Log-Hilbert-Schmidt distance between covariance operators and its approximation

Hà Quang Minh
Pattern Analysis and Computer Vision
Istituto Italiano di Tecnologia (IIT)
Via Morego 30, Genova, Italy
e-mail: minh.haquang@iit.it

One of the most commonly used Riemannian metrics on the set of symmetric, positive definite (SPD) matrices is the Log-Euclidean metric [1]. In this metric, the geodesic distance between two SPD matrices A and B is given by

$$d_{\mathrm{logE}}(A,B) = ||\log(A) - \log(B)||_F, \tag{1}$$

where log denotes the matrix principal logarithm.

Log-Hilbert-Schmidt distance. The generalization of the Log-Euclidean metric to the infinite-dimensional manifold $\Sigma(\mathcal{H})$ of positive definite Hilbert-Schmidt operators on a Hilbert space $\mathcal{H}$ has recently been given by [2]. In this metric, termed *Log-Hilbert-Schmidt (Log-HS) metric*, the distance between two positive definite Hilbert-Schmidt operators $A+\gamma I > 0$ and $B+\mu I > 0$, $A, B \in \mathrm{HS}(\mathcal{H})$, $\gamma, \mu > 0$, is given by

$$d_{\mathrm{logHS}}[(A+\gamma I),(B+\mu I)] = ||\log(A+\gamma I) - \log(B+\mu I)||_{\mathrm{eHS}}, \tag{2}$$

with the *extended Hilbert-Schmidt* norm defined by $||A+\gamma I||^2_{\mathrm{eHS}} = ||A||^2_{\mathrm{HS}} + \gamma^2$.

RKHS covariance operators. As examples of positive Hilbert-Schmidt operators, consider covariance operators in reproducing kernel Hilbert spaces (RKHS), which play an important role in machine learning and statistics. Let $\mathcal{X}$ be any nonempty set. Let K be a positive definite kernel on $\mathcal{X}\times\mathcal{X}$ and $\mathcal{H}_K$ its induced RKHS. Let $\mathcal{H}$ be any Hilbert feature space for K, assumed to be separable, which we identify with $\mathcal{H}_K$, with the corresponding feature map $\Phi : \mathcal{X} \to \mathcal{H}$, so that $K(x,y) = \langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}}$ $\forall (x,y) \in \mathcal{X}\times\mathcal{X}$. Let $\mathbf{x} = [x_1, \ldots, x_m]$ be a data matrix randomly sampled from $\mathcal{X}$ according to some probability distribution. The feature map Φ gives the (potentially infinite) data matrix $\Phi(\mathbf{x}) = [\Phi(x_1), \ldots, \Phi(x_m)]$ in $\mathcal{H}$. Formally, $\Phi(\mathbf{x})$ is a bounded linear operator $\Phi(\mathbf{x}) : \mathbb{R}^m \to \mathcal{H}$, defined by $\Phi(\mathbf{x})\mathbf{b} = \sum_{j=1}^m b_j \Phi(x_j)$, $\mathbf{b} \in \mathbb{R}^m$. The covariance operator for $\Phi(\mathbf{x})$ is defined by

$$C_{\Phi(\mathbf{x})} = \frac{1}{m}\Phi(\mathbf{x}) J_m \Phi(\mathbf{x})^* : \mathcal{H} \to \mathcal{H}, \quad J_m = I_m - \frac{1}{m}\mathbf{1}_m \mathbf{1}_m^T. \tag{3}$$

For $\gamma > 0, \mu > 0$, the Log-HS distance $d_{\mathrm{logHS}}[(C_{\Phi(\mathbf{x})} + \gamma I_{\mathcal{H}}), (C_{\Phi(\mathbf{y})} + \mu I_{\mathcal{H}})]$ between two regularized covariance operators $(C_{\Phi(\mathbf{x})} + \gamma I_{\mathcal{H}})$ and $(C_{\Phi(\mathbf{y})} + \mu I_{\mathcal{H}})$

$$d_{\mathrm{logHS}} = ||\log(C_{\Phi(\mathbf{x})} + \gamma I_{\mathcal{H}}) - \log(C_{\Phi(\mathbf{y})} + \mu I_{\mathcal{H}})||_{\mathrm{eHS}} \tag{4}$$

has a closed form in terms of the corresponding Gram matrices [2]. This distance is generally computationally intensive for large m, however.

