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

Wen Wang, *Student Member, IEEE,* Ruiping Wang, *Member, IEEE,* Zhiwu Huang, *Member, IEEE,* Shiguang Shan, *Senior Member, IEEE,* and Xilin Chen, *Fellow, IEEE*

I. Positive definiteness of the kernels for Gaussians

First, we recall from [1] and state a theorem that gives a necessary and sufficient condition for obtaining a positive definite (*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 positive definite (**pd**) for all $t > 0$ if and only if f is negative definite (**nd**).*

In [2], the following theorem is proved accordingly.

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 $f(x_i, x_j) = \|\psi(x_i) - \psi(x_j)\|^2_{\mathcal{V}}$ is **nd**.*

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

A. The Kullback-Leibler Kernel

Currently it is hard to theoretically prove the positive definiteness of the Kullback-Leibler kernel in Equation (3). But it can still be used as a valid kernel and the numerical stability is guaranteed by shifting the kernel width t as [3]. Our empirical study also shows that the Kullback-Leibler kernel with a proper value of t can be always guaranteed to be *pd* in the experiments.

B. The Bhattacharyya Kernel

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

$$BD(P, Q) = -\ln(BC(P, Q)), \tag{S1}$$

where BC is defined as:

$$BC(P, Q) = \int \sqrt{P(x)Q(x)}dx. \tag{S2}$$

According to Theorem 1, the 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*. This can be easily proved in the following.

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

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j BC(p_i, p_j) &= \int_X \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\sqrt{p_i(x)p_j(x)}dx \\ &= \int_X \left(\sum_{i=1}^{m}\alpha_i\sqrt{p_i(x)}\right)^2 dx \geq 0\end{aligned} \tag{S3}$$

Therefore, according to the definition of *pd*, $BC(P, Q)$ is *pd*.

C. The Hellinger Kernel

The Hellinger Distance (HD) is defined as follows:

$$HD(P, Q) = \sqrt{\frac{1}{2}\int_X \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2 dx}. \tag{S4}$$

According to Theorem 1, for proving the positive definiteness of the Hellinger kernel in Equation (7), we only need to prove that $HD^2(P, Q)$ 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

$$\begin{aligned}&\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j HD^2(p_i, p_j) \\ &= \frac{1}{2}\int_X \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\left(\sqrt{p_i(x)} - \sqrt{p_j(x)}\right)^2 dx \\ &= -\int_X \left(\sum_{i=1}^{m}\alpha_i\sqrt{p_i(x)}\right)^2 dx \leq 0\end{aligned} \tag{S5}$$

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

D. The kernel based on Lie Group

According to Theorem 1, for proving the positive definiteness of the kernel based on Lie Group in Equation (9), we only need to prove that $LGD^2(P_i, P_j)$ is *nd*. This is obviously true according to Theorem 2.

E. The kernel based on Mahalanobis distance and Log-Euclidean distance for covariance matrices

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 valid kernel.

II. Gradient derivation

To deduce the partial derivative of the objective function F with respect to the transformation F in Equation (22), we derive the numerical formulation of $\frac{\partial}{\partial F} Dist(\hat{g}_i, \hat{g}_j)$ here.

A. Kullback-Leibler Distance

In this subsection, we derive the corresponding gradients of $\frac{\partial}{\partial F} Dist(\hat{g}_i, \hat{g}_j)$ by referring $Dist(\cdot,\cdot)$ to the symmetric KLD. Formally, given the corresponding distributions $\hat{g}_i, \hat{g}_j$ by mapping two Gaussian components g_i, g_j with F, their KLD is of the following form.

$$\begin{aligned} &KLD(\hat{g}_i, \hat{g}_j) \\ &= \frac{1}{2}\Bigg(tr\left((F^T\Sigma_j F)^{-1}F^T\Sigma_i F + (F^T\Sigma_i F)^{-1}F^T\Sigma_j F\right) \\ &\quad + (\mu_i-\mu_j)^T F(F^T\Sigma_i F)^{-1}F^T(\mu_i-\mu_j) \\ &\quad + (\mu_i-\mu_j)^T F(F^T\Sigma_j F)^{-1}F^T(\mu_i-\mu_j)\Bigg) - D, \end{aligned} \tag{S6}$$

Let $J_1^i(F) = (\mu_i-\mu_j)^T F(F^T\Sigma_i F)^{-1}F^T(\mu_i-\mu_j)$ and $J_2^{ij}(F) = tr\left((F^T\Sigma_j F)^{-1}F^T\Sigma_i F\right)$, we can rewrite Equation (S6) as follows.

