STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **13**, April 2025, pp 1522–1535.
Published online in International Academic Press (www.IAPress.org)

Distance Measures for Hidden Markov Models Based on Hilbert Space Embeddings in Time Series Classification

Edgar Alirio Valencia-Angulo[1], Carlos Alberto Ramírez-Vanegas [1], Oscar Danilo Montoya [2]

[1]*Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad Tecnológica de Pereira, Pereira 660003, Colombia.*
[2]*Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá D.C. 110121, Colombia*

Abstract To build a classification scheme for sequences based on hidden Markov models (HMMs), an appropriate distance design is critical in both theory and practice. Kullback-Leibler (KL) and Hidden Markov stationary distance (HSD) measures have been used to build classification schemes for sequences based on HMMs. However, it has been widely recognized that the KL measure is not a true metric, and that the HSD metric is specifically designed for univariate data. Inspired by the recent emergence of probability distribution metrics in reproducible kernel Hilbert spaces (RKHS), this work introduces two new metrics between two stationary HMMs. The difference between the metrics based on RKHS, the HSD, and our proposal is that they can be calculated analytically and can be used for multivariate data. The performance of the two metrics in time series classification is assessed by applying them to a K-Nearest Neighbor (KNN) classifier. This evaluation uses the Massachusetts Eye and Ear Infirmary Disordered Voice Database, provided by Kay Elemetrics. The results indicate that the proposed metrics offer competitive classification accuracy in comparison with the KL, HSD, and dynamic time warping (DTW) measures.

Keywords Hidden Markov models, Reproducing kernel Hilbert spaces, Time series classification and Metrics between probability distributions.

AMS 2010 subject classifications 62J07

DOI:10.19139/soic-2310-5070-2184

List of acronyms

The following list presents some acronyms that will be used throughout the document:

BIC	Bayesian inference criterion
CV	Coefficient of variation
DTW	Dynamic time warping
ED	Euclidean distance
EEG	Electroencephalogram
EM	Expectation-maximization

*Correspondence to: O. D. Montoya (Email: odmontoyag@udistrital.edu.co). Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá D.C. 110121, Colombia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

GMM Gaussian mixture model

$HMMs$ Hidden Markov models

HSD Hidden Markov stationary distance

KEG Metric based on RKHS and the Gaussian kernel

KEL Metric based on RKHS and the Laplacian kernel

KL Kullback-Leibler

KNN K-Nearest Neighbor

MMD Maximum mean discrepancy

$RKHS$ Reproducible kernel Hilbert spaces

UCR Time series classification repository

1. Introduction

An HMM is a stochastic model that describes a sequence of random observations generated by an underlying Markov process with a finite number of hidden states, which are not directly observable [1]. HMMs can be adapted to address a variety of complex real-world issues [2, 3] and have been successfully used in fields such as automatic speech recognition [4, 5], bioinformatics [6], and computer vision [7]. Thanks to the Markovian property, HMMs are also attractive in terms of computational complexity, since they come equipped with efficient algorithms for parameter inference and prediction [8].

In the literature, three types of approaches are mainly used to establish metric spaces for HMMs: 1) defining distance measures between HMMs, 2) defining kernels over generative models (including HMMs as a particular case), and 3) building dissimilarity spaces, where the distance measures between HMMs are also used.

In the first group, the early work of [9] can be included, which defined the KL divergence between two HMMs together with a symmetric version of it. Another proposed measure is based on the Wasserstein metric [10] and computes the distance between two HMMs with state-conditional Gaussian distributions. A drawback of this metric, as with the original KL distance, is that it fails to satisfy the triangle inequality. In [11], a proper distance measure was introduced to build a classification scheme for sequences based on HMMs while employing stationary cumulative distribution functions. This approach is known as *HSD*, and, although it can be regarded as a real distance, its computation requires a numerical approximation of the integral associated with computing the difference between the cumulative distributions related to the stationary HMMs. Furthermore, the HSD measure was proposed while assuming one-dimensional input spaces, and its extension for multidimensional problems has not yet been explored.

