IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 25 February 2024, accepted 11 April 2024, date of publication 17 April 2024, date of current version 26 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3390149

THEORY

# Learning With Multiple Kernels

**MAHDI A. ALMAHDAWI [1,2] AND OMAR DE LA C. CABRERA [2]**

[1] Department of GRC, Applied College, King Abdulaziz University, Jeddah 21589, Saudi Arabia

[2] Department of Mathematical Sciences, College of Arts and Sciences, Kent State University, Kent, OH 44242, USA

Corresponding author: Mahdi A. Almahdawi (melmehdwy@kau.edu.sa)

**ABSTRACT** Over the last decades, learning methods using kernels have become very popular. The main reason is that real data analysis often requires nonlinear methods to detect the dependencies that allow successful predictions of properties of interest. Gaussian kernels have been used in many studies such as learning algorithms and data analysis. Most of these studies have shown that the parameter chosen for a Gaussian kernel could have a huge impact on the desired results. Therefore, it is essential to understand this impact on a theoretical level. The main contribution of this paper is to study the effect of the Gaussian kernel bandwidth parameter on how well an empirical operator defined from data approximates its continuous counterpart. Some results in spectral approximations are provided as well as some examples.

**INDEX TERMS** Gaussian kernel, radial kernels, kernel principal component analysis, reproducing kernel, support vector machine.

## I. INTRODUCTION

Gaussian kernels are one of the most popular choices in kernel methods. They can perform very efficiently in many learning algorithms such as support vector machines, and kernel principal component analysis (kernel PCA), when the appropriate requirements are met. Most kernels of interest are actually families of kernels, usually depending on one parameter that controls the "bandwidth," i.e., how "wide" is the kernel. A narrow bandwidth allows a kernel to distinguish between very close inputs, while farther inputs are all regarded as essentially infinitely far; a wide bandwidth allows distinctions between different levels of "farness" but loses granularity at close distances. Picking a bandwidth, in a way, selects the scale at which we will be able to study phenomena, see [1], [2], [3], [4], and [5]. One of the central points of this paper is how to study learning with kernels when we use kernels at different bandwidths at the same time, which is necessary when the data present different phenomena at different scales.

Our contribution in this paper is to find a theoretical technique that helps choosing such a good parameter. Therefore, we start by presenting some recent results on Reproducing Kernel Hilbert Spaces (RKHSs) of Gaussian kernels such as I. Steinnwart, and C. Scoval. A combination of these results and Rosasco's results, see [6], [7], and [8], we obtained a new bound shows the effect of the Gaussian kernel bandwidth parameter on how well an empirical operator defined from data approximates its continuous counterpart.

## II. INTEGRAL OPERATORS DEFINED BY GAUSSIAN REPRODUCING KERNEL

Gaussian kernel is one of the most popular and used kernels in learning algorithms such as Kernel PCA, clustering, and many other problems and Algorithms that make the use of kernels crucial. The choice of the parameter of Gaussian kernel could have a huge impact on these algorithms. For that reason, our contribution is to study the role of these parameters theoretically. In 2010, Rosasco considered the case when we have any positive kernel and made some bounds on eigenvalues and spectral projections, but he did not study the kernel parameter impact, see [7]. In our case, our results will be focused on on Gaussian kernels. We make some bounds that show the impact of such parameters.

First of all, let us assume that $X \subset \mathbb{R}^d$ and $k_\gamma : X \times X \to \mathbb{C}$ is the Gaussian reproducing kernel defined by

$$k_\gamma(x,t) = e^{-\frac{\|x-t\|^2}{\gamma^2}}, \; x, t \in X.$$

The associate editor coordinating the review of this manuscript and approving it for publication was Jianxiang Xi.

Let $p$ be a probability measure on $X$ and $\mathcal{L}^2(X,p)$ is the space of square integrable functions with norm

$$\|f\|_{\mathcal{L}^2(X,p)} = \langle f,f\rangle_{\mathcal{L}^2(X,p)} = \int_X |f(x)|^2 dp(x).$$

Let $L_{k_\gamma} : \mathcal{L}^2(X,p) \to \mathcal{L}^2(X,p)$ be an integral operator defined by

$$(L_{k_\gamma} f)(x) = \int_X e^{-\frac{\|x-t\|^2}{\gamma^2}} f(t)dp(t)$$

for all $x, t \in X$, $\gamma \in \mathbb{R}^+$, and $f \in \mathcal{L}^2(X,p)$. We know that

$$k_\gamma(x,t) = e^{-\frac{\|x-t\|^2}{\gamma^2}} \leq 1$$

for all $x, t \in X$. Therefore, $L_{k_\gamma}$ is a bounded operator, see [2], [6], [8], and [9].

Assume that we are given a set of points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ sampled i.id. according to $p$. The $n \times n$-kernel matrix $K_n$ is given by

$$K_{i,j} = \frac{1}{n} e^{-\frac{\|x_i-x_j\|^2}{\gamma^2}}.$$

Let $\mathbb{H}_\gamma(X)$ be the Gaussian reproducing kernel Hilbert space and define the operators $T_{\mathbb{H}}, T_n : \mathbb{H}_\gamma(X) \to \mathbb{H}_\gamma(X)$ by

$$T_{\mathbb{H}} = \int_X \langle \cdot, e^{-\frac{\|x-\cdot\|^2}{\gamma^2}} \rangle_{\mathbb{H}_\gamma(X)} e^{-\frac{\|x-\cdot\|^2}{\gamma^2}} dp(x), \tag{1}$$

$$T_n = \frac{1}{n}\sum_{i=1}^n \langle \cdot, e^{-\frac{\|x_i-\cdot\|^2}{\gamma^2}} \rangle_{\mathbb{H}_\gamma(X)} e^{-\frac{\|x_i-\cdot\|^2}{\gamma^2}}. \tag{2}$$

Let $i_n$ be the inclusion map $\mathbb{H}_\gamma(X) \hookrightarrow \mathcal{L}^2(X,p)$, then $i_n^*$ is its adjoint operator. The following proposition is a key in our results.

*Proposition 1:* Assume that $X = \mathbb{R}$, and $p(x)$ is a normal distribution with a density

$$\phi_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}.$$

Under the above assumptions, the operator $T_{\mathbb{H}}$ is a Hilbert-Schmidt operator. In particular,

$$\|T_{\mathbb{H}}\|_{HS}^2 = \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma^2}}}.$$

*Proof:* See Appendix A ■

*Corollary 1:* If $X = \mathbb{R}^d$, then it is also

$$\|T_{\mathbb{H}}\|^2 = \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma^2}}}.$$

The following theorem will provide a new boundness depends on $\sigma$ and the parameter $\gamma$ for the difference $T_{\mathbb{H}} - T_n$.

