma(\mathcal{H})$.

Corollary 1. *The following kernels defined on $\Sigma(\mathcal{H}) \times \Sigma(\mathcal{H})$ are positive definite:*

$$K[(A+\gamma I), (B+\mu I)] = (c + \langle A+\gamma I, B+\mu I\rangle_{\mathrm{logHS}})^d, \quad c > 0, \quad d \in \mathbb{N}, \tag{24}$$

$$K[(A+\gamma I), (B+\mu I)] = \exp(-d^p_{\mathrm{logHS}}[(A+\gamma I), (B+\mu I)]/\sigma^2), \quad 0 < p \leq 2. \tag{25}$$

4.2 Log-Hilbert-Schmidt metric between regularized positive operators

For our purposes in the present work, we focus on the following subset of $\Sigma(\mathcal{H})$:

$$\Sigma^+(\mathcal{H}) = \{A + \gamma I\ : A \in \mathrm{HS}(\mathcal{H}) \cap \mathrm{Sym}^+(\mathcal{H})\ ,\ \gamma > 0\} \subset \Sigma(\mathcal{H}). \tag{26}$$

Examples of operators in $\Sigma^+(\mathcal{H})$ are the regularized covariance operators $(C_{\Phi(\mathbf{x})} + \gamma I)$ with $\gamma > 0$. In this case the formulas in Theorems 2 and 3 have the following concrete forms.

Theorem 4. *Assume that $\dim(\mathcal{H}) = \infty$. Let $A, B \in \mathrm{HS}(\mathcal{H}) \cap \mathrm{Sym}^+(\mathcal{H})$. Let $\gamma, \mu > 0$. Then*

$$d^2_{\mathrm{logHS}}[(A+\gamma I), (B+\mu I)] = ||\log(\frac{1}{\gamma}A + I) - \log(\frac{1}{\mu}B + I)||^2_{\mathrm{HS}} + (\log\gamma - \log\mu)^2. \tag{27}$$

Their Log-Hilbert-Schmidt inner product is given by

$$\langle (A+\gamma I), (B+\mu I)\rangle_{\mathrm{logHS}} = \langle \log(\frac{1}{\gamma}A + I), \log(\frac{1}{\mu}B + I)\rangle_{\mathrm{HS}} + (\log\gamma)(\log\mu). \tag{28}$$

Finite dimensional case: As a consequence of the differences between the cases $\dim(\mathcal{H}) < \infty$ and $\dim(\mathcal{H}) = \infty$, we have different formulas for the case $\dim(\mathcal{H}) < \infty$, which depend on $\dim(\mathcal{H})$ and which are surprisingly more complicated than in the case $\dim(\mathcal{H}) = \infty$.

Theorem 5. *Assume that $\dim(\mathcal{H}) < \infty$. Let $A, B \in \mathrm{Sym}^+(\mathcal{H})$. Let $\gamma, \mu > 0$. Then*

$$d^2_{\mathrm{logHS}}[(A+\gamma I), (B+\mu I)] = ||\log(\frac{A}{\gamma} + I) - \log(\frac{B}{\mu} + I)||^2_{\mathrm{HS}}$$
$$+2(\log\gamma - \log\mu)\mathrm{tr}[\log(\frac{A}{\gamma} + I) - \log(\frac{B}{\mu} + I)] + (\log\gamma - \log\mu)^2 \dim(\mathcal{H}). \tag{29}$$

The Log-Hilbert-Schmidt inner product between $(A+\gamma I)$ and $(B+\mu I)$ is given by

$$\langle (A+\gamma I), (B+\mu I)\rangle_{\mathrm{logHS}} = \langle \log(\frac{A}{\gamma} + I), \log(\frac{B}{\mu} + I)\rangle_{\mathrm{HS}}$$
$$+(\log\gamma)\mathrm{tr}[\log(\frac{B}{\mu} + I)] + (\log\mu)\mathrm{tr}[\log(\frac{A}{\gamma} + I)] + (\log\gamma\log\mu)\dim(\mathcal{H}). \tag{30}$$

