Discriminant Analysis on Riemannian Manifold of Gaussian Distributions for Face Recognition with Image Sets
Supplementary Material

Wen Wang[1,2], Ruiping Wang[1], Zhiwu Huang[1,2], Shiguang Shan[1], Xilin Chen[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
{wen.wang, zhiwu.huang}@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

Here we give a proof that the defined kernels for Gaussian distributions in Sec.3 are valid kernels and analyze their positive definiteness. In the following, we start with reviewing some important definitions and theorems.

1. Background Theory

First, we give the definition of a positive definite (*pd*) kernel.

Definition 1. *Let $\mathcal{X}$ be a nonempty set. A function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a positive definite (**pd**) kernel on $\mathcal{X}$ if and only if K is symmetric and*

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

for any $n \in \mathbb{N}$, $c_1, ..., c_n \in \mathbb{R}$ and $x_1, ..., x_n \in \mathcal{X}$.

Then we define a related family of kernels which is called negative definite (*nd*) kernel.

Definition 2. *Let $\mathcal{X}$ be a nonempty set. A function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a negative definite (**nd**) kernel on $\mathcal{X}$ if and only if K is symmetric and*

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \leq 0, \tag{2}$$

for any $n \in \mathbb{N}$, $c_1, ..., c_n \in \mathbb{R}$ with $\sum_{i=1}^{n} c_i = 0$ and $x_1, ..., x_n \in \mathcal{X}$.

We next recall from [3] and state a theorem that gives a necessary and sufficient condition for obtaining a *pd* kernel from a distance function.

Theorem 1. *Let $\mathcal{X}$ be a nonempty set and $f : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a symmetric function. Then the kernel function $\exp(-tf(x_i, x_j))$ is **pd** for all $t > 0$ if and only if f is **nd**.*

From Theorem 1, Jayasumana *et al*. proved the following theorem in [1].

Theorem 2. *Let $\mathcal{X}$ be a nonempty set, $\mathcal{V}$ be an inner product space, and $\psi : \mathcal{X} \mapsto \mathcal{V}$ be a function. Then $f : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ defined by $f(x_i, x_j) = \|\psi(x_i) - \psi(x_j)\|_{\mathcal{V}}^2$ is **nd**.*

2. Positive definiteness of the kernels for Gaussians

Because *pd* kernels can define valid Reproducing Kernel Hilbert Space (RKHS) and further allow the kernel methods in Euclidean space to be generalized to nonlinear manifolds, in this section we try to give a rigorous proof that the proposed probabilistic kernels for Gaussian distributions are *pd*.

2.1. Kullback-Leibler Kernel

Formally, given continuous probability distributions P and Q, their Kullback-Leibler Divergence (KLD) is formulated by

$$KLD(P\|Q) = \int P(x) \ln \frac{P(x)}{Q(x)} dx. \tag{3}$$

Kullback-Leibler kernel is defined as follows:

$$K_{KLD}(P, Q) = \exp(-\frac{KLD(P\|Q) + KLD(Q\|P)}{2t^2}), \tag{4}$$

where t is the kernel width parameter. Hereinafter, it is similarly used in the following kernel functions.

According to Theorem 1, for proving that Kullback-Leibler kernel is *pd* for all $t \in \mathbb{R}$, we need to prove that the symmetric form of KLD, *i.e.* $KLD(P\|Q) + KLD(Q\|P) : \mathcal{P} \times \mathcal{P} \mapsto \mathbb{R}$, is *nd*. For any $p_1, ..., p_m \in \mathcal{P}$ and $\alpha_1, ..., \alpha_m \in$

$\mathbb{R}$ with $\sum_{i=1}^{m} \alpha_i = 0$, we have:

$$
\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j \left(KLD(p_i \| p_j) + KLD(p_j \| p_i)\right) \\
&= 2\int_X \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j \left(p_i(x)\ln\frac{p_i(x)}{p_j(x)}\right) dx \\
&= 2\int_X \left(\sum_{j=1}^{m}\alpha_j \sum_{i=1}^{m}\alpha_i p_i(x)\ln p_i(x)\right. \\
&\quad \left. - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j p_i(x)\ln p_j(x)\right) dx \\
&= -2\int_X \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j p_i(x)\ln p_j(x) dx
\end{aligned}
\tag{5}
$$

Since currently it is hard to theoretically prove that (5) is nonpositive, we cannot justify the positive definiteness of Kullback-Leibler kernel. But it can still be used as a valid kernel and the numerical stability is guaranteed by shifting the kernel width t as [2]. Our empirical study also shows that Kullback-Leibler kernel with a proper value of t can be always guaranteed to be *pd* in the experiments.