The second group could include the works of [12] and [13], which introduced a similarity measure between probability distributions known as a *probability product kernel*. In this measure, the output value of a pair of data points in the input space is obtained through the inner product of their corresponding probability distributions. Here, the probability product kernel is a similarity measure that can be computed in closed form for HMMs. An alternative approach to measuring the similarity between two HMMs is the Fisher kernel [14], which combines the benefits of generative and discriminative methods for pattern classification. The main idea of the Fisher kernel is to map data points to a gradient vector derived from a generative probability model [15]

Finally, the third group relies on dissimilarity measures. A measure of dissimilarity between a pair of data points is a numerical measure of the extent to which a couple of data points are different. When a dissimilarity

measure is defined by a kernel and such measure separates the classes, the nearest neighbor method exhibits a better performance. Several applications that employ HMMs for statistical data modeling require a dissimilarity measure to build a classification scheme for their sequences. Some of these measures are based on ED [16], the Bayes probability of error, kernel methods [17], or neural networks [18], among others.
This paper focuses on the first group to define metric spaces for HMMs and introduce true distance measures using the method of probability distribution embeddings in an RKHS. The Hilbert space embeddings of probability distributions are a recent development in the field of kernel methods, and they allow mapping a set of probability distributions into an RKHS via an injective operator that relies on a characteristic kernel [19]. RKHS-based techniques have become useful and flexible tools to deal with high-dimensional statistical models [20, 21]. They have enabled the non-parametric estimation of HMMs [22], electroencephalogram (EEG) mental recognition [23], and the development of distance measures between probability distributions [24, 25].

Additionally, stochastic processes consider random variables that take on values within a space of possible paths. However, reducing a stochastic process to a path-valued random variable overlooks its filtration, which represents the flow of information carried by the process over time. By conditioning the filtration process, works such as [26] and [27] introduce a family of higher-order probability distribution embedding into RKHS, generalizing the classical method for embedding these distributions by incorporating additional filtration-related information.

Contributions. In [28] and [29], a metric between probability distributions is proposed using RKHS, called the *MMD* metric. However, no metric between HMMs using RKHS is known. Following the ideas of [11] and inspired by the framework proposed in [30], this study developed two new metrics to build a classification scheme for sequences based on stationary HMMs via RKHS. Similar to the HSD metric, our approach is based on the stationary distributions of the observation sequences for each HMM. The problem regarding the calculation of metrics between HMMs in RKHS lies in computing integrals that are usually difficult to solve analytically. This paper solves this issue by assuming Gaussian and Laplacian kernels for the embedding. It also assumes that the emission probability is given by a Gaussian mixture model (GMM). The difference between the metrics based on RKHS, HSD, and our proposal is that they can be calculated analytically. Furthermore, the metrics based on the Gaussian and Laplacian kernels can be used for any type multivariate data, unlike the HSD, which is numerically calculated and used for univariate data. This document presents a use case of one of the metrics for multivariate data. The performance of the proposed metrics is evaluated using a KNN classifier on databases from the UCR Time Series Classification Archive [31] and Kay Elemetrics' voice database [32], *i.e.*, the subset of 226 voice records described by [32]. Although the discussion is limited to the use of Gaussian or Laplacian kernels, our approach is general, and other types of kernels can be assumed, thereby obtaining a large range of applicable metrics.

This paper is organized as follows. The materials and methods are described in section 2. This section briefly describes the theory behind HMMs and Hilbert space embeddings for probability distributions, and it introduces the new measures for several applications that employ HMMs for statistical data modeling and require a dissimilarity measure to build a sequence classification scheme via the probability distribution embedding method in RKHS. Section 3 describes the experiments employed in this study and discusses the results obtained using the RKHS-based measures with real-world data. The conclusions and proposals for future works are outlined in section 4.

2. Materials and methods

Note that uppercase letters (*e.g.*, X, Y) are used to represent random variables, while lowercase letters (*e.g.*, x, y) denote the specific values of said variables. Probability distributions are denoted by the letters $\mathbb{P}, \mathbb{Q}$.

2.1. Hidden Markov models

For the sake readability, univariate observation sequences will be assumed, but the framework can be generalized for multivariate cases [8, 33].

Let $\mathbf{X} = (X_1, X_2, \cdots, X_T)$ be an univariate observation sequence of length T that is driven by an unobservable Markov sequence $\mathbf{H} = (H_1, H_2, \cdots, H_T)$. The random variables X_t and H_t for $t = 1, 2, \cdots, T$ take values in the sets $\mathbb{R}$ and $\{h_1, h_2, \cdots, h_N\}$, respectively, where N is the number of states. An HMM is completely defined by the triplet $\rho = (\pi, A, B)$. The first term $\pi \in \mathbb{R}^N$ represents the initial state distribution, with $\pi_i = p(H_1 = h_i)$. The transition matrix $A \in \mathbb{R}^{N\times N}$ has the entries $a_{ij} = p(H_t = h_j | H_{t-1} = h_i)$, $\forall i, j = 1, 2, \cdots, N$,, which represents the probability of transitioning from state h_i to state h_j. The vector of emission probabilities $B \in \mathbb{R}^N$ has the entries

$$b_i(x) = p(X_t = x | H_t = h_i), \;\; i = 1, 2, \cdots, N,$$

and it represents the probability of observing x_t from the state h_j.