5 Log-Hilbert-Schmidt metric between regularized covariance operators

Let $\mathcal{X}$ be an arbitrary non-empty set. In this section, we apply the general results of Section 4 to compute the Log-Hilbert-Schmidt distance between covariance operators on an RKHS induced by a positive definite kernel K on $\mathcal{X} \times \mathcal{X}$. In this case, we have explicit formulas for d_{logHS} and the inner product $\langle \ , \ \rangle_{\mathrm{logHS}}$ via the corresponding Gram matrices. Let $\mathbf{x} = [x_i]_{i=1}^m$, $\mathbf{y} = [y_i]_{i=1}^m$, $m \in \mathbb{N}$, be two data matrices sampled from $\mathcal{X}$ and $C_{\Phi(\mathbf{x})}$, $C_{\Phi(\mathbf{y})}$ be the corresponding covariance operators induced by the kernel K, as defined in Section 2. Let $K[\mathbf{x}]$, $K[\mathbf{y}]$, and $K[\mathbf{x},\mathbf{y}]$ be the $m \times m$ Gram matrices defined by $(K[\mathbf{x}])_{ij} = K(x_i, x_j)$, $(K[\mathbf{y}])_{ij} = K(y_i, y_j)$, $(K[\mathbf{x},\mathbf{y}])_{ij} = K(x_i, y_j)$, $1 \leq i, j \leq m$. Let $A = \frac{1}{\sqrt{\gamma m}}\Phi(\mathbf{x})J_m : \mathbb{R}^m \to \mathcal{H}_K$, $B = \frac{1}{\sqrt{\mu m}}\Phi(\mathbf{y})J_m : \mathbb{R}^m \to \mathcal{H}_K$, so that

$$A^T A = \frac{1}{\gamma m} J_m K[\mathbf{x}] J_m, \quad B^T B = \frac{1}{\mu m} J_m K[\mathbf{y}] J_m, \quad A^T B = \frac{1}{\sqrt{\gamma \mu} m} J_m K[\mathbf{x},\mathbf{y}] J_m. \tag{31}$$

Let N_A and N_B be the numbers of nonzero eigenvalues of $A^T A$ and $B^T B$, 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

$$\frac{1}{\gamma m} J_m K[\mathbf{x}] J_m = U_A \Sigma_A U_A^T, \quad \frac{1}{\mu m} J_m K[\mathbf{y}] J_m = U_B \Sigma_B U_B^T. \tag{32}$$

In the following, let $\circ$ denote the Hadamard (element-wise) matrix product. Define

$$C_{AB} = \mathbf{1}_{N_A}^T \log(I_{N_A} + \Sigma_A)\Sigma_A^{-1}(U_A^T A^T B U_B \circ U_A^T A^T B U_B)\Sigma_B^{-1}\log(I_{N_B} + \Sigma_B)\mathbf{1}_{N_B}. \tag{33}$$

Theorem 6. *Assume that* $\dim(\mathcal{H}_K) = \infty$. *Let* $\gamma > 0$, $\mu > 0$. *Then*

$$d^2_{\mathrm{logHS}}[(C_{\Phi(\mathbf{x})} + \gamma I), (C_{\Phi(\mathbf{y})} + \mu I)] = \mathrm{tr}[\log(I_{N_A} + \Sigma_A)]^2 + \mathrm{tr}[\log(I_{N_B} + \Sigma_B)]^2 \\ -2C_{AB} + (\log\gamma - \log\mu)^2. \tag{34}$$

The Log-Hilbert-Schmidt inner product between $(C_{\Phi(\mathbf{x})} + \gamma I)$ *and* $(C_{\Phi(\mathbf{y})} + \mu I)$ *is*

$$\langle (C_{\Phi(\mathbf{x})} + \gamma I), (C_{\Phi(\mathbf{y})} + \mu I)\rangle_{\mathrm{logHS}} = C_{AB} + (\log\gamma)(\log\mu). \tag{35}$$

Theorem 7. *Assume that* $\dim(\mathcal{H}_K) < \infty$. *Let* $\gamma > 0$, $\mu > 0$. *Then*

$$d^2_{\mathrm{logHS}}[(C_{\Phi(\mathbf{x})} + \gamma I), (C_{\Phi(\mathbf{y})} + \mu I)] = \mathrm{tr}[\log(I_{N_A} + \Sigma_A)]^2 + \mathrm{tr}[\log(I_{N_B} + \Sigma_B)]^2 - 2C_{AB} \\ +2(\log\frac{\gamma}{\mu})(\mathrm{tr}[\log(I_{N_A} + \Sigma_A)] - \mathrm{tr}[\log(I_{N_B} + \Sigma_B)]) + (\log\frac{\gamma}{\mu})^2 \dim(\mathcal{H}_K). \tag{36}$$

The Log-Hilbert-Schmidt inner product between $(C_{\Phi(\mathbf{x})} + \gamma I)$ *and* $(C_{\Phi(\mathbf{y})} + \mu I)$ *is*

$$\langle (C_{\Phi(\mathbf{x})} + \gamma I), (C_{\Phi(\mathbf{y})} + \mu I)\rangle_{\mathrm{logHS}} = C_{AB} + (\log\mu)\mathrm{tr}[\log(I_{N_A} + \Sigma_A)] \\ +(\log\gamma)\mathrm{tr}[\log(I_{N_B} + \Sigma_B)] + (\log\gamma\log\mu)\dim(\mathcal{H}_K). \tag{37}$$

