Statistical Papers (2021) 62:1499–1518
https://doi.org/10.1007/s00362-019-01144-5

REGULAR ARTICLE

Admissible kernels for RKHS embedding of probability distributions

Liangzhi Chen[1] · Thomas Hotz[2] · Haizhang Zhang[3]

Received: 2 September 2018 / Revised: 18 October 2019 / Published online: 15 November 2019

Abstract

Similarity measurement of two probability distributions is important in many applications of statistics. Embedding such distributions into a reproducing kernel Hilbert space (RKHS) has many favorable properties. The choice of the reproducing kernel is crucial in the approach. We study this question by considering the similarity of two distributions of the same class. In particular, we investigate when the RKHS embedding is “admissible” in the sense that the distance between the embeddings should become smaller when the expectations are getting closer or when the variance is increasing to infinity. We give conditions on the widely-used translation-invariant reproducing kernels to be admissible. We also extend the study to multivariate non-symmetric Gaussian distributions.

Keywords Gaussian distributions · Reproducing kernels · RKHS embedding · Translation-invariant kernels · Radially decreasing functions

H. Zhang: Supported in part by NSFC under Grants 11571377 and 11971490, and by the Guangdong Natural Science Foundation (No. 2018A030313841). This work was completed while this author was visiting the Hong Kong University of Science and Technology.

✉ Haizhang Zhang
zhhaizh2@sysu.edu.cn

Liangzhi Chen
chenlzh29@mail.sysu.edu.cn

Thomas Hotz
thomas.hotz@tu-ilmenau.de

[1] School of Data and Computer Science, Sun Yet-sen University, Guangzhou, People’s Republic of China

[2] Institute of Mathematics, TU Ilmenau, Ilmenau, Germany

[3] School of Data and Computer Science and Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou, People’s Republic of China

1 Introduction

Distance between probability measures has many applications, including distribution testing, density estimation, signal detection, etc (Rachev 1991; Vajda 1989). In recent years, many kinds of distance between probability measures have been proposed (see, for example, Sriperumbudur et al. 2009, 2010). Many of them are built on the general approach of *integral probability metric* (*IPM*) (Müller 1997).

To introduce the approach, denote by $\mathcal{P}$ the set of all Borel probability measures on a probability space $(M, \mathcal{A})$. The IPM between $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}$ is defined as

$$\gamma_{\mathcal{F}}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{f \in \mathcal{F}} \left| \int_M f \, \mathrm{d}\mathbb{P}_1 - \int_M f \, \mathrm{d}\mathbb{P}_2 \right|, \tag{1}$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on M. Different choices of the class $\mathcal{F}$ yield different metrics $\gamma_{\mathcal{F}}$ on the given probability space. The followings are among the well-known examples in the literature:

1. Total variation distance: $\mathcal{F} = \mathrm{C}_{bu}(M)$, the space of all uniformly bounded continuous functions on M or $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$, where $\|f\|_\infty = \sup_{x \in M} |f(x)|$ (see, for example, Shorack 2000, Chapter 19);
2. The Kolmogorov distance: $\mathcal{F} = \{1_{(-\infty, t]} : t \in \mathbb{R}^d\}$, where 1_A denotes the characteristic function of a subset A of $\mathbb{R}^d$ (see, for example, Shorack 2000, Chapter 19);
3. The Kantorovich metric or Wasserstein distance: $\mathcal{F} := \{f : \|f\|_L \leq 1\}$ where $\|f\|_L := \sup\{|f(x) - f(y)|/\rho(x, y),\ x \neq y \in M\}$ with M being a metric space with metric ρ (see, Dudley 2002, Theorem 11.8.2);
4. Reproducing kernel Hilbert space embedding of measures: $\mathcal{F} = \{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \leq 1\}$, where $\mathcal{H}_K$ is the reproducing kernel Hilbert space of a reproducing kernel K on M (Gretton et al. 2007; Smola et al. 2007).
5. Reproducing kernel Banach space embedding of measures (Sriperumbudur et al. 2011): $\mathcal{F} = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq 1\}$, where $\mathcal{B}$ is a reproducing kernel Banach space on M (Song et al. 2013; Zhang et al. 2009).

This paper attempts to contribute to the approach 4 above. We introduce the concept of reproducing kernel Hilbert spaces and reproducing kernels (Aronszajn 1950).

Definition 1.1 Let M be a prescribed set. A reproducing kernel on M is a real-valued function $K : M \times M \to \mathbb{R}$ such that for all finite points $x_1, x_2, \ldots, x_n \in M$, the matrix

$$[K(x_j, x_k)]_{j,k=1}^n$$

is symmetric and positive semi-definite.

For a reproducing kernel K on M, there exists a unique associated Hilbert space denoted by $\mathcal{H}_K$ consisting of certain functions on M such that $K(x, \cdot) \in \mathcal{H}_K$ for all $x \in M$ and

$$f(x) = \langle f, K(x, \cdot)\rangle_{\mathcal{H}_K} \text{ for all } f \in \mathcal{H}_K,\ x \in M,$$

where $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$ denotes the inner product on $\mathcal{H}_K$. The space $\mathcal{H}_K$ is called the *reproducing kernel Hilbert space* (RKHS) of the reproducing kernel K. We introduce the notation

$$\gamma_K(\mathbb{P}_1,\mathbb{P}_2)=\sup_{f\in\mathcal{H}_K,\|f\|_{\mathcal{H}_K}\le 1}\left|\int_M f\,\mathrm{d}\mathbb{P}-\int_M f\,\mathrm{d}\mathbb{Q}\right|. \tag{2}$$

When only finite i.i.d. random samples $\{X_i : 1\le i\le m\}$ and $\{Y_j : 1\le j\le n\}$ drawn from unknown measures $\mathbb{P}_1,\mathbb{P}_2$ are available, one approximates $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively by

$$\mathbb{P}_{1m}:=\frac{1}{m}\sum_{i=1}^{m}\delta_{X_i}\ \text{and}\ \mathbb{P}_{2n}:=\frac{1}{n}\sum_{j=1}^{n}\delta_{Y_j}$$

and thereby approximating $\gamma_{\mathcal{F}}(\mathbb{P}_1,\mathbb{P}_2)$ by $\gamma_{\mathcal{F}}(\mathbb{P}_{1m},\mathbb{P}_{2n})$. By choosing $\mathcal{F}$ to be the unit ball of the reproducing kernel Hilbert space of a reproducing kernel K, the approach of reproducing kernel Hilbert space embedding of measures enjoys many advantages over other approaches (Gretton et al. 2007; Sriperumbudur et al. 2009; Weaver 1999). Firstly, $\gamma_K(\mathbb{P}_{1m},\mathbb{P}_{2n})$ is simply a sum of expectations of the kernel K and hence is much easier to compute compared to other choices. Secondly, $\gamma_K(\mathbb{P}_{1m},\mathbb{P}_{2n})$ is a $\sqrt{mn/(m+n)}$-consistent estimate of $\gamma_K(\mathbb{P}_1,\mathbb{P}_2)$ for all $\mathbb{P}_1,\mathbb{P}_2$ under the mild conditions that K is measurable and bounded (Gretton et al. 2007). Thirdly, when K is translation-invariant, the rate of approximating $\gamma_K(\mathbb{P}_1,\mathbb{P}_2)$ by $\gamma_K(\mathbb{P}_{1m},\mathbb{P}_{2n})$ is independent of the dimension (Sriperumbudur et al. 2009).

Despite many favorable properties, there is one critical question not well-addressed in the RKHS embedding of measures, which is the **choice of reproducing kernels**. An RKHS $\mathcal{H}_K$ is completely determined by its reproducing kernel K (Aronszajn 1950; Zhang and Zhao 2013). In fact, $\mathcal{H}_K$ is the completion of the linear space

$$\operatorname{span}\{K(x,\cdot) : x\in M\}$$

under the inner product

$$\left\langle\sum_{j=1}^{p}c_jK(x_j,\cdot),\sum_{k=1}^{q}d_kK(y_k,\cdot)\right\rangle_{\mathcal{H}_K}=\sum_{j=1}^{p}\sum_{k=1}^{q}c_jd_kK(x_j,y_k)\qquad c_j,d_k\in\mathbb{R}.$$

Therefore, the choice of the reproducing kernel K much affects the embedding of probability measures in $\mathcal{H}_K$. So far, studies in the literature have focused on **characteristic kernels** which ensure $\gamma_K(\mathbb{P}_1,\mathbb{P}_2)$ to be a metric on $(M,\mathcal{A})$ (see, for example, Berlinet and Thomas-Agnan 2004; Chen et al. 2016; Fukumizu et al. 2009, 2008; Gretton et al. 2007; Sriperumbudur et al. 2011, 2009, 2010; Steinwart 2001). Being characteristic can be viewed as a preliminary requirement on the reproducing kernel. Our attempt in this paper is to impose another **admissibility criterion** on the reproducing kernel in measuring the similarity of a class of probability distributions. Let us make our objective clear.