*Theorem 1:* $T_{\mathbb{H}}$ and $T_n$ are Hilbert Schmidt operators. Under the above assumptions with probability $1 - 2e^{-\tau}$

$$\|T_{\mathbb{H}} - T_n\|_{HS} \leq \sqrt{\frac{2\tau}{n}}[1 + \left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}].$$

*Proof:* Assume that $(\xi)_{i=1}^n$ is a sequence of random variables in the Hilbert space of Hilbert-Schmidt operators defined by

$$\xi_i = \langle \cdot, e^{-\frac{\|x_i-\cdot\|^2}{\gamma^2}} \rangle_{\mathbb{H}_\gamma(\mathbb{R})} e^{-\frac{\|x_i-\cdot\|^2}{\gamma^2}} - T_{\mathbb{H}}.$$

From (1) $E(\xi_i) = 0$. By a simple computation we obtain that

$$\left\| \langle \cdot, e^{-\frac{\|x-\cdot\|^2}{\gamma^2}} \rangle_{\mathbb{H}_\gamma(\mathbb{R})} e^{-\frac{\|x-\cdot\|^2}{\gamma^2}} \right\|_{HS}^2 = \left\| e^{-\frac{\|x-\cdot\|^2}{\gamma^2}} \right\|_{\mathbb{H}_\gamma(\mathbb{R})}^4 \leq 1.$$

From the last proposition, we have

$$\|T_{\mathbb{H}}\|_{HS} = \left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}$$

and thus

$$\|\xi_i\|_{HS} \leq 1 + \left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}.$$

Using the concentration inequality, see [10], [11], [12], [13], and [14] in Hilbert spaces with confidence $1 - 2e^{-\tau}$

$$\left\| \frac{1}{n}\sum_{i=1}^n \xi_i \right\|_{HS} = \|T_{\mathbb{H}} - T_n\|_{HS} \leq \sqrt{\frac{2\tau}{n}}[1 + \left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}].$$

■

Theorem 1 shows that the Hilbert-Schmidt of the difference of the operators $T_{\mathbb{H}}$ and $T_n$ is bounded by

$$\sqrt{\frac{2\tau}{n}}[1 + \left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}].$$

This means that the operators $T_{\mathbb{H}}$ and $T_n$ become closer and closer, when the bound above becomes smaller and smaller. It is obvious that a smaller bandwidth $\gamma$ can result in a smaller bound, while a larger bandwidth $\gamma$ will result in a larger bound. The closeness of operators $T_{\mathbb{H}}$ and $T_n$ is the mathematical interpretation of using different bandwidth parameters in our experiment, which we will see in the last section.

*Corollary 2:* The same results in Theorem 1 hold true if $p$ is sub-Gaussian probability distribution.

At this point, since we have bounded the difference between the operators $T_{\mathbb{H}}$ and $T_n$, we shall be able to introduce the next proposition which gives a bound for the $\ell_2$-distance between the spectrum of the operator $K_n$ and the spectrum of the operator $L_{k_\gamma}$.

*Proposition 2:* There exists an extended enumeration $\{\sigma_j\}_{j\geq 1}$ of discrete eigenvalues for $L_{k_\gamma}$ and an extended enumeration $\{\hat{\sigma}_j\}_{j\geq 1}$ of discrete eigenvalues for $K_n$ such that

$$\sum_{j\geq 1}(\sigma_j - \hat{\sigma}_j)^2 \leq \frac{2\tau}{n}[1 + \left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}]^2,$$

with a probability greater than $1 - 2e^{-\tau}$.

*Proof:* The same technique in [6] is performed. We see that the the extended enumeration of discrete eigenvalues for $L_{k_\gamma}$ is also an extended enumeration of discrete eigenvalues for $T_{\mathbb{H}}$, and the same relationship holds for $T_n$ and $K_n$. Therefore, we obtain that

$$\sum_{j\geq 1}(\sigma_j - \hat{\sigma}_j)^2 \leq \|T_{\mathbb{H}} - T_n\|_{HS}^2$$

Now if $(\sigma_j)_{j\geq 1}$, and $(\hat{\sigma}_j)_{j\geq 1}$ are two suitable extended enumerations of discrete eigenvalues for $T_{\mathbb{H}}$ and $T_n$ respectively. From Theorem 1 we obtain

$$\sum_{j\geq 1}(\sigma_j - \hat{\sigma}_j)^2 \leq \frac{2\tau}{n}[1 + \left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}]^2$$

which proves the claim. ∎

*Theorem 2:* Let $X = \mathbb{R}^d$, and $\alpha_1 \geq \ldots \geq \alpha_s > \alpha_{s+1}$ eigenvalues for the operator $L_{k_\gamma}$. Let $s$ be the sum of the multiplicities of the first $S$ distinct eigenvalues. Let us call $\mathrm{P}_S$ to the orthogonal projection from the Hilbert space $\mathcal{L}^2(\mathbb{R}^d, p)$ onto the space spanned by the eigenfunctions corresponding to the eigenvalues $\alpha_1, \ldots, \alpha_s, \alpha_{s+1}$. Let $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_s$ be the eigenvectors of the kernel matrix $K_n$, which has the rank $r$ corresponding to the nonzero eigenvalues in a non-increasing order, and $\hat{v}_1, \ldots, \hat{v}_s \in \mathbb{H}_\gamma(\mathbb{R}^d)$ be their corresponding Nystrom extension. Assume that we have $n$ examples such that

$$n > \frac{32\tau}{(\alpha_s - \alpha_{s+1})^2}\left[1 + \left(\left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}\right)\right]^2$$

for a given $\tau > 0$, then

$$\sum_{j=1}^{s}\|(I - \mathrm{P}_S)\hat{v}_j\|_{\mathcal{L}^2(\mathbb{R}^d,p)}^2 + \sum_{j=s+1}^{r}\|\mathrm{P}_S\hat{v}_j\|_{\mathcal{L}^2(\mathbb{R}^d,p)}^2$$
$$\leq \frac{16\tau}{n(\alpha_s - \alpha_{s+1})^2}\left[1 + \left(\left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}\right)\right]^2$$

with a confidence greater than $1 - 2\,e^{-\tau}$.