6 Experimental results

This section demonstrates the empirical performance of the Log-HS metric on the task of multi-category image classification. For each input image, the original features extracted from the image are implicitly mapped into the infinite-dimensional RKHS induced by the Gaussian kernel. The covariance operator defined on the RKHS is called the GaussianCOV and is used as the representation for the image. In a classification algorithm, the distance between two images is the Log-HS distance between their corresponding GaussianCOVs. This is compared with the directCOV representation, that is covariance matrices defined using the original input features. In all of the experiments, we employed LIBSVM [7] as the classification method. The following algorithms were evaluated in our experiments: *Log-E* (directCOV and Gaussian SVM using the Log-Euclidean metric), *Log-HS* (GaussianCOV and Gaussian SVM using the Log-HS metric), *Log-HS*$_\Delta$ (GaussianCOV and SVM with the Laplacian kernel $K(x, y) = \exp(-\frac{||x-y||}{\sigma})$). For all experiments, the kernel parameters were chosen by cross validation, while the regularization parameters were fixed to be $\gamma = \mu = 10^{-8}$. We also compare with empirical results by the different algorithms in [10], namely J-SVM and S-SVM (SVM with the Jeffreys and Stein divergences between directCOVs, respectively), $J_{\mathcal{H}}$-SVM and $S_{\mathcal{H}}$-SVM (SVM with the Jeffreys and Stein divergences between GaussianCOVs, respectively), and results of the Covariance Discriminant Learning (CDL) technique of [25], which can be considered as the state-of-the-art for COV-based classification. All results are reported in Table1.

Table 1: Results over all the datasets

	Methods	Kylberg texture	KTH-TIPS2b	KTH-TIPS2b (RGB)	Fish
GaussianCOV	$Log\text{-}HS$	$\mathbf{92.58}\%(\pm\mathbf{1.23})$	$\mathbf{81.91}\%(\pm\mathbf{3.3})$	$\mathbf{79.94}\%(\pm\mathbf{4.6})$	$\mathbf{56.74}\%(\pm\mathbf{2.87})$
	$Log\text{-}HS_{\triangle}$	$92.56\%(\pm1.26)$	$81.50\%(\pm3.90)$	$77.53\%(\pm5.2)$	$56.43\%(\pm3.02)$
	$S_{\mathcal{H}}$-**SVM**[10]	$91.36\%(\pm1.27)$	$80.10\%(\pm4.60)$	-	-
	$J_{\mathcal{H}}$-**SVM**[10]	$91.25\%(\pm1.33)$	$79.90\%(\pm3.80)$	-	-
directCOV	$Log\text{-}E$	$87.49\%(\pm1.54)$	$74.11\%(\pm7.41)$	$74.13\%(\pm6.1)$	$42.70\%(\pm3.45)$
	S-**SVM**[10]	$81.27\%(\pm1.07)$	$78.30\%(\pm4.84)$	-	-
	J-**SVM**[10]	$82.19\%(\pm1.30)$	$74.70\%(\pm2.81)$	-	-
	CDL [25]	$79.87\%(\pm1.06)$	$76.30\%(\pm5.10)$	-	-

Texture classification: For this task, we used the Kylberg texture dataset [13], which contains 28 texture classes of different natural and man-made surfaces, with each class consisting of 160 images. For this dataset, we followed the validation protocol of [10], where 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 point, we extracted a set of $n = 5$ low-level features $\mathbf{F}(x, y) = [I_{x,y}, |I_x|, |I_y|, |I_{xx}|, |I_{yy}|]$, 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 test data, repeating the entire procedure 10 times. We report the mean and the standard deviation values for the classification accuracies for the different experiments over all 10 random training/testing splits.

Material classification: For this task, we used the KTH-TIPS2b dataset [6], 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. We applied the same protocol as used for the previous dataset [10], extracting 23 low-level dense features: $\mathbf{F}(x, y) = \left[R_{x,y}, G_{x,y}, B_{x,y}, \left|G^{0,0}_{x,y}\right|, \ldots \left|G^{4,5}_{x,y}\right|\right]$, 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. We report the mean and the standard deviation values for all the 4 splits of the dataset.

Fish recognition: The third dataset used is the Fish Recognition dataset [5]. The fish data are acquired from a live video dataset resulting in 27370 verified fish images. The whole dataset is divided into 23 classes. The number of images per class ranges from 21 to 12112, with a medium resolution of roughly 150×120 pixels. The significant variations in color, pose and illumination inside each class make this dataset very challenging. We apply the same protocol as used for the previous datasets, extracting the 3 color intensities from each image to show the effectiveness of our method: $\mathbf{F}(x, y) = [R_{x,y}, G_{x,y}, B_{x,y}]$. We randomly selected 5 images from each class for training and 15 for testing, repeating the entire procedure 10 times.