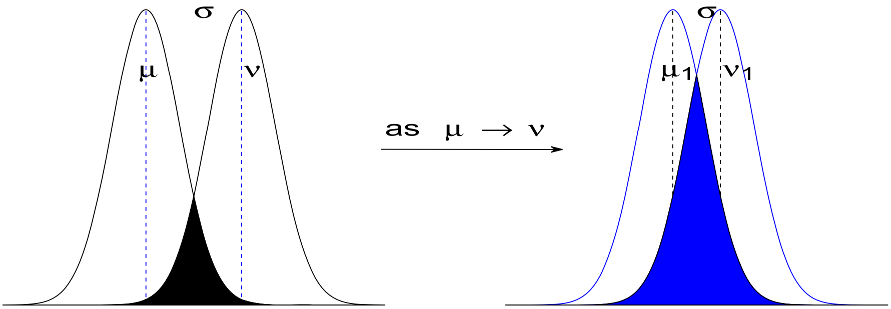

Fig. 1 Let $\mu < \nu$ and σ be fixed. Then the regions under the curves of the density functions of $\mathbb{P}_{\mu,\sigma}$ and $\mathbb{P}_{\nu,\sigma}$ have a larger overlapping area when μ, ν become closer

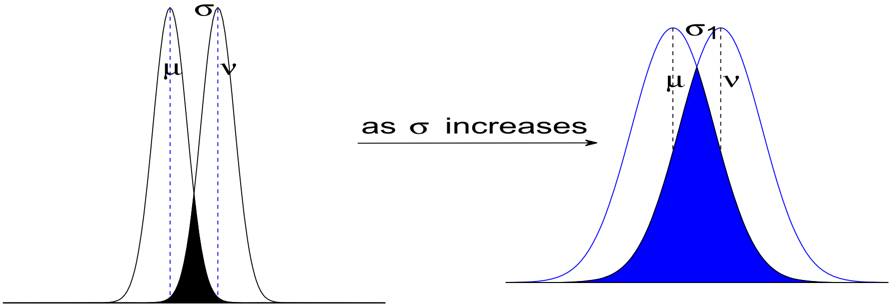

Fig. 2 Let $\mu < \nu$ be fixed. Then the regions under the curves of the density functions of $\mathbb{P}_{\mu,\sigma}$ and $\mathbb{P}_{\nu,\sigma}$ have a larger overlapping area as σ increases

Assume that K is a characteristic kernel. Thus $\gamma_K(\mathbb{P}_1, \mathbb{P}_2)$ is a metric and can be used to measure the similarity between two probability measures $\mathbb{P}_1, \mathbb{P}_2$. Consider the most important class of Gaussian distributions measures

$$\mathrm{d}\mathbb{P}_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x, \quad x \in \mathbb{R}.$$

with mean μ and standard deviation σ. Naturally, two Gaussian measures $\mathbb{P}_{\mu_1,\sigma_1}$ and $\mathbb{P}_{\mu_2,\sigma_2}$ should be closer in the following two cases (see Figs. 1 and 2 for illustration and explanation):

(i) when the means are getting closer, that is, μ_1 tends to μ_2;
(ii) when $\sigma_1 = \sigma_2$ are increasing while the means are different but fixed.

To summarize, we shall study conditions on the reproducing kernel K that is **admissible** for the Gaussian distributions in the following sense. Denote by $\|\cdot\|$ the standard Euclidean norm on $\mathbb{R}^d$.

Definition 1.2 Let $\mathbb{Q}$ be a Borel probability measure on $\mathbb{R}^d$. A reproducing kernel K on $\mathbb{R}^d$ is said to be **admissible** for the class of distributions

$$\mathrm{d}\mathbb{P}_{\mu,\sigma}(x) = \frac{1}{\sigma^d}\,\mathrm{d}\mathbb{Q}\left(\frac{x-\mu}{\sigma}\right), \qquad x \in \mathbb{R}^d, \mu \in \mathbb{R}^d, \sigma > 0$$

(or simply, $\mathbb{Q}$-admissible), if the following two conditions are satisfied:

(A1) $\gamma_K(\mathbb{P}_{\mu_1,\sigma}, \mathbb{P}_{\mu_2,\sigma})$ is strictly decreasing as $\|\mu_1 - \mu_2\|$ decreases;
(A2) $\gamma_K(\mathbb{P}_{\mu_1,\sigma}, \mathbb{P}_{\mu_2,\sigma})$ is strictly decreasing as σ increases.

We present sufficient conditions for a translation-invariant kernel K to be $\mathbb{Q}$-admissible in Sect. 3. The concrete examples of Gaussian distributions is then investigated. To this end, we present necessary preliminaries on reproducing kernels and RKHS embeddings of probability measures in Sect. 2. Section 3 is devoted to non-symmetric multivariate Gaussian distributions. We remark that by the illustration of our motivation in Figs. 1 and 2, the notion of admissible kernels introduced in the paper seems useful only for probability distributions of a single mode such as the Gaussian distributions. Kernel methods to evaluate the distance between probability distributions of multiple modes would be an interesting question for us in the future.

2 Admissible kernels

Let K be a reproducing kernel on $\mathbb{R}^d$ that is translation-invariant in the sense

$$K(x, y) = K(x - z, y - z)$$

for all $x, y, z \in \mathbb{R}^d$. It is easy to see

$$K(x, y) = \psi(x - y), \quad x, y \in \mathbb{R}^d \tag{3}$$

for some function $\psi(x) = K(x, 0)$ on $\mathbb{R}^d$. By the celebrated Bochner theorem (Bochner 1959), if ψ is continuous on $\mathbb{R}^d$ then $K(x, y) = \psi(x - y)$ makes a reproducing kernel on $\mathbb{R}^d$ if and only if there exists a finite positive Borel measure ρ on $\mathbb{R}^d$ such that

$$\psi(x) = \int_{\mathbb{R}^d} e^{-ix\cdot t}\,\mathrm{d}\rho(t), \quad x \in \mathbb{R}^d. \tag{4}$$

It was shown in Sriperumbudur et al. (2010) that K is a characteristic kernel, that is, γ_K is a metric, if and only if ρ is supported on the whole $\mathbb{R}^d$. For simplicity, we also assume that ρ is symmetric about the origin so that ψ and K are real-valued.

Denote by $\mathcal{P}(\mathbb{R}^d)$ the set of all Borel probability measures on $\mathbb{R}^d$. Let $\mathbb{R}_+ := [0, +\infty)$ and let $\mathbb{N}$ be the set of positive integers. We shall need the convolution of a bounded continuous function f on $\mathbb{R}^d$ and a measure $\rho \in \mathcal{P}(\mathbb{R}^d)$ given as

$$(f * \rho)(x) := \int_{\mathbb{R}^d} f(x - y)\,\mathrm{d}\rho(y), \quad x \in \mathbb{R}^d.$$

and the convolution of two probability measures $\varrho, \lambda \in \mathcal{P}(\mathbb{R}^d)$

$$(\varrho * \lambda)(E) := \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} 1_E(x - y)\, \mathrm{d}\varrho(x)\, \mathrm{d}\lambda(y).$$

For a probability distribution $\mathbb{Q}$, denote

$$\bar{\mathbb{Q}} := \mathbb{Q} * \tilde{\mathbb{Q}}, \tag{5}$$

where $\mathrm{d}\tilde{\mathbb{Q}}(x) := \mathrm{d}\mathbb{Q}(-x)$.