*Proof:* First, we are allowed to assume that $u_1, \ldots, u_s$ are the eigenfunctions of $L_{k_n}$ with strictly positive eigenvalues $\alpha_1, \ldots, \alpha_s$ without loss of generality. Assume that we have the two families of eigenfunctions of the operator $T_{\mathbb{H}}$ the family $\{v_j\}_{j\geq 1}$, and the famliy $\hat{v}_1, \ldots, \hat{v}_r$ obtained by the Nystrom extension. Complete both families to orthonrmal basis of the RKHS $\mathbb{H}_\gamma(\mathbb{R}^d)$, and assume that we $n$ examples such that

$$n > \frac{32\tau}{(\alpha_s - \alpha_{s+1})^2}\left[1 + \left(\left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}\right)\right]^2. \tag{3}$$

From Kato's theorem,and Zwald and Blachard, see [13], [14], [15], and [16], we obtain

$$\|T_n - T_{\mathbb{H}}\|_{HS}^2 \leq \frac{2\tau}{n}\left[1 + \left(\left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}\right)\right]^2 \leq \frac{(\alpha_s - \alpha_{s+1})^2}{16}$$

$$\|\mathrm{P}^{T_n} - \mathrm{P}^{T_{\mathbb{H}}}\|_{HS}^2 \leq \frac{4}{(\alpha_s - \alpha_{s+1})^2}\|T_n - T_{\mathbb{H}}\|_{HS}^2$$
$$\leq \frac{8\tau}{n(\alpha_s - \alpha_{s+1})^2}\left[1 + \left(\left(\frac{1}{1+8\frac{\sigma^2}{\gamma^2}}\right)^{\frac{1}{4}}\right)\right]^2,$$

with high probability, where

$$\mathrm{P}^{T_{\mathbb{H}}} = \sum_{j=1}^{s}\langle\cdot, v_j\rangle_{\mathbb{H}_\gamma(\mathbb{R}^d)}v_j$$

and

$$\mathrm{P}^{T_n} = \sum_{j=1}^{s}\langle\cdot, \hat{v}_j\rangle_{\mathbb{H}_\gamma(\mathbb{R}^d)}\hat{v}_j.$$

Now we have both $\{v_i\}_{j\geq 1}$ and $\{\hat{v}_i\}_{j\geq 1}$ are orthonormal bases for $\mathbb{H}_\gamma(\mathbb{R}^d)$

$$\begin{aligned}\|\mathrm{P}^{T_n} - \mathrm{P}^{T_{\mathbb{H}}}\|_{HS}^2 &= \sum_{i,j\geq 1}|\langle \mathrm{P}^{T_n}v_i - \mathrm{P}^{T_{\mathcal{H}}}v_i, \hat{v}_j\rangle_{\mathbb{H}_\gamma(\mathbb{R}^d)}|^2\\ &= \sum_{j=1}^{s}\sum_{i\geq s+1}|\langle v_i, \hat{v}_j\rangle_{\mathbb{H}_\gamma(\mathbb{R}^d)}|^2\\ &\quad + \sum_{j\geq s+1}\sum_{i=1}^{s}|\langle v_i, \hat{v}_j\rangle_{\mathbb{H}_\gamma(\mathbb{R}^d)}|^2\\ &\geq \sum_{j=1}^{s}\sum_{\substack{i\geq s+1\\ T_{\mathbb{H}}v_i\neq o}}|\langle v_i, \hat{v}_j\rangle_{\mathbb{H}_\gamma(\mathbb{R}^d)}|^2\\ &\quad + \sum_{j\geq s+1}^{r}\sum_{i=1}^{s}|\langle v_i, \hat{v}_j\rangle_{\mathbb{H}_\gamma(\mathbb{R}^d)}|^2\end{aligned}$$

Mercer's theorem implies that $\langle v_i, \hat{v}_j\rangle_{\mathbb{H}_\gamma(\mathbb{R}^d)} = \langle v_i, \hat{v}_j\rangle_{\mathcal{L}^2(\mathbb{R}^d,p)}$, when the sum of on $i$ with respect to the eigenfunctions of $T_{\mathbb{H}}$ with nonzero eigenvalue. The last observation is that

$$\sum_{i=1}^{s}|\langle u_i, \hat{v}_j\rangle_{\mathcal{L}^2(\mathbb{R}^d,p)}|^2 = \|\mathrm{P}_S\hat{v}_j\|_{\mathcal{L}^2(\mathbb{R}^d,p)}^2$$

and

$$\begin{aligned}\sum_{\substack{i\geq s+1\\ T_{\mathbb{H}}v_i\neq 0}}|\langle u_i, \hat{v}_j\rangle_{\mathcal{L}^2(\mathbb{R}^d,p)}|^2 &= \sum_{\substack{i\geq s+1\\ L_K u_i\neq 0}}|\langle u_i, \hat{v}_j\rangle_{\mathcal{L}^2(\mathbb{R}^d,p)}|^2\\ &= \|(I - \mathrm{P}_S)\hat{v}_j\|_{\mathcal{L}^2(\mathbb{R}^d,p)}^2\end{aligned}$$

where we used that $\ker T_{\mathbb{H}} \subset \ker T_n$. Therefore, $\hat{v}_j \in \ker L_{k_\gamma}^{\perp}$ with probability 1. ■

## III. INTEGRAL OPERATORS DEFINED BY A SUM OF GAUSSIAN REPRODUCING KERNELS

In the previous section, we studied the parameter impact of only one Gaussian kernel, and how that can affect our estimation for an operator defined on the Hilbert space of square integrable functions by an operator defined on an empirical data. In this section, we will apply the same technique when we have a sum of two Gaussian kernels.

Let $X$ be a subset of $\mathbb{R}^d$, and $k_{\gamma_1}, k_{\gamma_2}$ are two Gaussian kernels with reproducing kernel Hilbert spaces $\mathbb{H}_{\gamma_1}(X), \mathbb{H}_{\gamma_2}(X)$ consecutively. It is known that the sum of two kernels is a kernel. Thus, we have $k_\gamma = k_{\gamma_1} + k_{\gamma_2}$ is a kernel, and its RKHSs $\mathbb{H}_\gamma(X)$ is given by

$$\mathbb{H}_\gamma(X) = \{f_1 + f_2 | f_1 \in \mathbb{H}_{\gamma_1}(X), f_2 \in \mathbb{H}_{\gamma_2}(X)\}$$

with a norm

$$\|f\|^2_{\mathbb{H}_\gamma(X)} = \inf_{\substack{f=f_1+f_2 \\ f_1\in\mathbb{H}_{\gamma_1}(X), f_2\in\mathbb{H}_{\gamma_2}(X)}} \left( \|f_1\|^2_{\mathbb{H}_{\gamma_1}(X)} + \|f_2\|^2_{\mathbb{H}_{\gamma_2}(X)} \right)$$

for all $f \in \mathbb{H}_\gamma(X)$, see [9], [17], and [18]. In addition, the reproducing property is defined as follows,

$$f(x) = \langle k_{\gamma_1}(x,\cdot), f_1 \rangle_{\mathbb{H}_{\gamma_1}} + \langle k_{\gamma_2}(x,\cdot), f_2 \rangle_{\mathbb{H}_{\gamma_2}},$$