Discussion of results: As one can observe in Table1, in all of the datasets, the Log-HS framework, operating on GaussianCOVs, significantly outperforms approaches based on directCOVs computed using the original input features, including those using Log-Euclidean, Stein and Jeffreys divergences. Across all datasets, our improvement over the Log-Euclidean metric is up to 14% in accuracy. This is consistent with kernel-based learning theory, because GaussianCOVs, defined on the infinite-dimensional RKHS, can better capture nonlinear input correlations than directCOVs, as we expected. To the best of our knowledge, our results in the Texture and Material classification experiments are the new state of the art results for these datasets. Furthermore, our results, which are obtained using a theoretically rigorous framework, also consistently outperform those of [10]. *The computational complexity of our framework, its two-layer kernel machine interpretation, and other discussions are given in the Supplementary Material.*

Conclusion and future work

We have presented a novel mathematical and computational framework, namely Log-Hilbert-Schmidt metric, that generalizes the Log-Euclidean metric between SPD matrices to the infinite-dimensional setting. Empirically, on the task of image classification, where each image is represented by an infinite-dimensional RKHS covariance operator, the Log-HS framework substantially outperforms other approaches based on covariance matrices computed directly on the original input features. Given the widespread use of covariance matrices, we believe that the Log-HS framework can be potentially useful for many problems in machine learning, computer vision, and other applications. Many more properties of the Log-HS metric, along with further applications, will be reported in a longer version of the current paper and in future work.

References

[1] E. Andruchow and 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. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. on Matrix An. and App.*, 29(1):328–347, 2007.

[3] R. Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

[4] D. A. Bini and B. Iannazzo. Computing the Karcher mean of symmetric positive definite matrices. *Linear Algebra and its Applications*, 438(4):1700–1710, 2013.

[5] B. J. Boom, J. He, S. Palazzo, P. X. Huang, C. Beyan, H.-M. Chou, F.-P. Lin, C. Spampinato, and R. B. Fisher. A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. *Ecological Informatics*, in press, 2013.

[6] B. Caputo, E. Hayman, and P. Mallikarjuna. Class-specific material categorisation. In *ICCV*, pages 1597–1604, 2005.

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[8] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos. Jensen-Bregman LogDet divergence with application to efficient similarity search for covariance matrices. *TPAMI*, 35(9):2161–2174, 2013.

[9] I.L. Dryden, A. Koloydenko, and D. Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics*, 3:1102–1123, 2009.

[10] M. Harandi, M. Salzmann, and F. Porikli. Bregman divergences for infinite dimensional covariance matrices. In *CVPR*, 2014.

[11] S. Jayasumana, R. Hartley, M. Salzmann, Hongdong Li, and M. Harandi. Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In *CVPR*, 2013.

[12] B. Kulis, M. A. Sustik, and I. S. Dhillon. Low-rank kernel learning with Bregman matrix divergences. *The Journal of Machine Learning Research*, 10:341–376, 2009.

[13] 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.

[14] G. Larotonda. *Geodesic Convexity, Symmetric Spaces and Hilbert-Schmidt Operators*. PhD thesis, Universidad Nacional de General Sarmiento, Buenos Aires, Argentina, 2005.

[15] G. Larotonda. Nonpositive curvature: A geometrical approach to Hilbert–Schmidt operators. *Differential Geometry and its Applications*, 25:679–700, 2007.

[16] J. D. Lawson and Y. Lim. The geometric mean, matrices, metrics, and more. *The American Mathematical Monthly*, 108(9):797–812, 2001.

[17] P. Li, Q. Wang, W. Zuo, and L. Zhang. Log-Euclidean kernels for sparse representation and dictionary learning. In *ICCV*, 2013.

[18] G.D. Mostow. Some new decomposition theorems for semi-simple groups. *Memoirs of the American Mathematical Society*, 14:31–54, 1955.

[19] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[20] W.V. Petryshyn. Direct and iterative methods for the solution of linear operator equations in Hilbert spaces. *Transactions of the American Mathematical Society*, 105:136–175, 1962.

[21] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5), July 1998.

[22] S. Sra. A new metric on the manifold of kernel matrices with application to matrix geometric means. In *NIPS*, 2012.

[23] D. Tosato, M. Spera, M. Cristani, and V. Murino. Characterizing humans on Riemannian manifolds. *TPAMI*, 35(8):1972–1984, Aug 2013.

[24] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on Riemannian manifolds. *TPAMI*, 30(10):1713–1727, 2008.

[25] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503, 2012.

[26] S. K. Zhou and R. Chellappa. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space. *TPAMI*, 28(6):917–929, 2006.