Let K be given by (3) and (4), where ρ is a finite positive Borel measure on $\mathbb{R}^d$. Then K is bounded on $\mathbb{R}^d \times \mathbb{R}^d$. Consequently, the function

$$f_\varrho(x) := \int_{\mathbb{R}^d} K(x, t)\, \mathrm{d}\varrho(t), \quad x \in \mathbb{R}^d$$

is well-defined for each $\varrho \in \mathcal{P}(\mathbb{R}^d)$. An important observation made in Sriperumbudur et al. (2010) is that $f_\varrho \in \mathcal{H}_K$ for all $\varrho \in \mathcal{P}(\mathbb{R}^d)$ and that

$$\gamma_K(\varrho, \lambda) = \left\| f_\varrho - f_\lambda \right\|_{\mathcal{H}_K}, \quad \varrho, \lambda \in \mathcal{P}(\mathbb{R}^d). \tag{6}$$

Moreover,

$$\langle f_\varrho, g \rangle_{\mathcal{H}_K} = \int_{\mathbb{R}^d} g(x)\, \mathrm{d}\varrho(x), \quad \varrho \in \mathcal{P}(\mathbb{R}^d), \ g \in \mathcal{H}_K. \tag{7}$$

Returning to our main theme, we let $\mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$ and define the associated class of probability measures

$$\mathrm{d}\mathbb{P}_{\mu,\sigma}(x) = \frac{1}{\sigma^d}\, \mathrm{d}\mathbb{Q}\left(\frac{x - \mu}{\sigma}\right) \ x \in \mathbb{R}^d, \quad \mu \in \mathbb{R}^d, \sigma > 0. \tag{8}$$

We first give an initial result for the reproducing kernel K given by (3), (4) to satisfy the two admissible requirements **(A1)** and **(A2)**. To this end, we introduce the following definitions.

Definition 2.1 Let f be a function on $\mathbb{R}^d$. Then it is said to be

- *radial* provided that $f(x) = f(y)$ whenever $\|x\| = \|y\|$;
- *radially decreasing* if f is radial and $f(x) \leq f(y)$ whenever $\|x\| > \|y\|$;
- *strictly radially decreasing* if f is radial and $f(x) < f(y)$ whenever $\|x\| > \|y\|$;
- *radially increasing* (*strictly radially increasing*) if $-f$ is radially decreasing (strictly radially decreasing).

It can be verified that the convolution of two radial functions remains radial. Interested readers are referred to Lieb and Loss (2001) for more properties about radial functions.

Lemma 2.2 *Let K be the translation-invariant kernel given by (3) and (4), where ρ is supported on the whole $\mathbb{R}^d$. Define the function*

$$\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(x) := (\psi_\sigma * \bar{\mathbb{Q}})(x), \quad x \in \mathbb{R}^d, \tag{9}$$

where

$$\psi_\sigma(x) := \psi(\sigma x), \quad x \in \mathbb{R}^d.$$

Then K is admissible for the class of distributions (8) if and only if $\mathcal{G}_{\psi_\sigma,\mathbb{Q}}$ is strictly decreasing as $\|x\|$ increases for any $\sigma > 0$ and as σ increases for fixed $\|x\|$.

Proof By (6) and (7), we have for two measures $\varrho, \lambda \in \mathcal{P}(\mathbb{R}^d)$

$$\begin{aligned}
\big(\gamma_K(\varrho,\lambda)\big)^2 &= \| f_\varrho - f_\lambda \|^2_{\mathcal{H}_K} \\
&= \langle f_\varrho, f_\varrho \rangle_{\mathcal{H}_K} + \langle f_\lambda, f_\lambda \rangle_{\mathcal{H}_K} - 2\langle f_\varrho, f_\lambda \rangle_{\mathcal{H}_K} \\
&= \int_{\mathbb{R}^d} f_\varrho(x)\,\mathrm{d}\varrho(x) + \int_{\mathbb{R}^d} f_\lambda(x)\,\mathrm{d}\lambda(x) - 2\int_{\mathbb{R}^d} f_\varrho(x)\,\mathrm{d}\lambda(x) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K(x,y)\,\mathrm{d}\varrho(x)\,\mathrm{d}\varrho(y) + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K(x,y)\,\mathrm{d}\lambda(x)\,\mathrm{d}\lambda(y) \\
&\quad - 2\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K(x,y)\,\mathrm{d}\varrho(x)\,\mathrm{d}\lambda(y).
\end{aligned} \tag{10}$$

Specifying the distributions $\mathbb{P}_{\mu_1,\sigma}, \mathbb{P}_{\mu_2,\sigma}$, we compute

$$\begin{aligned}
\gamma_K^2(\mathbb{P}_{\mu_1,\sigma}, \mathbb{P}_{\mu_2,\sigma}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \psi(x-y)\,\mathrm{d}\mathbb{P}_{\mu_1,\sigma}(y)\,\mathrm{d}\mathbb{P}_{\mu_1,\sigma}(x) \\
&\quad + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \psi(x-y)\,\mathrm{d}\mathbb{P}_{\mu_2,\sigma}(y)\,\mathrm{d}\mathbb{P}_{\mu_2,\sigma}(x) \\
&\quad - 2\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \psi(x-y)\,\mathrm{d}\mathbb{P}_{\mu_2,\sigma}(y)\,\mathrm{d}\mathbb{P}_{\mu_1,\sigma}(x) \\
&= \sigma^{-2d}\Big[\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \psi(\sigma x + \mu_1 - \sigma y - \mu_1)\,\mathrm{d}\mathbb{Q}(y)\,\mathrm{d}\mathbb{Q}(x) \\
&\quad + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \psi(\sigma x + \mu_2 - \sigma y - \mu_2)\,\mathrm{d}\mathbb{Q}(y)\,\mathrm{d}\mathbb{Q}(x) \\
&\quad - \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \psi(\sigma x + \mu_1 - \sigma y - \mu_2)\,\mathrm{d}\mathbb{Q}(y)\,\mathrm{d}\mathbb{Q}(x)\Big] \\
&= \sigma^{-2d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \big[\psi\,(\sigma(x-y)) - \psi\big(\sigma(x-y) + \mu_1 - \mu_2\big)\big]\,\mathrm{d}\mathbb{Q}(y)\,\mathrm{d}\mathbb{Q}(x) \\
&= \sigma^{-2d}\left[\big(\psi_\sigma * \bar{\mathbb{Q}}\big)(0) - (\psi_\sigma * \bar{\mathbb{Q}})\left(\frac{\mu_1 - \mu_2}{\sigma}\right)\right].
\end{aligned}$$

Since ρ is supported on the whole real line, γ_K is a metric on $\mathcal{P}(\mathbb{R}^d)$. Then by the above calculation, we have that

$$(\psi_\sigma * \bar{\mathbb{Q}})(t) < (\psi_\sigma * \bar{\mathbb{Q}})(0) \ \text{ for } t \neq 0.$$

As a result, the requirements **(A1)** and **(A2)** are satisfied if and only if $(\psi_\sigma * \bar{\mathbb{Q}})(t)$ is monotonically decreasing as $\|t\|$ increases for fixed $\sigma > 0$ and as σ increases for fixed $\|t\|$. □

Specifying a certain class of probability distributions, we are able to show that not all translation-invariant characteristic kernels are admissible for the RKHS embedding of probability distributions.

Example 2.3 Consider dimension $d = 1$. Let $\mathbb{Q}$ be the Bernoulli distribution with success probability $\frac{1}{2}$, i.e. $\mathbb{Q}(0) = \mathbb{Q}(1) = \frac{1}{2}$. Then we have $\bar{\mathbb{Q}}(-1) = \bar{\mathbb{Q}}(1) = \frac{1}{4}$ and $\bar{\mathbb{Q}}(0) = \frac{1}{2}$. With ψ being given by (4), we have

$$\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(t) = \psi_\sigma * \bar{\mathbb{Q}}(t) = \frac{1}{2}\psi(\sigma t) + \frac{1}{4}\psi(\sigma t - \sigma) + \frac{1}{4}\psi(\sigma t + \sigma)$$

Now let ψ be an even function on $\mathbb{R}$ which is strictly decreasing on $\mathbb{R}_+$ and converging to 0 (examples including $e^{-|t|}$, e^{-t^2}, etc.). Choose σ large enough such that

$$\psi\Big(\frac{\sigma}{2}\Big) < \frac{1}{4}\psi(0).$$

Then

$$\begin{aligned}
\mathcal{G}_{\psi_\sigma,\mathbb{Q}}\left(\frac{1}{2}\right) &= \tfrac{1}{4}\psi_\sigma\big(-\tfrac{1}{2}\big) + \tfrac{1}{2}\psi_\sigma\big(\tfrac{1}{2}\big) + \tfrac{1}{4}\psi_\sigma\big(\tfrac{3}{2}\big) \\
&\leq \psi_\sigma\big(\tfrac{1}{2}\big) \\
&< \tfrac{1}{4}\psi_\sigma(0) + \tfrac{1}{2}\psi_\sigma(1) + \tfrac{1}{4}\psi_\sigma\big(2\big) = \mathcal{G}_{\psi_\sigma,\mathbb{Q}}(1).
\end{aligned}$$

Therefore, the function $\mathcal{G}_{\psi_\sigma,\mathbb{Q}}$ is not monotonically decreasing for this example. By Lemma 2.2, the kernel K is not admissible.