Assume the sequence of observations $\mathbf{X} = (X_1, \; X_2, \; \cdots, \; X_T)$ to be described by an HMM. The probability of a given sequence $p\left(X_1 = x_1, \; \cdots, \; X_T = x_T\right) = p\left(x_1, \; \cdots, \; x_T\right)$ is defined as follows:

$$p\left(x_1, \; \cdots, \; x_T\right) = \sum_{i_1=1}^{N} \cdots \sum_{i_T=1}^{N} \pi_{i_1} b_{i_1}(x_1) \cdots a_{i_{T-1}, i_T} b_{i_T}(x_T).$$

This work focuses on the stationary distribution of the resulting HMM, which is expressed as follows:

$$p(x) = \sum_{i=1}^{N} \pi_{s,i} b_i(x), \tag{1}$$

where $\pi_{s,i}$ is the stationary probability of state i. That is to say that $\pi_{t,i} = \pi_{t-1,i} = \pi_{s,i}$, $t \geq t_s$, with t_s being the transient time. It is assumed that the emission probability $b_i(x)$ is given by a GMM with $M_{\mathbb{P}}$ mixture components.

$$b_i(x) = \sum_{j=1}^{M_{\mathbb{P}}} \alpha_{i,j} \mathcal{N}\left(x | \mu_{i,j}, \; \Sigma_{i,j}\right), \;\; \text{with} \;\; \sum_{j=1}^{M_{\mathbb{P}}} \alpha_{i,j} = 1 \;\; \text{and} \;\; \alpha_{i,j} \geq 0,$$

where $\alpha_{i,j}$ is the previous probability for the component j of state i; $\mu_{i,j}$ is the mean parameter for the component j of state i; and $\Sigma_{i,j}$ is the variance parameter for the component j of state i.

2.2. Hilbert space embeddings for probability distributions

Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions in a non-empty set $\mathcal{X}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The function $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}, \forall x, x' \in \mathcal{X}$ is a reproducing kernel in $\mathcal{H}$, which is a RKHS [34].

1. $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$.
2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

Let $\mathcal{P}$ be the space of all probability distributions and X a random variable with a distribution function $\mathbb{P} \in \mathcal{P}$. The function mean map μ_X from a probability distribution $\mathbb{P} \in \mathcal{P}$ for the RKHS $\mathcal{H}$ is defined as $\mu_X(\mathbb{P}) = \mathbb{E}_X[k(X, \cdot)] = \int_{\mathcal{X}} k(\cdot, x) d\mathbb{P}(x)$, where the kernel $k(x, x')$ is characteristic[†]. Examples of characteristic kernels in the set $\mathbb{R}^d$ can be seen in [30]. This paper uses Gaussian and Laplacian kernels because the resulting measures can be computed in closed forms.

Figure 1 shows the shapes of the Gaussian and Laplacian kernels. Note that the tails of the Gaussian kernel are heavier than those of the Laplacian kernel, making values far from the mean less probable.

[†]A characteristic kernel is a reproducing kernel for which $\mu_X(\mathbb{P}) = \mu_Y(\mathbb{Q}) \iff \mathbb{P} = \mathbb{Q}, \mathbb{P}, \mathbb{Q} \in \mathcal{P}$, where $\mathcal{P}$ denotes the set of all Borel probability measures in a topological space $(\mathcal{X}, \mathcal{A})$.