$f_1 \in \mathbb{H}_{\gamma_1}(X), f_2 \in \mathbb{H}_{\gamma_2}(X)$. In particular, we have

$$\begin{aligned} \left[k_\gamma(x,x)\right]^2 &= \left\|k_\gamma(x,\cdot)\right\|^2_{\mathbb{H}_\gamma(X)} \\ &= \inf_{\substack{k_\gamma = k_{\gamma_1}+k_{\gamma_2} \\ k_{\gamma_1}\in\mathbb{H}_{\gamma_1}(X), \\ k_{\gamma_2}\in\mathbb{H}_{\gamma_2}(X)}} \left( \left|k_{\gamma_1}(x,x)\right|^2 + \left|k_{\gamma_2}(x,x)\right|^2 \right) \end{aligned}$$

The inner product of any two functions $f, g$ in $\mathbb{H}_\gamma(X)$ is given by

$$\langle f, g\rangle_{\mathbb{H}_\gamma(X)} = \inf_{\substack{f=f_1+f_2 \\ g=g_1+g_2 \\ f_1,g_1\in\mathbb{H}_{\gamma_1}(X) \\ f_2,g_2\in\mathbb{H}_{\gamma_2}(X)}} \left( \langle f_1, g_1\rangle_{\mathbb{H}_{\gamma_1}(X)} + \langle f_2, g_2\rangle_{H_{\gamma_2}(X)} \right)$$

Now we will define two operators $T_{n,\gamma}, T_{\mathbb{H},\gamma} : \mathbb{H}_\gamma(X) \to \mathbb{H}_\gamma(X)$ as follows,

$$\begin{aligned} (T_{\mathbb{H},\gamma}f)(x) &= \int_X \langle k_{\gamma_1}(x,\cdot), f_1\rangle_{\mathbb{H}_{\gamma_1}(X)} k_{\gamma_1}(x,\cdot) dp(x) \\ &\quad + \int_X \langle k_{\gamma_2}(x,\cdot), f_2\rangle_{\mathbb{H}_{\gamma_2}(X)} k_{\gamma_2}(x,\cdot) dp(x) \end{aligned}$$

for all $f \in \mathbb{H}_\gamma(X), f_1 \in \mathbb{H}_{\gamma_1}(X)$, and $f_2 \in \mathbb{H}_{\gamma_2}(X)$.

$$\begin{aligned} (T_{n,\gamma}f)(x) &= \sum_{i=1}^n \langle k_{\gamma_1}(x_i,\cdot), f_1\rangle_{\mathbb{H}_{\gamma_1}(X_i)} k_{\gamma_1}(x_i,\cdot) \\ &\quad + \sum_{i=1}^n \langle k_{\gamma_2}(x_i,\cdot), f_2\rangle_{\mathbb{H}_{\gamma_2}(X_i)} k_{\gamma_2}(x_i,\cdot) \end{aligned}$$

where $\{x_1, \ldots, x_n\}$ sampled i.i.d from $X$ with a probability $p$. The kernel matrix $K_n$ whose entry $ij$ is given by $K_{ij} = \frac{1}{n}k_\gamma(x_i, x_j)$, is an operator from $\mathbb{R}^n$ to $\mathbb{R}^n$. The last operator we need is the integral operator $L_{k_\gamma} : \mathcal{L}^2(X,p) \to \mathcal{L}^2(X,p)$, which is given by

$$(L_{k_\gamma}f)(x) = \int_X f_1(t)k_{\gamma_1}(x,t)\, dp(t) + \int_X f_2(t)k_{\gamma_2}(x,t)\, dp(t)$$

where $f, f_1, f_2 \in \mathcal{L}^2(X,p)$, and $f = f_1 + f_2$ with the norm

$$\|f\|^2_{\mathcal{L}^2(X,p)} = \inf_{\substack{f=f_1+f_2 \\ f_1,f_2\in\mathcal{L}^2(X,p)}} \left( \|f_1\|^2_{\mathcal{L}^2(X,p)} + \|f_2\|^2_{\mathcal{L}^2(X,p)} \right).$$

Again we need to bound the difference between the operators $T_{\mathbb{H},\gamma}, T_{n,\gamma}$, which we use to connect the operators $L_{k_\gamma}$ and, $K_n$. The following proposition shows a bound.

*Proposition 3:* $T_{\mathbb{H},\gamma}, T_{n,\gamma}$ are Hilbert Schmidt operators. Moreover, if $X = \mathbb{R}^d$,and $p(x)$ is a normal distribution with a density

$$\phi_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}.$$

Then, with confidence $1 - 2\, e^{-\tau}$ the following inequality holds true.

$$\begin{aligned} &\|T_{\mathbb{H},\gamma} - T_{n,\gamma}\|_{HS} \\ &\quad \leq \sqrt{\frac{2\tau}{n}} \left[ \sqrt{2} + \left( \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma_1^2}}} + \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma_2^2}}} \right)^{\frac{1}{2}} \right]. \end{aligned}$$

with a probability $1 - 2e^{-\tau}$.

*Proof:* To prove this proposition we use the same approach in Theorem 1 as well as using the following facts,

$$\begin{aligned} &\|\langle k_\gamma(x_i,\cdot), \cdot\rangle_{\mathbb{H}_\gamma(X)} k_\gamma(x_i,\cdot)\|^2_{HS} \\ &\quad \leq \|\langle k_{\gamma_1}(x_i,\cdot), \cdot\rangle_{\mathbb{H}_{\gamma_1}(X)} k_{\gamma_1}(x_i,\cdot)\|^2_{HS} \\ &\qquad + \|\langle k_{\gamma_2}(x_i,\cdot), \cdot\rangle_{\mathbb{H}_{\gamma_2}(X)} k_{\gamma_2}(x_i,\cdot)\|^2_{HS} \\ &\quad \leq 2 \\ &\|\langle k_\gamma(x_i,\cdot), \cdot\rangle_{\mathbb{H}_\gamma(X)} k_\gamma(x_i,\cdot)\|_{HS} \leq \sqrt{2} \end{aligned}$$