Next we shall present a sufficient condition and a necessary condition guaranteeing admissibility, which covers a large class of reproducing kernels and probability distributions.

Theorem 2.4 *Let ψ and $\bar{\mathbb{Q}}$ be defined by (4) and (5), respectively, with ρ supported on $\mathbb{R}^d$. Suppose that $\bar{\mathbb{Q}}$ has a Lebesgue integrable density function f. If both ψ and f are radially decreasing with at least one of them being strictly radially decreasing then the kernel K given by (3) is admissible for the class of distributions (8).*

Conversely, suppose that f and ψ are radial and that f is radially decreasing. If the kernel K is admissible, then ψ is radially decreasing.

Proof For the first part, by Lemma 2.2, we need to show that the function $\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(t)$ defined by (9) is monotonically decreasing as $\|t\|$ and σ increase.

We only consider the case when σ is fixed, the case when σ varies can be handled similarly. Let δ, Δ be two points in $\mathbb{R}^d$ with $\|\delta\| < \|\Delta\|$. As ψ_σ, f are both radial, so is $\mathcal{G}_{\psi_\sigma,\mathbb{Q}} = \psi_\sigma * f$. We may hence assume $\delta = (\delta_1, 0, \ldots, 0)$, $\Delta = (\Delta_1, 0, \ldots, 0)$, where $0 \le \delta_1 < \Delta_1$. Also, we may assume further that ψ is the one which is strictly radially decreasing.

Set $H_t := \{x \in \mathbb{R}^d : x_1 < t\}$, $t \in \mathbb{R}$. We first write

$$\begin{aligned}
&\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(\delta) - \mathcal{G}_{\psi_\sigma,\mathbb{Q}}(\Delta) = (\psi_\sigma * f)(\delta) - (\psi_\sigma * f)(\Delta)\\
&\quad = \int_{\mathbb{R}^d} \big[\psi_\sigma(\delta - x) - \psi_\sigma(\Delta - x)\big] f(x)\,\mathrm{d}x\\
&\quad = \int_{H_{\frac{\delta_1+\Delta_1}{2}}} \big[\psi_\sigma(\delta - x) - \psi_\sigma(\Delta - x)\big] f(x)\,\mathrm{d}x\\
&\qquad - \int_{\mathbb{R}^d \setminus H_{\frac{\delta_1+\Delta_1}{2}}} \big[\psi_\sigma(\Delta - x) - \psi_\sigma(\delta - x)\big] f(x)\,\mathrm{d}x.
\end{aligned}$$

We then apply the substitution $x = \delta + \Delta - t$ to the second integral above and use the radiality of ψ and f to get

$$\begin{aligned}
&\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(\delta) - \mathcal{G}_{\psi_\sigma,\mathbb{Q}}(\Delta)\\
&\quad = \int_{H_{\frac{\delta_1+\Delta_1}{2}}} \big[\psi_\sigma(\delta - x) - \psi_\sigma(\Delta - x)\big]\big[f(x) - f(x - \delta - \Delta)\big]\,\mathrm{d}x. \qquad (11)
\end{aligned}$$

Note that for $x \in H_{\frac{\delta_1+\Delta_1}{2}}$,

$$\|\delta - x\| < \|\Delta - x\| \ \text{ and } \ \|x\| < \|x - \delta - \Delta\|.$$

Since ψ_σ is strictly radially decreasing and f is radially decreasing,

$$\psi_\sigma(\delta - x) - \psi_\sigma(\Delta - x) > 0 \text{ and } f(x) - f(x - \delta - \Delta) \ge 0, \quad x \in H_{\frac{\delta_1+\Delta_1}{2}}.$$

As a consequence, we get by (11) that $\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(\delta) - \mathcal{G}_{\psi_\sigma,\mathbb{Q}}(\Delta) \ge 0$.

To obtain $\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(\delta) - \mathcal{G}_{\psi_\sigma,\mathbb{Q}}(\Delta) > 0$, we have to show that the set $E = \{x \in H_{\frac{\delta_1+\Delta_1}{2}} : f(x) \ne f(x - \delta - \Delta)\}$ has a positive Lebesgue measure on $\mathbb{R}^d$. Assume on the contrary that the Lebesgue measure of E equals 0. Then f is a periodic function on $H_{\frac{\delta_1+\Delta_1}{2}} \setminus E$. Since f is radially decreasing, it must be constant on $H_{\frac{\delta_1+\Delta_1}{2}} \setminus E$. This together with the assumption that f is radial implies that it equals a constant almost everywhere on $\mathbb{R}^d$. It is hence impossible to be the density function of a probability distribution, which is a contradiction.

For the second part, we only have to prove that if f is radially decreasing, then the kernel K being admissible implies ψ being radially decreasing. Since f is radially decreasing and is the density function of $\bar{\mathbb{Q}}$, we have $\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = 1$. Define

$$\begin{aligned}\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(x) &= \int_{\mathbb{R}^d} \psi_\sigma(x-y) f(y)\, \mathrm{d}y\\ &= \frac{1}{\sigma^d}\int_{\mathbb{R}^d} \psi(\sigma x - y) f\left(\frac{y}{\sigma}\right) \mathrm{d}y,\end{aligned}$$

where ψ is a bounded and continuous function on $\mathbb{R}^d$. The density function $\frac{1}{\sigma^d} f(\frac{y}{\sigma})$ clearly converges in distribution to the Dirac mass δ_0 in y when $\sigma \to 0$. Hence $\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(x) \to \psi(x)$ when $\sigma \to 0$. Now assume on the contrary that ψ is not radially decreasing. Then there exist two points x_0, y_0 such that $\|x_0\| \le \|y_0\|$ and $\psi(x_0) < \psi(y_0)$. Let ε be a real number such that $0 < \varepsilon < \psi(y_0) - \psi(x_0)$. Then by the continuity of ψ and the fact that $\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(x) \to \psi(x)$ $(as\ \sigma \to 0)$, there exists σ small enough such that

$$\mathcal{G}_{\psi_\sigma,\mathbb{Q}}(x_0/\sigma) \le \psi(x_0) + \varepsilon/2 < \psi(y_0) - \varepsilon/2 \le \mathcal{G}_{\psi_\sigma,\mathbb{Q}}(y_0/\sigma).$$

Therefore, by Lemma 2.2 we know that K is not admissible. So if K is admissible, then ψ must be radially decreasing. □

Note that the kernel which satisfies the assumptions in Theorem 2.4 is of the form

$$K(x, y) = \psi(x-y) = \phi(\|x-y\|), \quad x, y \in \mathbb{R}^d, \tag{12}$$

where ϕ is decreasing on $\mathbb{R}_+$. Kernels of the above form are called radial basis functions (Wendland 2005; Wu 1995). Let ϕ be a function on $\mathbb{R}_+$ such that $\phi(\|x-y\|)$ is a reproducing kernel on $\mathbb{R}^d$. It is quite natural to ask when ϕ is strictly decreasing. A fundamental result on radial basis functions due to Schoenberg (1938) states that $\phi(\|x-y\|)$ makes a reproducing kernel on $\mathbb{R}^d$ for all dimensions $d \in \mathbb{N}$ if and only if there is a finite positive Borel measure μ on $\mathbb{R}_+$ such that

$$\phi(t) = \int_0^{+\infty} \exp(-st^2)\, \mathrm{d}\mu(s), \quad t \in \mathbb{R}_+.$$

In this case, ϕ is automatically decreasing on $\mathbb{R}_+$ and is strictly decreasing as long as $\operatorname{supp}\mu \ne \{0\}$, which is equivalent to say that the radial kernel K in (12) is characteristic (Sriperumbudur et al. 2011, Proposition 5). In conclusion, if for all $d \in \mathbb{N}$, K is a nontrivial reproducing kernel on $\mathbb{R}^d$ and a characteristic kernel, then we have by Theorem 2.4 that K is $\mathbb{Q}$-admissible provided that the density function f of $\mathbb{Q}$ is radially decreasing.