2.2. Bhattacharyya Kernel

Formally, given continuous probability distributions P and Q, their Bhattacharyya Distance (BD) is closely related to their Bhattacharyya Coefficient (BC):

$$BD(P,Q) = -\ln(BC(P,Q)), \tag{6}$$

where BC is defined as:

$$BC(P,Q) = \int \sqrt{P(x)Q(x)}dx. \tag{7}$$

Bhattacharyya kernel is defined as follows:

$$K_{BD}(P,Q) = \exp\left(-\frac{BD(P,Q)}{2t^2}\right). \tag{8}$$

According to Theorem 1, Bhattacharyya kernel is *pd* for all $t \in \mathbb{R}$ if and only if $BD(P,Q)$ is *nd*, which can be proved if $BC(P,Q)$ is *pd*. As for the positive definiteness of $BC(P,Q)$, we give a theorem in the following.

Theorem 3. *Let $\mathcal{P}$ be a family of continuous probability distributions defined over a nonempty set X. Then Bhattacharyya Coefficient $BC(P,Q) : \mathcal{P} \times \mathcal{P} \mapsto \mathbb{R}$ defined by (7) is symmetric and **pd**.*

Proof. It is obvious that $BC(P,Q)$ is symmetric.

For any $p_1, ..., p_m \in \mathcal{P}$ and $\alpha_1, ..., \alpha_m \in \mathbb{R}$, we have

$$
\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j BC(p_i,p_j) = \int_X \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\sqrt{p_i(x)p_j(x)}dx \\
&= \int_X \sum_{i=1}^{m}\alpha_i\sqrt{p_i(x)}\sum_{j=1}^{m}\alpha_j\sqrt{p_j(x)}dx \\
&= \int_X \left(\sum_{i=1}^{m}\alpha_i\sqrt{p_i(x)}\right)^2 dx \geq 0
\end{aligned}
$$

Therefore, according to Definition 1, $BC(P,Q)$ is *pd*. □

2.3. Hellinger Kernel

For two continuous probability distributions P and Q, Hellinger Distance (HD) is defined as follows:

$$HD(P,Q) = \sqrt{\frac{1}{2}\int_X \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2 dx}. \tag{9}$$

According to the definition of BC and HD, HD can be also associated with BC as follows:

$$HD(P,Q) = \sqrt{1 - BC(P,Q)}. \tag{10}$$

We define Hellinger kernel as follows:

$$K_{HD}(P,Q) = \exp\left(-\frac{HD^2(P,Q)}{2t^2}\right). \tag{11}$$

According to Theorem 1, for proving the positive definiteness of Hellinger kernel, we only need to prove that $HD^2(P,Q)$ is *nd*.

Theorem 4. *Let $\mathcal{P}$ be a family of continuous probability distributions defined over a nonempty set X. Then the square of Hellinger Distance $HD(P,Q) : \mathcal{P} \times \mathcal{P} \mapsto \mathbb{R}$ defined by (10) is symmetric and **nd**.*

Proof. It is obvious that $HD^2(P,Q)$ is symmetric.