We have

$$\|T_{\mathbb{H},\gamma}\|^2_{HS} \leq \|T_{\mathbb{H},\gamma_1}\|^2_{HS} + \|T_{\mathbb{H},\gamma_2}\|^2_{HS}$$

where $T_{\mathbb{H},\gamma_1} : \mathbb{H}_{\gamma_1}(X) \to \mathbb{H}_{\gamma_1}(X)$, and $T_{\mathbb{H},\gamma_2} : \mathbb{H}_{\gamma_2}(X) \to \mathbb{H}_{\gamma_2}(X)$. Now we know from Proposition 1 that

$$\|T_{\mathbb{H},\gamma_1}\|^2_{HS} = \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma_1^2}}}, \quad \|T_{\mathbb{H},\gamma_2}\|^2_{HS} = \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma_2^2}}}$$

Therefore, we obtain

$$\|T_{\mathbb{H},\gamma}\|^2_{HS} \leq \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma_1^2}}} + \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma_2^2}}}$$

$$\|T_{\mathbb{H},\gamma}\|_{HS} \leq \left( \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma_1^2}}} + \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma_2^2}}} \right)^{\frac{1}{2}}.$$

Using Hoeffding inequality as in Theorem 1, we will have that

$$\|T_{\mathbb{H},\gamma} - T_{n,\gamma}\|_{HS} \leq \sqrt{\frac{2\tau}{n}}\left[\sqrt{2} + \left(\frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma_1^2}}} + \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma_2^2}}}\right)^{\frac{1}{2}}\right].$$

holds with a probability $1-2e^{-\tau}$. ■

A similar result holds when we have if $k_\gamma(x,t) = \sum_{j=1}^m k_{\gamma_j}(x,t)$.

## IV. LEARNING WITH A FAMILY OF GAUSSIAN KERNELS

In this section, our goal is to extend our study to an infinite number of Gaussian Kernels. One important thing to consider is how the corresponding RKHS of a family of Gaussian kernels is going to be defined. In 2010, Clint Scovel, Don Hush, Ingo Steinwart, and James Theiler introduced what they called RKHS of mixture, see [9]. In particular, they discussed the RKHS of the radial kernel when it is a family of Gaussain Kernels. We shall be able to relate the empirical operator to its continuous part as we will see below.

### A. RKHS OF RADIAL KERNELS

Let $X \subset \mathbb{R}^d$, and $k : X \times X \to \mathbb{R}$ be a radial kernel given by

$$k(x,t) = \int_{\mathbb{R}^+} k_\gamma(x,t)\, d\mu(\gamma),$$

where $k_\gamma(x,t)$ is a Gaussian kernel with a parameter $\gamma$, $\mu(\gamma)$ is a finite Borel measure on $\mathbb{R}^+$, $\gamma \in \mathbb{R}^+$, and $x, t \in \mathbb{R}^d$. Let $\Gamma \subset \mathbb{R}^+$, and $\mathbb{H}_{k_\gamma}(X)$ represents the reproducing kernel Hilbert space corresponding to the Gaussian kernel $k_\gamma(x,t)$, then for any $f_\gamma \in \mathbb{H}_{k_\gamma}(X)$, $f_\gamma = \langle f_\gamma, k_\gamma \rangle_{\mathbb{H}_{k_\gamma}(X)}$. Denote $\mathbb{H}_k(X)$ to the RKHS of the radial kernel $k(x,t)$, and for the sake of simplicity let $\mathbf{E}_{\gamma,\mu} = \int_\Gamma k_\gamma(x,t)\, d\mu(\gamma)$, then the RKHS corresponding to $k(x,t)$ is given by

$$\mathbb{H}_k(X) = \{\mathbf{E}_{\gamma,\,\mu} f_\gamma,\ f_\gamma \in \mathbb{H}_{k_\gamma}(X),\ \forall\, \gamma \in \Gamma\}$$

with the norm

$$\|f\|^2_{\mathbb{H}_k(X)} = \inf_{\substack{f=\mathbf{E}_{\gamma,\mu} f_\gamma \\ f_\gamma \in \mathbb{H}_{k_\gamma}(X),\ \gamma\in\Gamma}} \mathbf{E}_{\gamma,\mu}\|f_\gamma\|^2_{\mathbb{H}_{k_\gamma}(X)}.$$

### B. INTEGRAL OPERATOR DEFINED BY RADIAL REPRODUCING KERNELS

First of all, consider all the above assumptions. Let $L_k : \mathcal{L}^2(X,p) \to \mathcal{L}^2(X,p)$ be an integral operator defined by

$$(L_k f)(x) = \int_X f(t)\, \mathbf{E}_{\gamma,\mu}(x,t)\, dp(t),$$

where $\mathcal{L}^2(X,p)$ is the space of square integrable functions with a probability measure $p(x)$ with the norm

$$\|f\|^2 = \inf_{\substack{f=\mathbf{E}_{\gamma,\mu} f_\gamma \\ f_\gamma \in \mathcal{L}^2(X,p),\ \gamma\in\Gamma}} \mathbf{E}_{\gamma,\mu}\|f_\gamma\|^2_{\mathcal{L}^2(X,p)}.$$

It is easy to show that $L_K$ is a bounded and well-defined operator. Let $\kappa = \sup_{x\in X} k(x,x)$, then

$$\kappa = \mathbf{E}_{\gamma,\mu} e^{-\frac{\|x-x\|^2}{\gamma^2}} = \int_\Gamma e^{-\frac{\|x-x\|^2}{\gamma^2}}\, d\mu(\gamma) = \int_\Gamma d\mu(\gamma) = \mu(\Gamma).$$

Now let $\{x_1, \ldots, x_n\} \subset X$ a set of points sampled i.i.d. Then, the kernel matrix $K_n$ whose entry $ij$ is given by

$$K_{ij} = \frac{1}{n} k(x_i, x_j) = \frac{1}{n}\mathbf{E}_{\gamma,\mu}(x_i,x_j) = \frac{1}{n}\int_\Gamma k_\gamma(x_i,x_j)\, d\mu(\gamma).$$

The next step is to introduce two operators $T_{\mathbb{H},\Gamma}, T_{n,\Gamma} : \mathbb{H}_k(X) \to \mathbb{H}_k(X)$ using the reproducing property as follows,

$$\begin{aligned} T_{\mathbb{H},\Gamma} &= \int_X \mathbf{E}_{\gamma,\mu}\langle \cdot, k_\gamma(x,\cdot)\rangle_{\mathbb{H}_{k_\gamma}(X)}\, k_\gamma(x,\cdot)\, dp(x) \\ &= \int_\Gamma \int_X \langle \cdot, k_\gamma(x,\cdot)\rangle_{\mathbb{H}_{k_\gamma}(X)}\, k_\gamma(x,\cdot)\, dp(x)\, d\mu(\gamma) \\ T_{n,\Gamma} &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\gamma,\mu}\langle \cdot, k_\gamma(x_i,\cdot)\rangle_{\mathbb{H}_{k_\gamma}(X_i)}\, k_\gamma(x_i,\cdot) \\ &= \frac{1}{n}\int_\Gamma \sum_{i=1}^n \langle \cdot, k_\gamma(x_i,\cdot)\rangle_{\mathbb{H}_{k_\gamma}(X_i)}\, k_\gamma(x_i,\cdot)\, d\mu(\gamma). \end{aligned}$$

The next proposition will show how the latter operators approach each other.