Things are different if $\phi(\|x-y\|)$ is only a kernel on certain dimensions. We present an explicit example to illustrate this. It was proved in Schoenberg (1938) that for a fixed dimension d, $\phi(\|x-y\|)$ is a kernel on $\mathbb{R}^d$ if and only if

$$\phi(t) = \int_0^{+\infty} \Omega_d(ts)\, \mathrm{d}\mu(s), \quad t \ge 0,$$

where μ is a finite positive Borel measure on $\mathbb{R}_+$ and

$$\Omega_d(r) := \frac{\int_0^\pi e^{ir\cos\theta}\sin^{d-2}\theta\,\mathrm{d}\theta}{\int_0^\pi \sin^{d-2}\theta\,\mathrm{d}\theta}, \quad d \geq 2,\ r \geq 0.$$

Setting $d = 3$ leads to

$$\Omega_3(r) = \frac{\sin r}{r}.$$

We then choose the measure μ such that $\operatorname{supp}\mu = [0, 2\pi]$ and $\mathrm{d}\mu(s) = s\,\mathrm{d}s$ for $s \in [0, 2\pi]$. The resulting function ϕ is

$$\phi(t) = \int_0^{2\pi} \frac{\sin st}{st}\,\mathrm{d}\mu(s) = \frac{1 - \cos 2\pi t}{t^2}, \quad t \geq 0,$$

which is not decreasing since $\phi(1) = 0$ while $\phi(\frac{3}{2}) = \frac{8}{9}$. Therefore by Theorem 2.4 $K(x, y) = \phi(\|x - y\|)$ is not admissible for any class of radially decreasing distributions. In particular, it is not admissible for the Gaussian distributions.

Nevertheless, there exists a large class of compactly supported decreasing ϕ so that (12) defines a kernel that satisfies the conditions in Theorem 2.4. These are the compactly supported radial basis functions of minimal degree constructed in Wendland (2005) and Wu (1995). Examples for dimension $d = 3$ include

$$\phi(r) := (1 - r)_+^3 \text{ and } \phi(r) := (1 - r)_+^4(1 + 4r), \quad r \in \mathbb{R}_+,$$

where $(1 - r)_+ := \max\{0, 1 - r\}$. More examples are available in (Wendland 2005, Chapter 9).

Going back to our main objective, we shall use Theorem 2.4 to establish admissibility for the RKHS embedding of Gaussian distributions.

Theorem 2.5 *Let K be a reproducing kernel on $\mathbb{R}^d$ of the form $K(x, y) = \psi(x - y) = \phi(\|x - y\|)$ where ϕ is deceasing on $\mathbb{R}_+$ and such that ρ in (4) is supported on all of $\mathbb{R}^d$. Then K is admissible for the class of Gaussian distributions*

$$\mathrm{d}\mathbb{P}_{\mu,\sigma}(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left(-\frac{\|x - \mu\|^2}{2\sigma^2}\right)\mathrm{d}x, \quad x \in \mathbb{R}^d, \quad \mu \in \mathbb{R}^d, \sigma > 0 \quad (13)$$

and for the class of generalized Gaussian distributions

$$\mathrm{d}E_{\mu,\sigma}(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d c_\omega \int_0^\infty \exp\left(-\frac{\|x - \mu\|^2}{2\tau\sigma^2}\right)\mathrm{d}\omega(\tau)\,\mathrm{d}x,$$
$$x \in \mathbb{R}^d, \quad \mu \in \mathbb{R}^d, \sigma > 0 \quad (14)$$

where ω is a nontrivial finite positive Borel measure on $\mathbb{R}_+$ with $\operatorname{supp}\omega \neq \{0\}$*, and* c_w *is a positive constant such that*

$$c_\omega \int_0^\infty \tau^{\frac{d}{2}}\, \mathrm{d}\omega(\tau) = 1.$$

Proof It suffices to verify that the conditions in Theorem 2.4 is satisfied. Firstly, $K(x, y) = \psi(x - y)$ where $\psi(x) = \phi(\|x\|)$ is radial and radially decreasing. For the Gaussian distributions, we see that

$$\mathrm{d}\mathbb{P}_{\mu,\sigma}(x) = \mathrm{d}\mathbb{Q}\left(\frac{x-\mu}{\sigma}\right)$$

where $\mathbb{Q}$ has the density function

$$g(x) := \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left(-\frac{\|x\|^2}{2}\right), \quad x \in \mathbb{R}^d.$$

Thus, the density function for $\bar{\mathbb{Q}} = \mathbb{Q} * \tilde{\mathbb{Q}}$ is

$$f(x) := (g * g)(x) := \left(\frac{1}{2\sqrt{\pi}}\right)^d \exp\left(-\frac{\|x\|^2}{4}\right), \quad x \in \mathbb{R}^d,$$

which is radial and strictly radially decreasing. Therefore K is admissible for the Gaussian distributions (13).

For the generalized Gaussian distributions (14), the density function of $\bar{\mathbb{Q}}$ is

$$g(x) := \left(\frac{1}{\sqrt{2\pi}}\right)^d c_\omega^2 \int_0^\infty \int_0^\infty \sqrt{\frac{ts}{t+s}} \exp\left(-\frac{\|x\|^2}{2(t+s)}\right) \mathrm{d}\omega(s)\, \mathrm{d}\omega(t), \quad x \in \mathbb{R}^d,$$

which is also radial and strictly radially decreasing. □

Generalized Gaussian distributions have found many applications in image processing (Mallat 1989; Moulin and Liu 1999) and the field of engineering (Miller and Thomas 1972; Beaulieu and Young 2009). Classical probability density functions for generalized Gaussian distributions are of the following form

$$f_p(x) = \frac{c_p}{\sigma^d} \exp\left(-\frac{\|x-\mu\|^p}{2\sigma^p}\right), \quad x \in \mathbb{R}^d,$$

where $p > 0$ and c_p is the constant that makes $f_p(x)$ a density function. The existence of the measure $\omega(\tau)$ can be guaranteed by theoretic results in Bochner (1937).

Specifying the measure ω, we have the following examples.

Example 2.6 Let the Borel measure $\mathrm{d}\omega(\tau) = \mathrm{d}\tau/\tau^2,\ \tau \in [1, 2]$ in (14). Then the corresponding generalized Gaussian distribution is

$$\begin{aligned}\mathrm{d}E_{\mu,\sigma}(x) = \frac{2c_\omega}{(2\pi)^{d/2}\sigma^{d-2}\|x-\mu\|^2}\left(\exp\left(-\frac{\|x-\mu\|^2}{4\sigma^2}\right)\right.\\ \left.-\exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right)\right)\mathrm{d}x, \quad x \in \mathbb{R}^d.\end{aligned}$$

Let the Borel measure $\omega = \sum_{i=1}^m \alpha_i\delta_{\tau_i}$, where $\alpha_i \geq 0$ and δ_{τ_i} is the dirac measure at point $\tau_i > 0$. Then the corresponding generalized Gaussian distribution is simply the linear combination of Gaussian distributions with the same expectation. That is,

$$\mathrm{d}E_{\mu,\sigma}(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d c_\omega \sum_{i=1}^m \alpha_i \exp\left(-\frac{\|x-\mu\|^2}{2\tau_i\sigma^2}\right)\mathrm{d}x, \quad x \in \mathbb{R}^d.$$

In particular, the Wendland functions (Wendland 2005) and the Gaussian kernels are admissible for the Gaussian distributions. We remark that the latter observation can also be made from direct computation as done in Sriperumbudur et al. (2009), where it was shown that for two Gaussian distributions $\mathbb{P}_{\mu,\sigma}$, $\mathbb{P}_{\nu,\theta}$ and for the Gaussian kernel

$$K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\tau^2}\right), \quad x, y \in \mathbb{R}^d, \tau > 0,$$

it holds

$$\begin{aligned}\gamma_K^2(\mathbb{P}_{\mu,\sigma}, \mathbb{P}_{\nu,\theta}) = \left(\frac{\tau}{\sqrt{2\sigma^2+\tau^2}}\right)^d\\ +\left(\frac{\tau}{\sqrt{2\theta^2+\tau^2}}\right)^d - 2\prod_{i=1}^d \frac{\tau\exp\left(-\frac{(\mu_i-\nu_i)^2}{2(\sigma^2+\theta^2+\tau^2)}\right)}{\sqrt{\sigma^2+\theta^2+\tau^2}}.\end{aligned}$$

Thus, when $\sigma = \theta$,

$$\gamma_K^2(\mathbb{P}_{\mu,\sigma}, \mathbb{P}_{\nu,\sigma}) = 2\left(\frac{\tau}{\sqrt{2\sigma^2+\tau^2}}\right)^d\left(1-\exp\left(-\frac{\|\mu-\nu\|^2}{2(2\sigma^2+\tau^2)}\right)\right). \qquad (15)$$

Clearly, the two admissibility requirements are satisfied.