For any $p_1, ..., p_m \in \mathcal{P}$ and $\alpha_1, ..., \alpha_m \in \mathbb{R}$ with

$\sum_{i=1}^m \alpha_i = 0$, we have

$$\begin{aligned}
&\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j HD^2(p_i, p_j) \\
&= \frac{1}{2} \int_X \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j \left(\sqrt{p_i(x)} - \sqrt{p_j(x)} \right)^2 dx \\
&= \frac{1}{2} \int_X \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j \left(p_i(x) + p_j(x) - 2\sqrt{p_i(x) p_j(x)} \right) dx \\
&= \int_X \left[\sum_{i=1}^m \alpha_i \sum_{j=1}^m \alpha_j p_j(x) - \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j \sqrt{p_i(x) p_j(x)} \right] dx \\
&= - \int_X \sum_{i=1}^m \alpha_i \sqrt{p_i(x)} \sum_{j=1}^m \alpha_j \sqrt{p_j(x)} dx \\
&= - \int_X \left(\sum_{i=1}^m \alpha_i \sqrt{p_i(x)} \right)^2 dx \leq 0
\end{aligned}$$

Therefore, according to Definition 2, $HD^2(P, Q)$ is *nd*. □

2.4. Kernel based on Lie Group

Given two D-dimensional Gaussian distributions g_i, g_j, i.e. $g_i(x) = \mathcal{N}(x|\mu_i, \Sigma_i), g_j(x) = \mathcal{N}(x|\mu_j, \Sigma_j)$, let P_i and P_j denote the $(D+1) \times (D+1)$ SPD matrices corresponding to them respectively. We review the definition of the distance based on Lie Group (LGD) as follows:

$$LGD(P_i, P_j) = \| \log(P_i) - \log(P_j) \|_F, \tag{12}$$

where $P = |\Sigma|^{-\frac{1}{D+1}} \begin{pmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{pmatrix}$.

Then we define Kernel based on Lie Group in the following to measure the similarity between Gaussians g_i and g_j.

$$K_{LGD}(g_i, g_j) = \exp\left(-\frac{LGD^2(P_i, P_j)}{2t^2} \right) \tag{13}$$

According to Theorem 1, for proving the positive definiteness of Kernel based on Lie Group, we only need to prove that $LGD^2(P_i, P_j)$ is *nd*. This is obviously true according to Theorem 2.

2.5. Kernel based on Mahalanobis distance and Log-Euclidean distance

Given two Gaussian distributions g_i, g_j, i.e. $g_i(x) = \mathcal{N}(x|\mu_i, \Sigma_i), g_j(x) = \mathcal{N}(x|\mu_j, \Sigma_j)$, we first recall the kernel for mean vectors and covariance matrices respectively. The kernel based on Mahalanobis Distance (MD) for mean vectors is formulated as

$$K_{MD}(\mu_i, \mu_j) = \exp\left(-\frac{MD^2(\mu_i, \mu_j)}{2t^2} \right), \tag{14}$$

where

$$MD(\mu_i, \mu_j) = \sqrt{(\mu_i - \mu_j)^T (\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j)}, \tag{15}$$

and the kernel based on Log-Euclidean distance (LED) for covariance matrices is defined by

$$K_{LED}(\Sigma_i, \Sigma_j) = \exp\left(-\frac{LED^2(\Sigma_i, \Sigma_j)}{2t^2} \right), \tag{16}$$

where

$$LED(\Sigma_i, \Sigma_j) = \| \log(\Sigma_i) - \log(\Sigma_j) \|_F. \tag{17}$$

Then we linearly combine them to form the kernel based on MD and LED for Gaussian distributions g_i and g_j.

$$\begin{aligned} &K_{MD+LED}(g_i, g_j) \\ &= \gamma_1 K_{MD}(\mu_i, \mu_j) + \gamma_2 K_{LED}(\Sigma_i, \Sigma_j). \end{aligned} \tag{18}$$

where γ_i and γ_j are the combination coefficients.

We first need to state that the kernel based on MD and that based on LED are both valid kernels. Since the square of LED is obviously *nd*, according to Theorem 2, the kernel based on LED is *pd*. Though there is little understanding about the positive definiteness of the kernel based on MD, we can make it *pd* by properly choosing the kernel width parameter similar to the Kullback-Leibler Kernel. Finally according to [4], the superposition of the two *pd* kernels is a new valid kernel.

References

[1] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel methods on the riemannian manifold of symmetric positive definite matrices. In *IEEE Computer Society on Computer Vision and Pattern Recognition (CVPR)*, pages 73–80, 2013. 1

[2] P. J. Moreno, P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in Neural Information Processing Systems (NIPS)*, 2003. 2

[3] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938. 1

[4] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004. 3