*Proposition 4:* $T_{\mathbb{H},\Gamma}, T_{n,\Gamma}$ are Hilbert Schmidt operators. Moreover,if $X = \mathbb{R}^d$, and $p(x)$ is a normal distribution with a density

$$\phi_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}.$$

Then, with confidence $1 - 2\, e^{-\tau}$ the following inequality holds true:

$$\|T_{\mathbb{H},\Gamma} - T_{n,\Gamma}\|_{HS} \leq \sqrt{\frac{2\tau}{n}}[\mu(\Gamma) + \left(\mathbf{E}_{\gamma,\mu}\frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma^2}}}\right)^{\frac{1}{2}}.$$

*Proof:* Assume that $(\xi)_{i=1}^n$ is a sequence of random variables in the Hilbert space of Hilbert-Schmidt operators by

$$\xi_i = \langle \cdot, \int_\Gamma e^{-\frac{\|x_i-\cdot\|^2}{\gamma^2}}\, d\mu(\gamma)\rangle_{\mathbb{H}_{k_\gamma}(\mathbb{R})} \int_\Gamma e^{-\frac{\|x_i-\cdot\|^2}{\gamma^2}}\, \mu(\gamma) - T_{\mathbb{H},\Gamma}.$$

We have $E(\xi_i) = 0$. By a simple computation we obtain that

$$\begin{aligned} &\left\| \langle \cdot, \int_\Gamma e^{-\frac{\|x-\cdot\|^2}{\gamma^2}}\, d\mu(\gamma)\rangle_{\mathbb{H}_\gamma(\mathbb{R})} \int_\Gamma e^{-\frac{\|x-\cdot\|^2}{\gamma^2}}\, d\mu(\gamma)\right\|^2_{HS} \\ &= \left\| \int_\Gamma e^{-\frac{\|x-\cdot\|^2}{\gamma^2}}\, d\mu(\gamma)\right\|^4_{\mathbb{H}_{k_\gamma}(\mathbb{R}^d)} \leq \mu(\Gamma)^2. \end{aligned}$$

We can easily compute that

$$\|T_{\mathbb{H},\Gamma}\|_{HS} = \left(\int_\Gamma \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma^2}}} d\mu(\gamma)\right)^{\frac{1}{2}},$$

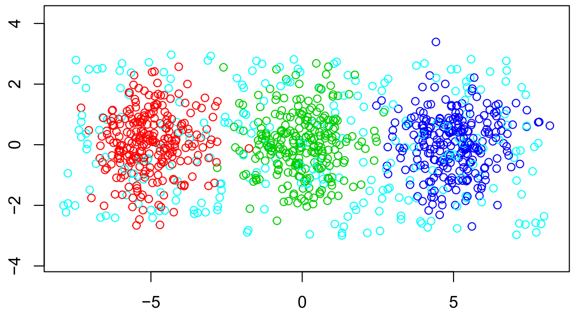


FIGURE 1. Four populations represented with different colors.

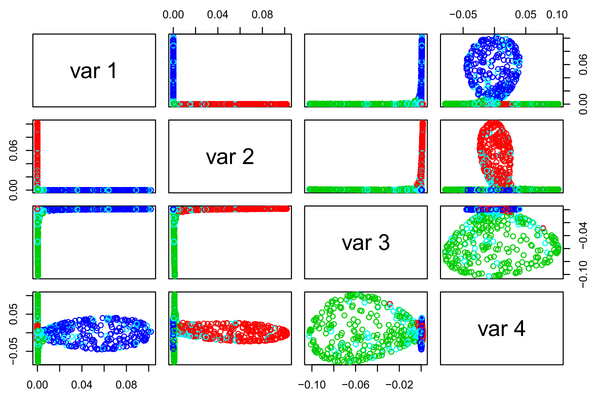



FIGURE 2. The first four eigenvectors of Gaussian kernel matrix with $\lambda = 1$.

and

$$\|\xi_i\|_{HS} \leq \mu(\Gamma) + \left(\int_\Gamma \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma^2}}} d\mu(\gamma)\right)^{\frac{1}{2}}.$$

Thus, the inequality

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right\|_{HS} = \|T_{\mathbb{H},\Gamma} - T_{n,\Gamma}\|_{HS}$$

$$\leq \sqrt{\frac{2\tau}{n}}[\mu(\Gamma) + \left(\int_\Gamma \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma^2}}} d\mu(\gamma)\right)^{\frac{1}{2}}].$$

holds true with confidence $1 - 2\,e^{-\tau}$. ■

At this point, we shall be able to bound the difference between the eigenvalues of the operators $L_k$, and $K_n$ when using radial kernels.

*Proposition 5:* Consider all the assumptions in section IV, then there exists an extended enumeration $\{\sigma_j\}_{j\geq 1}$ of discrete eigenvalues for $L_k$ and an extended enumeration $\{\hat{\sigma}_j\}_{j\geq 1}$ of discrete eigenvalues for $K_n$ such that

$$\sum_{j\geq 1}(\sigma_j - \hat{\sigma}_j)^2 \leq \frac{2\tau}{n}[\mu(\Gamma) + \left(\int_\Gamma \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma^2}}} d\mu(\gamma)\right)^{\frac{1}{2}}]^2,$$

with a probability greater than $1 - 2e^{-\tau}$.

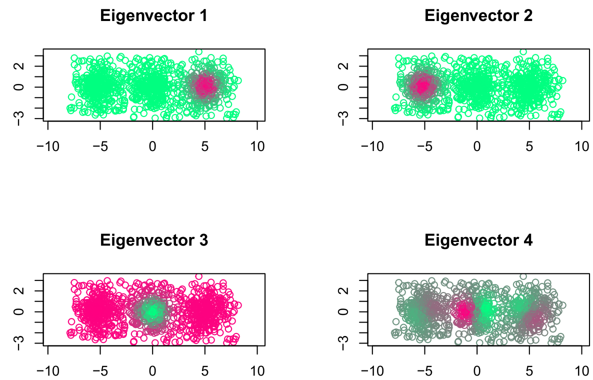



FIGURE 3. Another representation for the first four eigenvectors of Gaussian kernel matrix with $\lambda = 1$.