3 Non-radial Gaussian distributions

In this section, we study similarity of two multivariate non-radial Gaussian distributions under RKHS embedding. Such distributions appear widely in probability and statistics. They are of the general form

$$\mathrm{d}\mathbb{P}_{\mu,\Sigma}(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^{d} (\det \Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \mathrm{d}x, \quad x \in \mathbb{R}^d \tag{16}$$

where $\mu \in \mathbb{R}^d$ and Σ is a radial and positive-definite $d \times d$ matrix. To fulfill our purpose, we need to introduce the definition of multivariate monotonic functions from (Engelking 1989).

Definition 3.1 Let h be a function from $\mathbb{R}^n$ to $\mathbb{R}^k$. We say that h is *monotonic* provided that for any $y \in \mathbb{R}^k$, $h^{-1}(y)$ is connected in $\mathbb{R}^n$.

Recall that a set C in $\mathbb{R}^n$ is *connected* if there do not exist two disjoint open subsets $U, V \in \mathbb{R}^n$ such that $C \subseteq U \cup V$ and both $C \cap U$ and $C \cap V$ are nonempty.

Obviously, a constant function f on $\mathbb{R}^n$ is monotonic as for every $c \in \mathbb{R}$, $f^{-1}(c)$ is either empty or the entire $\mathbb{R}^n$. We give some other nontrivial examples of multivariate monotonic functions to help comprehend this definition.

We first point out that the above seemingly abstract definition coincides with the ordinary one for continuous univariate monotonic functions.

Example 3.2 Let f be a continuous function on $\mathbb{R}$. If f is monotonic in the ordinary sense then it is easy to see that it satisfies Definition 3.1. On the other hand, assume that it is monotonic according to Definition 3.1. In other words, for each $c \in \mathbb{R}$, $f^{-1}(c)$ is connected in $\mathbb{R}$. One shows by the intermediate value theorem for continuous functions that f must be monotonic in the ordinary sense.

Example 3.3 A linear function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) := a_1x_1 + a_2x_2 + \cdots + a_nx_n$, where $a_i \in \mathbb{R}$ is a monotonic function. This is because for all $y \in \mathbb{R}$, $f^{-1}(y) = \{x \in \mathbb{R}^n : \sum_{i=1}^n a_ix_i = y\}$ is a hyperplane in $\mathbb{R}^n$, which is connected in $\mathbb{R}^n$.

The following example will appear in our discussion.

Lemma 3.4 *Let $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \exp(\alpha_1x_1^2+\alpha_2x_2^2+\cdots+\alpha_nx_n^2)$, where α_i are simultaneously all negative or all positive. Then f is a monotonic function.*

Proof Without loss of generality, we assume $\alpha_1, \alpha_2, \ldots, \alpha_n > 0$. Then $f(x) \geq 1$ and $f^{-1}(1) = \{0\}$, which is a connected set. For $c > 1$, the set $f^{-1}(c) = \{(x_1, x_2, \ldots, x_n) : \alpha_1x_1^2 + \alpha_2x_2^2 + \alpha_nx_n^2 = \ln c\}$ is an n-dimensional ellipsoid, which is connected in $\mathbb{R}^n$. Therefore we have by Definition 3.1 that f is a monotonic function. □

Before proceeding to the next lemma, two classical results are needed, see van Mill (1989).

1. A continuous injective function f from a compact subset of $\mathbb{R}^n$ to $\mathbb{R}^n$ is a homeomorphism (van Mill 1989, Excise 1.1.4).
2. The Brouwer Invariance of Domain Theorem: If two sets $X, Y \subseteq \mathbb{R}^n$ are homeomorphic then so are their interiors (van Mill 1989, Theorem 4.6.7, Corollary 4.6.6).

Lemma 3.5 *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuous monotonic function and $G : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous injective mapping. Then their composition $F \circ G : \mathbb{R}^n \to \mathbb{R}$ is monotonic.*

Proof It is easy to see that $F \circ G$ is a continuous function. By definition, we have to show that for every $c \in \mathbb{R}$, the set $(F \circ G)^{-1}(c)$ is connected in $\mathbb{R}^n$. Since F is monotonic, we know that $F^{-1}(c)$ is a connected set in $\mathbb{R}^n$. By the fact that the continuous image of a connected set is still connected, the set $(F \circ G)^{-1}(c) = G^{-1}(F^{-1}(c))$ is connected provided that G^{-1} is continuous.

We then prove that if G is continuous and injective, then its inverse G^{-1} is a continuous function from $G(\mathbb{R}^n)$ to $\mathbb{R}^n$. For any sequence $\{y_k : k = 1, 2, \ldots\} \subseteq G(\mathbb{R}^n)$ that converges to $y_0 \in G(\mathbb{R}^n)$, denote their preimages by $\{x_k : k = 1, 2, \ldots\}$ and x_0, respectively. Let B_r be a closed ball with radius r centered at x_0, which is clearly compact. Using the result 1 above we have the fact that B_r must be homeomorphic to $G(B_r)$. Then we have by the Brouwer Invariance of Domain Theorem that $\operatorname{int} B_r$ and $\operatorname{int} G(B_r)$ are homeomorphic. Therefore for sufficiently large k, y_k must be located in $\operatorname{int} G(B_r)$. Again by the fact that G is injective, x_k must be contained in B_r for all sufficient large k. Since the radius r is arbitrary, we have that x_k converges to x_0, namely, G^{-1} is continuous. □

We are ready to present the main result of this section about the similarity of two general d-dimensional Gaussian distributions.

Theorem 3.6 *The Gaussian reproducing kernel $K(x, y) = \exp(-\frac{\|x-y\|_2^2}{2\tau^2})$, $\tau > 0$ is admissible for the class of d-dimensional Gaussian distributions given by (16) in the sense that*

(A1') *As a function of μ, ν, $\gamma_K(\mathbb{P}_{\mu,\Sigma}, \mathbb{P}_{\nu,\Sigma})$ decreases monotonically to 0 as μ tends to ν,*

(A2') *As a function of the eigenvalues of Σ^{-1}, $\gamma_K(\mathbb{P}_{\mu,\Sigma}, \mathbb{P}_{\nu,\Sigma})$ decreases monotonically to 0 as $\det \Sigma$ tends to infinity.*

Proof Let $\mathbb{P}_1$, $\mathbb{P}_2$ be two d-dimensional Gaussian distributions given by

$$\mathrm{d}\mathbb{P}_1(x) = \frac{1}{(2\pi)^{\frac{d}{2}}(\det \Sigma_1)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma_1^{-1}(x-\mu)\right) \mathrm{d}x,$$
$$\mathrm{d}\mathbb{P}_2(x) = \frac{1}{(2\pi)^{\frac{d}{2}}(\det \Sigma_2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\nu)^T \Sigma_2^{-1}(x-\nu)\right) \mathrm{d}x,$$

where Σ_1, Σ_2 are two positive-definite matrices.