Approximation by finite-dimensional Log-Euclidean distances. To reduce the computational cost, we consider computing an *explicit approximate feature map* $\hat{\Phi}_D : \mathcal{X} \to \mathbb{R}^D$, where D is finite and $D << \dim(\mathcal{H})$, so that

$$\langle \hat{\Phi}_D(x), \hat{\Phi}_D(y) \rangle_{\mathbb{R}^D} = \hat{K}_D(x,y) \approx K(x,y), \quad \text{with} \lim_{D \to \infty} \hat{K}_D(x,y) = K(x,y), \tag{5}$$

$\forall (x,y) \in \mathcal{X} \times \mathcal{X}$. With the approximate feature map $\hat{\Phi}_D$, we have the matrix $\hat{\Phi}_D(\mathbf{x}) = [\hat{\Phi}_D(x_1), \ldots, \hat{\Phi}_D(x_m)] \in \mathbb{R}^{D \times m}$ and the approximate covariance operator

$$C_{\hat{\Phi}_D(\mathbf{x})} = \frac{1}{m} \hat{\Phi}_D(\mathbf{x}) J_m \hat{\Phi}_D(\mathbf{x})^T : \mathbb{R}^D \to \mathbb{R}^D. \tag{6}$$

We then consider the following as an approximate version of the Log-HS distance given in Formula (4):

$$\left\| \log\left(C_{\hat{\Phi}_D(\mathbf{x})} + \gamma I_D\right) - \log\left(C_{\hat{\Phi}_D(\mathbf{y})} + \mu I_D\right) \right\|_F. \tag{7}$$

Key theoretical question. We need to determine whether Formula (7) is truly a finite-dimensional approximation of Formula (4), in the sense that

$$\lim_{D \to \infty} \left\| \log(C_{\hat{\Phi}_D(\mathbf{x})} + \gamma I_D) - \log(C_{\hat{\Phi}_D(\mathbf{y})} + \mu I_D) \right\|_F$$
$$= ||\log(C_{\Phi(\mathbf{x})} + \gamma I_{\mathcal{H}}) - \log(C_{\Phi(\mathbf{y})} + \mu I_{\mathcal{H}})||_{\mathrm{eHS}}. \tag{8}$$

The following results shows that in general, this is *not* possible.

Theorem 1. *Assume that* $\gamma \neq \mu$, $\gamma > 0$, $\mu > 0$. *Then*

$$\lim_{D \to \infty} \left\| \log(C_{\hat{\Phi}_D(\mathbf{x})} + \gamma I_D) - \log(C_{\hat{\Phi}_D(\mathbf{y})} + \mu I_D) \right\|_F = \infty.$$

In practice, however, it is reasonable to assume that we can use the same regularization parameter for both $C_{\hat{\Phi}_D(\mathbf{x})}$ and $C_{\hat{\Phi}_D(\mathbf{y})}$, that is to set $\gamma = \mu$. In this setting, we obtain the necessary convergence, as follows.

Theorem 2. *Assume that* $\gamma = \mu > 0$. *Then*

$$\lim_{D \to \infty} \left\| \log(C_{\hat{\Phi}_D(\mathbf{x})} + \gamma I_D) - \log(C_{\hat{\Phi}_D(\mathbf{y})} + \gamma I_D) \right\|_F$$
$$= ||\log(C_{\Phi(\mathbf{x})} + \gamma I_{\mathcal{H}}) - \log(C_{\Phi(\mathbf{y})} + \gamma I_{\mathcal{H}})||_{\mathrm{eHS}}. \tag{9}$$

References

[1] Arsigny, V., Fillard, P., Pennec, X. and Ayache, N., 2007. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1), pp.328-347.

[2] Minh, H.Q., San Biagio, M. and Murino, V., 2014. Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces. In *Advances in Neural Information Processing Systems* (pp. 388-396).