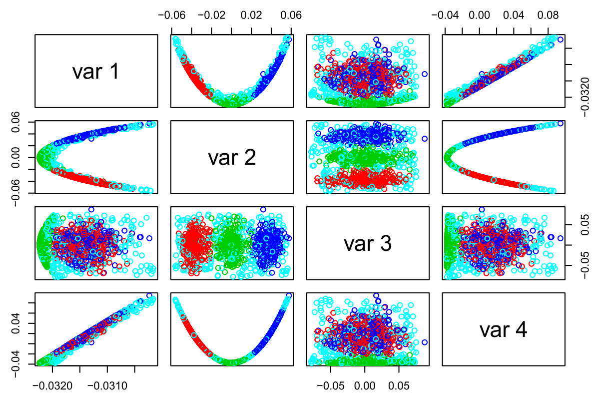



FIGURE 4. The first four eigenvectors of Gaussian kernel matrix with $\lambda = 1000$.

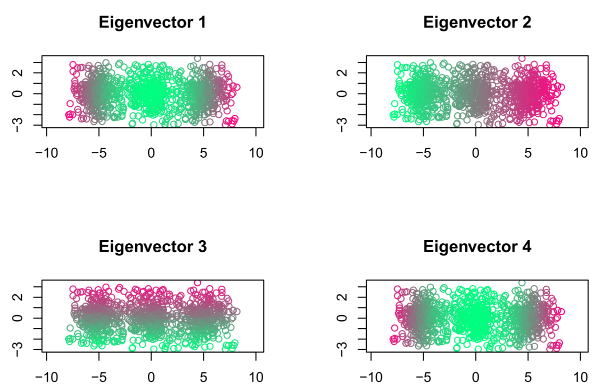



FIGURE 5. Another representation for the first four eigenvectors of Gaussian kernel matrix with $\lambda = 1000$.

## V. KERNEL PCA EXPERIMENTS

We create three clusters using standard bivariate normals shifted to have centers at the points $(-5, 0)$, $(0, 0)$, and $(5, 0)$. We also have a fourth population, uniformly distributed on the rectangle $[-8, 8] \times [-3, 3]$, see Fig. 1.

First, we use Kernel PCA, see [19], [20], [21], and [22], with the Gaussian kernel, with parameter $\lambda = 1$, see Fig. 2,3.

We repeat Kernel PCA, with the Gaussian kernel, with parameter $\lambda = 1000$, as can be seen in Fig. 4,5. We can see that Eigenvector 2 tracks very well the $x$ coordinate while Eigenvector 3 tracks very well the $y$ coordinate.

This shows that a smaller bandwidth works well for identifying clusters (areas of higher density), but loses track of the geometric location of the clusters with respect to each other, while a larger bandwidth can recover such relationships.

## VI. CONCLUSION

We studied the effect of the Gaussian kernel bandwidth parameter $\gamma$ on estimating an integral operator $L_{k_\gamma}$ defined on the space of square integrable functions by its empirical counterpart, which is the kernel matrix $K_n$ at a theoretical level. The proof technique, adapted from [6], involved establishing two operators, $T_{\mathcal{H}}$ and $T_n$ and define them on the Reproducing Kernel Hilbert space $\mathcal{H}_k$ with a reproducing kernel $k(x, t)$. Estimating $L_{k_\gamma}$ by $K_n$ would depend on how close the operators $T_{\mathcal{H}}$ and $T_n$ to each other. Therefore, we bounded the norm of the difference of these operators. Our results show that when the parameter $\gamma$ becomes smaller and smaller, the operators $T_{\mathcal{H}}$ and $T_n$ become closer and closer. The bounds we found for the Gaussian case improve on the general bounds found in [6], and allow us to show that this bound changes by a factor of less than 2 for all positive values of the bandwidth parameter. We have also shown how this translates to estimated spectral decompositions for different values of the bandwidth parameter.

An experiment on kernel PCA was performed to test the impact of the Gaussian kernel parameter. The results show that a small bandwidth tells us more about clusters, while a large one can recover the locations of these clusters. These results clearly support our claim that using Gaussian kernels at different bandwidth at the same time can help learning different things about the data.

## APPENDIX A
## PROOF OF PROPOSITION 1

First of all, for all $x \in \mathbb{R}$ we have that

$$\begin{aligned}(T_{\mathbb{H}}f)(x) &= \langle k_\gamma(x,\cdot), T_{\mathbb{H}}f\rangle_{\mathbb{H}_\gamma(\mathbb{R})}\\ &= \langle k_\gamma(x,\cdot), i_n^* i_n f\rangle_{\mathbb{H}_\gamma(\mathbb{R})}\\ &= \langle i_n k_\gamma(x,\cdot), i_n f\rangle_{\mathbb{L}^2(\mathbb{R})}\\ &= \int_{\mathbb{R}} k_\gamma(x,t) f(t)dp(t).\end{aligned}$$

We have

$$(T_{\mathbb{H}}f)(x) = \int_{\mathbb{R}} e^{-\frac{\|x-t\|^2}{\gamma^2}} f(t)dp(t). \tag{4}$$

We know that $\{e_n(x) = \sqrt{\frac{2^n}{\gamma^{2n} n!}} x^n e^{-\frac{x^2}{\gamma^2}},\ n = 0, 1, \ldots\}$ is an orthonormal basis for the Gaussian RKHS $\mathbb{H}_\gamma(\mathbb{R})$. Now we can calculate $\|T_{\mathbb{H}}\|_{HS}^2$ as follows,