If $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$ then there exists an orthogonal $d \times d$ matrix B such that

$$\Lambda_1 = B^T \Sigma_1^{-1} B = \begin{pmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \ddots & \\ & & & \alpha_d \end{pmatrix}, \quad \Lambda_2 = B^T \Sigma_2^{-1} B = \begin{pmatrix} \beta_1 & & & \\ & \beta_2 & & \\ & & \ddots & \\ & & & \beta_d \end{pmatrix},$$

where $\alpha_i, \beta_i, i = 1, 2, \ldots, d$ are all positive. Then

$$\begin{aligned}
&\iint_{\mathbb{R}^d \times \mathbb{R}^d} K(x, y) \, \mathrm{d}\mathbb{P}_1(x) \, \mathrm{d}\mathbb{P}_1(y) \\
&= \frac{1}{(2\pi)^d \det \Sigma_1} \iint_{\mathbb{R}^d \times \mathbb{R}^d} K(x, y) \\
&\quad \exp\left\{-\frac{1}{2}\left[(x-\mu)^T \Sigma_1^{-1}(x-\mu) + (y-\mu)^T \Sigma_1^{-1}(y-\mu)\right]\right\} \mathrm{d}x \, \mathrm{d}y \\
&\overset{(\spadesuit)}{=} \frac{1}{(2\pi)^d \det \Sigma_1} \iint_{\mathbb{R}^d \times \mathbb{R}^d} \\
&\quad \exp\left(-\frac{\|x-y\|_2^2}{2\tau^2}\right) \exp\left\{-\frac{1}{2}(x^T \Sigma_1^{-1} x + y^T \Sigma_1^{-1} y)\right\} \mathrm{d}x \, \mathrm{d}y \\
&\overset{(\heartsuit)}{=} \frac{1}{(2\pi)^d \det \Sigma_1} \iint_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left(-\frac{\|x-y\|_2^2}{2\tau^2}\right) \\
&\quad \exp\left\{-\frac{1}{2}(x^T \Lambda_1 x + y^T \Lambda_1 y)\right\} \mathrm{d}x \, \mathrm{d}y \\
&= \frac{1}{(\det \Sigma_1)^{\frac{1}{2}}} \prod_{i=1}^{d} \frac{\tau}{\sqrt{2 + \alpha_i \tau^2}},
\end{aligned}$$

the equality ($\spadesuit$) holds since the kernel $K(x, y)$ is translation invariance, and ($\heartsuit$) follows if we replace x and y with Bx and By, respectively.

Similarly, we have

$$\iint_{\mathbb{R}^d \times \mathbb{R}^d} K(x, y) \, \mathrm{d}\mathbb{P}_2(x) \, \mathrm{d}\mathbb{P}_2(y) = \frac{1}{(\det \Sigma_2)^{\frac{1}{2}}} \prod_{i=1}^{d} \frac{\tau}{\sqrt{2 + \beta_i \tau^2}}$$

and

$$\iint_{\mathbb{R}^d \times \mathbb{R}^d} K(x, y) \, \mathrm{d}\mathbb{P}_1(x) \, \mathrm{d}\mathbb{P}_2(y) = \frac{1}{(\det(\Sigma_1 \Sigma_2))^{\frac{1}{2}}} \prod_{i=1}^{d} \frac{\tau e^{-\frac{\alpha_i \beta_i (\mu_i - \nu_i)^2}{2(\alpha_i + \beta_i + \alpha_i \beta_i \tau^2)}}}{\sqrt{\alpha_i + \beta_i + \alpha_i \beta_i \tau^2}}.$$

As a result,

$$\begin{aligned}
\gamma_K^2(\mathbb{P}_1, \mathbb{P}_2) &= \frac{1}{(\det \Sigma_1)^{\frac{1}{2}}} \prod_{i=1}^{d} \frac{\tau}{\sqrt{2 + \alpha_i \tau^2}} + \frac{1}{(\det \Sigma_2)^{\frac{1}{2}}} \prod_{i=1}^{d} \frac{\tau}{\sqrt{2 + \beta_i \tau^2}} \\
&\quad - \frac{2}{(\det(\Sigma_1 \Sigma_2))^{\frac{1}{2}}} \prod_{i=1}^{d} \frac{\tau e^{-\frac{\alpha_i \beta_i (\mu_i - \nu_i)^2}{2(\alpha_i + \beta_i + \alpha_i \beta_i \tau^2)}}}{\sqrt{\alpha_i + \beta_i + \alpha_i \beta_i \tau^2}}.
\end{aligned}$$

In particular, we have

$$\gamma_K^2(\mathbb{P}_{\mu,\Sigma},\mathbb{P}_{\nu,\Sigma}) = \frac{2\tau^d}{(\det\Sigma)^{\frac{1}{2}}}\frac{1-\prod_{i=1}^d e^{-\frac{\alpha_i(\mu_i-\nu_i)^2}{2(2+\alpha_i\tau^2)}}}{\prod_{i=1}^d\sqrt{2+\alpha_i\tau^2}}, \tag{17}$$

where the α_i's are the eigenvalues of Σ^{-1}. Based on this formula, we shall verify that the properties **(A1')** and **(A2')** hold true.

It is easy to see that $\gamma_K(\mathbb{P}_{\mu,\Sigma},\mathbb{P}_{\nu,\Sigma}) \to 0$ as $\|\mu-\nu\| \to 0$. To show the monotonicity of γ_K with respect to $\mu-\nu$, we just have to verify that

$$D(x) := \prod_{i=1}^d \exp\left(-\frac{\alpha_i x_i^2}{2(2+\alpha_i\tau^2)}\right),\ x\in\mathbb{R}^d$$

is monotonic on $\mathbb{R}^d$ according to Definition 3.1. This falls into Lemma 3.4. Thus, **(A1')** is verified.

Next, let μ,ν be fixed. By Formula (17), $\gamma_K(\mathbb{P}_{\mu,\Sigma},\mathbb{P}_{\nu,\Sigma}) \to 0$ as $\det\Sigma\to\infty$. It remains to show that γ_K is monotonic with respect to $(\alpha_1,\alpha_2,\ldots,\alpha_d)$. Set

$$g(s) := \sqrt{\frac{s}{2+\tau^2 s}},\ s\geq 0 \text{ and } G(t) := (g(t_1), g(t_2),\ldots,g(t_d)),\ t\in\mathbb{R}_+^d.$$

Then g is continuous and strictly increasing on $\mathbb{R}_+$. Therefore G is continuous and injective on $\mathbb{R}^n$. Denote

$$F_c(t_1,t_2,\ldots,t_d) = t_1t_2\cdots t_d(1-\exp[-(c_1t_1^2+c_2t_2^2+\cdots+c_dt_d^2)]),$$

where $c=(c_1,c_2,\ldots,c_d)$ and $t=(t_1,t_2,\ldots,t_d)$ are both in $\mathbb{R}_+^d$. Then by Formula (17) we have

$$\gamma_K^2(\mathbb{P}_{\mu,\Sigma},\mathbb{P}_{\nu,\Sigma}) = 2\tau^d F_\omega(G(\alpha_1,\alpha_2,\ldots,\alpha_d))$$

where $\omega = (-\frac{(\mu_1-\nu_1)^2}{2}, -\frac{(\mu_2-\nu_2)^2}{2},\ldots,-\frac{(\mu_d-\nu_d)^2}{2})$. For every $s\in\mathbb{R}_+$, the set

$$\begin{aligned}F_\omega^{-1}(s) &= \{x\in\mathbb{R}_+^d : F_\omega(x)=s\}\\ &= \{(x_1,\ldots,x_d) : (x_1\cdots x_d)\\ &\qquad \left(1-\prod_{i=1}^d \exp\left[-\left(\frac{(\mu_i-\nu_i)^2}{2}x_i^2\right)\right]\right) = s\right\}\end{aligned}$$

is connected in $\mathbb{R}_+^d$ (Please see the proof in the Appendix). By Lemma 3.5, $2\tau^d F_\omega\circ G(\alpha_1,\ldots,\alpha_d)$ is monotonic on α, which confirms **(A2')**. □

4 Conclusion

Measuring the similarity and distance between two probability distributions is important in many applications of statistics. The approach of RKHS embedding has many advantages over other integral probability metrics. Due to the one-to-one correspondence between reproducing kernels and reproducing kernel Hilbert spaces, the choice of the reproducing kernel is critical in the approach. Past studies have been focusing on when the kernel is characteristic. We investigate an admissibility criterion on the kernel to ensure that the similarity among the RKHS embeddings of the same class of distributions would satisfy two natural requirements. Sufficient and necessary conditions are provided. In particular, we find that radially decreasing radial basis functions are admissible for Gaussian distributions. We remark that the study can be extended to other classes of probability distributions and to other norms on the Euclidean space.

Appendix

We now prove the connectivity of the set $F_\omega^{-1}(s)$ in Theorem 3.6.