$$\begin{aligned} &KLD(\hat{g}_i, \hat{g}_j) \\ &= \frac{1}{2}\Bigg(J_2^{ij}(F) + J_2^{ji}(F) + J_1^i(F) + J_1^j(F)\Bigg) - D. \end{aligned} \tag{S7}$$

Firstly, we consider $\frac{\partial J_1^i(F)}{\partial F}$ and compute its p-th row and q-th column element in the following.

$$\begin{aligned} &\frac{\partial J_1^i(F)}{\partial F_{pq}} \\ &= \left[\frac{\partial}{\partial F_{pq}}\left(F^T(\mu_i-\mu_j)\right)\right]^T (F^T\Sigma_i F)^{-1}F^T(\mu_i-\mu_j) \\ &\quad + (\mu_i-\mu_j)^T F\left[\frac{\partial}{\partial F_{pq}}(F^T\Sigma_i F)^{-1}\right]F^T(\mu_i-\mu_j) \\ &\quad + (\mu_i-\mu_j)^T F(F^T\Sigma_i F)^{-1}\left[\frac{\partial}{\partial F_{pq}}\left(F^T(\mu_i-\mu_j)\right)\right]. \end{aligned} \tag{S8}$$

We can easily obtain

$$\frac{\partial}{\partial F_{pq}}\left(F^T(\mu_i-\mu_j)\right) = (\mu_i-\mu_j)_q \tag{S9}$$

and

$$\begin{aligned} &\frac{\partial}{\partial F_{pq}}(F^T\Sigma_i F)^{-1} \\ &= -(F^T\Sigma_i F)^{-1}\left[\frac{\partial}{\partial F_{pq}}(F^T\Sigma_i F)\right](F^T\Sigma_i F)^{-1} \\ &= -(F^T\Sigma_i F)^{-1}\left(F^T\Sigma_i E^{pq} + E^{qp}\Sigma_i F\right)(F^T\Sigma_i F)^{-1}, \end{aligned} \tag{S10}$$

where E^{pq} is the single-entry matrix with 1 at (p,q) and 0 elsewhere. Hence, $\frac{\partial J_1^i(F)}{\partial F}$ can be finally formulated by substituting Equation (S9) and (S10) into (S8).

Next, we compute the partial derivative of $J_2^{ij}(F)$ with respect to F.

$$\begin{aligned} \frac{\partial J_2^{ij}(F)}{\partial F_{pq}} &= \left[\frac{\partial}{\partial F_{pq}}(F^T\Sigma_j F)^{-1}\right]F^T\Sigma_i F \\ &\quad + (F^T\Sigma_j F)^{-1}\left[\frac{\partial}{\partial F_{pq}}\left(F^T\Sigma_i F\right)\right], \end{aligned} \tag{S11}$$

where $\frac{\partial}{\partial F_{pq}}(F^T\Sigma_j F)^{-1}$ and $\frac{\partial}{\partial F_{pq}}\left(F^T\Sigma_i F\right)$ have been given in Equation (S10).

B. Bhattacharyya Distance

Here we give the derivation when $Dist(\cdot,\cdot)$ refers to BD in Equation (4). By defining

$$J_1(F) = (\mu_i-\mu_j)^T F(F^T\Sigma F)^{-1}F^T(\mu_i-\mu_j), \tag{S12}$$

we can see that it has a similar numerical formulation with $J_1^i(F)$. Then noting that

$$\ln(\det(F^T\Sigma F)) = 2\Sigma F(F^T\Sigma F)^{-1}, \tag{S13}$$

the partial derivative of $BD(\hat{g}_i, \hat{g}_j)$ with respect to F can be formulated as follows.

$$\begin{aligned} \frac{\partial}{\partial F}BD(\hat{g}_i, \hat{g}_j) &= \frac{1}{8}\frac{\partial}{\partial F}J_1(F) + \Sigma F(F^T\Sigma F)^{-1} \\ &\quad - \frac{1}{2}F(F^T\Sigma_i F)^{-1} - \frac{1}{2}F(F^T\Sigma_j F)^{-1}. \end{aligned} \tag{S14}$$

C. Hellinger Distance

For HD, it can be formulated with BD accordingly.

$$HD^2(\hat{g}_i, \hat{g}_j) = 1 - \exp\left\{-BD(\hat{g}_i, \hat{g}_j)\right\}, \tag{S15}$$

Hence, the derivation is similar with that of BD.

References

[1] I. J. Schoenberg, "Metric spaces and positive definite functions," in *Transactions of the American Mathematical Society*, vol. 44, no. 3, 1938, pp. 522–536.

[2] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. T. Harandi, "Kernel methods on the riemannian manifold of symmetric positive definite matrices," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 73–80.

[3] 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, pp. 1385–1392.

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