$$\begin{aligned}\|T_{\mathbb{H}}\|_{HS}^2 &= \sum_{n=0}^{\infty} \langle T_{\mathbb{H}} e_n, T_{\mathbb{H}} e_n\rangle_{\mathbb{H}_\gamma(\mathbb{R})}\\ &= \sum_{n=0}^{\infty} \langle T_{\mathbb{H}} e_n, i_n^* i_n e_n\rangle_{\mathbb{H}_\gamma(\mathbb{R})}\\ &= \sum_{n=0}^{\infty} \langle i_n T_{\mathbb{H}} e_n, i_n e_n\rangle_{\mathcal{L}^2(\mathbb{R})}\\ &= \sum_{n=0}^{\infty} \langle T_{\mathbb{H}} e_n, e_n\rangle_{\mathcal{L}^2(\mathbb{R})}\\ &= \sum_{n=0}^{\infty} \int_{\mathbb{R}} T_{\mathbb{H}} e_n(x) \cdot e(x) p(x)\\ &= \sum_{n=0}^{\infty} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} k_\gamma(x,t) e_n(t) dp(t)\right) e_n(x) dp(x)\\ &= \sum_{n=0}^{\infty} \int_{\mathbb{R}} \int_{\mathbb{R}} e_n(t) k_\gamma(x,t) e_n(x) dp(t) dp(x)\\ &= \frac{1}{2\sigma^2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} \sum_{n=0}^{\infty} \frac{2^n}{\gamma^{2n} n!} t^n e^{-\frac{t^2}{\gamma^2}} e^{-\frac{(x-t)^2}{\gamma^2}}\\ &\quad \times x^n e^{-\frac{x^2}{\gamma^2}} e^{-\frac{t^2}{2\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dp(t) dp(x)\\ &= \frac{1}{2\sigma^2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-(\frac{1}{\gamma^2}+\frac{1}{2\sigma^2})x^2 + \frac{4xt}{\gamma^2} - (\frac{1}{\gamma^2}+\frac{1}{2\sigma^2})t^2} dxdt\\ &= \frac{1}{\sqrt{1+8\frac{\sigma^2}{\gamma^2}}}.\end{aligned}$$

which is the end of the proof.

## ACKNOWLEDGMENT

The authors would like to thank King Abdulaziz University and Kent State University for providing the opportunity to complete this work.

## REFERENCES

[1] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. London, U.K.: MIT Press, 2002.

[2] I. Steinwart and A. Christmann, *Support Vector Machines*, M. Jordan, J. Kleinberg, and B. Schölkopf, Eds. New York, NY, USA: Springer, 2008.

[3] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, M. J. Ablowitz, S. H. Davis, E. J. Hinch, A. Iserles, J. Ockenden, and P. J. Olver, Eds. New York, NY, USA: Cambridge Univ. Press, 2007.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

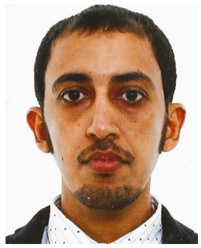
[5] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge Univ. Press, 2004.

[6] L. Rosasco, M. Belkin, and E. D. Vito, "On learning with integral operators," *J. Mach. Learn. Res.*, vol. 11, no. 30, pp. 905–934, 2010.

[7] E. D. Vito, L. Rosasco, A. Caponnetto, U. D. Giovannini, and F. Odone, "Learning from examples as an inverse problem," *J. Mach. Learn. Res.*, vol. 6, pp. 883–904, May 2005.

[8] I. Steinwart, D. Hush, and C. Scovel, "An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4635–4643, Oct. 2006, doi: 10.1109/TIT.2006.881713.

[9] C. Scovel, D. Hush, I. Steinwart, and J. Theiler, "Radial kernels and their reproducing kernel Hilbert spaces," *J. Complex.*, vol. 26, no. 6, pp. 641–660, Dec. 2010, doi: 10.1016/j.jco.2010.03.002.
[10] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics*, 12th ed., J. Siemons, Ed. Cambridge, U.K.: Cambridge Univ. Press, 1989, pp. 148–188.
[11] L. Pinelis, "An approach to inequalities for the distributions of infinite-dimensional martingales," in *Progress in Probability*, vol. 30, R. M. Dudley, M. G. Hahn, and J. Kuelbs, Eds. Boston, MA, USA: BirkhäUser, 1992, pp. 128–134.
[12] S. Axler, *Measure, Integration & Real Analysis*, P. Hersh, R. Vakil, and J. Wunsch, Eds. Cham, Switzerland: Springer, 2020.
[13] T. Kato, "Variation of discrete spectra," *Commun. Math. Phys.*, vol. 111, no. 3, pp. 501–504, Sep. 1987, doi: 10.1007/bf01238911.
[14] T. Kato, *Perturbation Theory for Linear Operators*, vol. 623, 2nd ed. Berlin, Germany: Springer, 1995.
[15] L. Zwald and G. Blanchard, "On the convergence of eigenspaces in kernel principal component analysis," in *Proc. Neural Inf. Process. Syst.*, 2005, pp. 1649–1656. [Online]. Available: https://api.semanticscholar.org/CorpusID:959869
[16] L. Lan, K. Zhang, H. Ge, W. Cheng, J. Liu, A. Rauber, X.-L. Li, J. Wang, and H. Zha, "Low-rank decomposition meets kernel learning: A generalized Nyström method," *Artif. Intell.*, vol. 250, pp. 1–15, Sep. 2017, doi: 10.1016/j.artint.2017.05.001.
[17] N. Aronszajn, *Theory of Reproducing Kernels*, vol. 60. Providence, RI, USA: Transactions of the American Mathematical Society, 1950.
[18] M. G. Genton, "Classes of kernels for machine learning: A statistics perspective," *J. Mach. Learn. Res.*, vol. 2, pp. 299–312, Dec. 2001.
[19] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.
[20] L. Zwald, O. Bousquet, and G. Blanchard, "Statistical properties of kernel principal component analysis," *Proc. 17th Annu. Conf. Learn. Theory*, Jul. 2004, pp. 594–608.
[21] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.
[22] T. Arnold, M. Kane, and B. W. Lewis, "Dimensionality reduction," in *A Computational Approach To Statistical Learning*, 1st ed. New York, NY, USA: CRC Press, 2019, pp. 261–295.



**MAHDI A. ALMAHDAWI** received the M.Sc. and Ph.D. degrees in applied mathematics from Kent State University, OH, USA, in 2021 and 2023, respectively. Currently, he is an Assistant Professor with King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include learning theory, statistical methods, data analysis, and kernel methods.

**OMAR DE LA C. CABRERA** received the B.Sc. degree in mathematics from Universidad Lisandro Alvarado, Barquisimeto, Venezuela, in 1991, the M.Sc. degree in mathematics from Venezuelan Institute for Scientific Research, Caracas, Venezuela, in 1996, the Ph.D. degree in mathematics from the University of Florida, Gainesville, FL, USA, in 2000, and the Ph.D. degree in statistics from the University of Chicago, Chicago, IL, USA, in 2008. From 2008 to 2011, he was a Postdoctoral Scholar with Stanford University. Currently, he is an Associate Professor with Kent State University. His research interests include statistical genetics, geometric approaches to data analysis, high dimensional data, and kernel methods. Previously, his field of work in pure mathematics was set theory.