For if $\omega = \mathbf{0}$, then $F_{\mathbf{0}}(x_1, x_2, \ldots, x_d) = 0$ for any $x \in \mathbb{R}_+^d$. Therefore, $F_{\mathbf{0}}^{-1}(0) = \mathbb{R}_+^d$ and for any $s \neq 0$, $F_{\mathbf{0}}^{-1}(s)$ is an empty set. In both cases, the set $F_{\mathbf{0}}^{-1}(s)$ is connected.

If $\omega \neq \mathbf{0}$, then there exists at least one coordinate $\omega_i \neq 0$. For the case $s = 0$, it is easy to see that the set

$$F_\omega^{-1}(0) = \{(x_1, x_2, \ldots, x_d) \in \mathbb{R}_+^d : \text{at least one coordinate } x_i \text{ equals to zero}\}$$

is clearly a connected subset in $\mathbb{R}_+^d$.

Now assume $s > 0$, by scaling and permuting the coordinates we can assume further that $\omega_1 = \cdots = \omega_n = 1$ and $\omega_{n+1} = \cdots = \omega_d = 0$, where $n \leq d$. We then use the polar coordinates to simplify the problem. That is, let

$$\begin{aligned} x_1 &= r\cos\theta_1 \\ x_2 &= r\sin\theta_1\cos\theta_2 \\ &\vdots \\ x_{n-1} &= r\sin\theta_1\sin\theta_2\cdots\sin\theta_{n-2}\cos\theta_{n-1} \\ x_n &= r\sin\theta_1\sin\theta_2\cdots\sin\theta_{n-2}\sin\theta_{n-1} \end{aligned}$$

where $r > 0$ and $\theta_1, \ldots, \theta_{n-1} \in (0, \pi/2)$. Then we have

$$F_\omega^{-1}(s) = \Big\{x = (x_1, \ldots, x_d) \in \mathbb{R}_+^d : (r^n\Theta(\theta_1, \ldots, \theta_{n-1}) \\ \cdot x_{n+1}\cdots x_d)\left(1 - e^{-r^2}\right) = s\Big\},$$

where $\Theta(\theta_1, \ldots, \theta_{n-1}) = \sin^{n-1}\theta_1 \sin^{n-2}\theta_2 \cdots \sin\theta_{n-1} \cdot \cos\theta_1 \cdots \cos\theta_{n-1}$.

For every fixed point $(\theta_1, \ldots, \theta_{n-1}) \in (0, \pi/2)^{n-1}$ and fixed product $p = x_{n+1} \cdots x_d$, there exists a unique solution $r_s = r(\theta_1, \ldots, \theta_{n-1}, p) > 0$ to the equation

$$(\Theta(\theta_1, \ldots, \theta_{n-1}) \cdot p) r_s^n \left(1 - e^{-r_s^2}\right) = s.$$

The inverse mapping theorem shows that $r(\theta_1, \ldots, \theta_{n-1}, p)$ is a continuous function with respect to the variables $\theta_1, \ldots, \theta_{n-1}, p$. We now show that this implies that the following set

$$F_\omega^{-1}(s) = \left\{x(r_s, \theta_1, \ldots, \theta_{n-1}) \in \mathbb{R}_+^d : \theta_i \in (0, \pi/2), x_{n+1} \cdots x_d = p\right\}$$

is path connected, and therefore connected. Indeed, for any two distinct points $a, b \in F_\omega^{-1}(s)$, $s > 0$, assume their corresponding polar coordinates (without the r coordinate) are $\alpha = (\theta_1', \ldots, \theta_{n-1}', a_{n+1}, \ldots, a_d)$ and $\beta = (\theta_1'', \ldots, \theta_{n-1}'', b_{n+1}, \ldots, b_d)$, respectively. Let $\gamma_t = (1-t)\alpha + t\beta$. Then by the definition of r_s, we have for every $t \in [0, 1]$, $(r_s(\gamma_t), \gamma_t) \in F_\omega^{-1}(s)$, and therefore $(r_s(\gamma_t), \gamma_t)$, $t \in [0, 1]$ is a path from a to b.

References

Aronszajn N (1950) Theory of reproducing kernels. Trans Am Math Soc 68:337–404

Beaulieu NC, Young DJ (2009) Designing time-hopping ultrawide bandwidth receivers for multiuser interference environments. Proc IEEE 97(2):255–284

Berlinet A, Thomas-Agnan C (2004) Reproducing kernel Hilbert spaces in probability and statistics. Kluwer, Dordrecht

Bochner S (1937) Stable laws of probability and completely monotone functions. Duke Math J 3(4):726–728

Bochner S (1959) Lectures on Fourier integrals with an author's supplement on monotonic functions, Stieltjes integrals, and harmonic analysis. Annals of mathematics studies, vol 42. Princeton University, New Jersey

Chen W, Wang B, Zhang H (2016) Universalities of reproducing kernels revisited. Appl Anal 95:1776–1791

Dudley RM (2002) Real analysis and probability. Cambridge University Press, Cambridge, UK

Engelking R (1989) Gerneral topology, 2nd edn. Heldermann-Verlag, Berlin

Fukumizu K, Gretton A, Sun X, Schölkopf B (2008) Kernel measures of conditional dependence. In: Advances in neural information processing systems, vol 20. MIT Press, Cambridge, pp 489–496

Fukumizu K, Bach FR, Jordan MI (2009) Kernel dimension reduction in regression. Ann Stat 37:1871–1905

Gretton A, Borgwardt K, Rasch B, Schölkopf B, Smola A (2007) A kernel methods for the two sample problem. In: Advances in neural information processing systems, vol 19. MIT Press, Cambridge, pp 513–520

Lieb EH, Loss M (2001) Analysis. American Mathematical Society, New York

Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11(7):674–693

Miller J, Thomas JB (1972) Detectors for discrete-time signals in non-Gaussian noise. IEEE Trans Inf Theory 18(2):241–250

Moulin P, Liu J (1999) Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors. IEEE Trans Inf Theory 45(3):909–919

Müller A (1997) Integral probability metrics and their generating classes of functions. Adv Appl Probab 29:429–443

Rachev ST (1991) Probability metrics and the stability models. Wiley, Chichester

Schoenberg IJ (1938) Metric spaces and completely monotone functions. Ann. Math. (2) 39:811–841

Shorack GR (2000) Probability for statisticians. Springer, New York

Smola AJ, Gretton A, Song L, Schölkopf B (2007) A Hilbert space embedding for distributions. In: Proc. 18th international conference on algorithmic learning theory. Springer, Berlin, pp 13–31

Song G, Zhang H, Hickernell FJ (2013) Reproducing kernel banach spaces with the ℓ^1 norm. Appl Comput Harmon Anal 34:96–116

Sriperumbudur BK, Gretton A, Fukumizu K, Schölkopf B, Lanckriet GRG (2009) On integral probability metrics, ϕ-divergences and binary classification. Computing Research Repository. arXiv: 0901.2698v4

Sriperumbudur BK, Gretton A, Fukumizu K, Schölkopf B, Lanckriet GRG (2010) Hilbert space embeddings and metrics on probability measures. J Mach Learn Res 11:1517–1561

Sriperumbudur BK, Fukumizu K, Lanckriet GRG (2011) Learning in Hilbert vs. Banach spaces: a measure embedding viewpoint. In: Advances in neural information processing systems, vol 24. MIT Press, pp 1773–1781

Sriperumbudur BK, Fukumizu K, Lanckriet GRG (2011) Universality, characteristic kernels and RKHS embedding of measures. J Mach Learn Res 12:2389–2410

Steinwart I (2001) On the influence of the kernel on the consistency of support vector machines. J Mach Learn Res 2:67–93

Vajda I (1989) Theory of statistical inference and information. Kluwer Academic Publishers, Boston

van Mill J (1989) Infinite-dimensional topology, prerequisites and introduction. North-Holland math. library, vol 43. Elsevier, Amsterdam

Weaver N (1999) Lipschitz algebras. World Scientific Publishing Company, Singapore

Wendland H (2005) Scattered data approximation. Cambridge University Press, Cambridge

Wu ZM (1995) Compactly supported positive definite radial functions. Adv Comput Math 4(3):283–292

Zhang H, Zhao L (2013) On the inclusion relation of reproducing kernel Hilbert spaces. Anal Appl 11, 1350014

Zhang H, Xu Y, Zhang J (2009) Reproducing kernel Banach spaces for machine learning. J Mach Learn Res 10:2741–2775

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.