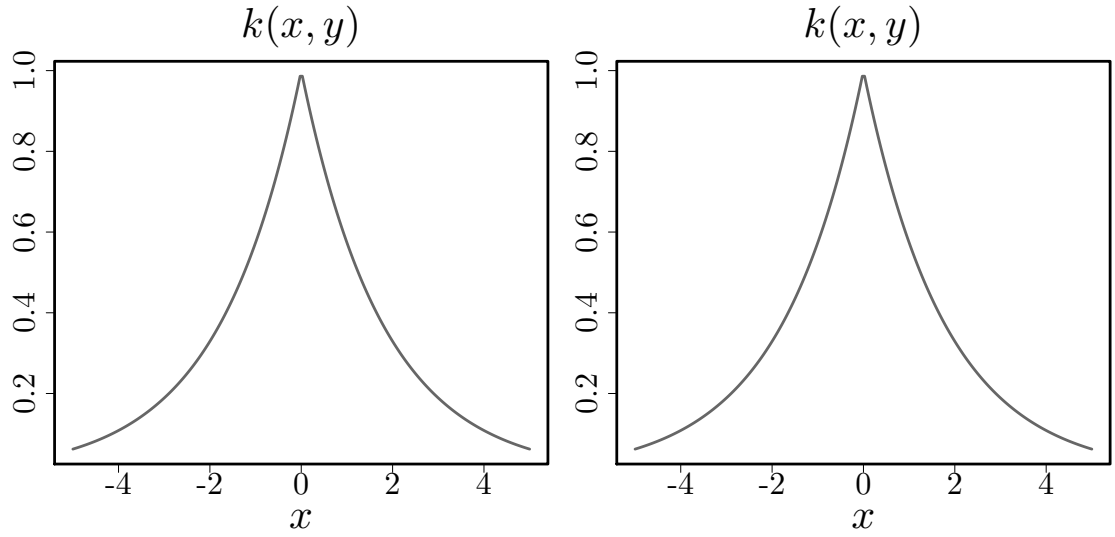

Figure 1. Shapes of the Gaussian (left) and Laplacian (right) kernels

2.3. Metric for the distance between probability distributions in RKHS

RKHS allow computing distances between probability distributions [35]. According to [30], the RKHS-based distance $\gamma_k^2(\mathbb{P}, \mathbb{Q})$ over the probability measures $\mathbb{P}$ and $\mathbb{Q}$ with the characteristic kernel k is given by

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \left\| \int_{\mathcal{X}} k(\cdot, x) d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot, y) d\mathbb{Q}(y) \right\|_{\mathcal{H}}^2. \tag{2}$$

If the probability distributions $\mathbb{P}(x)$ and $\mathbb{Q}(y)$ admit the density functions $p(x)$ and $q(y)$, respectively, then $d\mathbb{P}(x) = p(x)dx$ and $d\mathbb{Q}(y) = q(y)dy$. Hence, Equation (2) can be written as

$$\begin{aligned} \gamma_k^2(\mathbb{P}, \mathbb{Q}) = & \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) p(x) p(y) dx dy + \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) q(x) q(y) dx dy \\ & - 2 \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) p(x) q(y) dx dy. \end{aligned} \tag{3}$$

From Equation 3, the distance between probability distributions can be obtained by appropriately selecting the characteristic kernels and the probability functions p and q. In [28], the MMD metric is constructed while assuming any characteristic kernel and empirical distributions for p and q. In the next section, Equation 3 is used to obtain two metrics for HMMs. To this effect, it is assumed that the characteristic kernels are Gaussian and Laplacian, and that the distributions of the observations of the stationary HMMs are Gaussian mixtures. Under these assumptions, the integrals of Equation 3 are analytically solved, obtaining two metrics between HMMs. The novelty of this approach is that there are no known metrics between HMMs which use RKHS.

2.4. RKHS-based distance measures between stationary HMMs using characteristic kernels

The probability distributions for RKHS-based distance (p and q) are given by the stationary distribution of Equation (1), where GMMs provide the emission probabilities. Note that the solution to Equation (3) depends on the type of characteristic kernel $k(\cdot, \cdot)$ selected for the RKHS. Since both the Gaussian and the Laplacian kernels provide closed expressions when computing (3), they are used as characteristic kernels.

Assume that $\widehat{p}$ and $\widehat{q}$ are the estimators of the HMM stationary distributions p and q, respectively. Then, according to Equation (1), the following is obtained

$$\widehat{p}(x) = \sum_{i=1}^{N_{\mathbb{P}}} \pi_{s,i}^{\mathbb{P}} \widehat{b_i^{\mathbb{P}}}(x), \quad \text{and} \quad \widehat{q}(y) = \sum_{i=1}^{N_{\mathbb{Q}}} \pi_{s,i}^{\mathbb{Q}} \widehat{b_i^{\mathbb{Q}}}(y), \tag{4}$$

where $\pi_{s,i}^{\mathbb{P}}$ and $\pi_{s,j}^{\mathbb{Q}}$ are the stationary probabilities of the probability distributions $\mathbb{P}$ and $\mathbb{Q}$. Now, since the emission probabilities are modeled using Gaussian mixture distributions, the following can be obtained:

$$\widehat{b_i^{\mathbb{P}}}(x) = \sum_{j=1}^{M_{\mathbb{P}}} \alpha_{i,j} \mathcal{N}(x|\mu_{i,j}, \Sigma_{i,j}), \quad \text{and} \quad \widehat{b_i^{\mathbb{Q}}}(y) = \sum_{j=1}^{M_{\mathbb{Q}}} \beta_{i,j} \mathcal{N}(y|\nu_{i,j}.\Lambda_{i,j}).$$

Here, $\beta_{i,j}$ is the previous probability, $\nu_{i,j}$ is the mean parameter, and $\Lambda_{i,j}$ is the variance parameter for the component j of state i. Thus, the mean parameter for the component j of state i and $\Lambda_{i,j}$ represent the variance parameter for the component j of state i. The mean maps for the distributions are given by

$$\widehat{\mu}_X(\mathbb{P}) = \int_{\mathcal{X}} k(\cdot, x)\widehat{p}(x)dx, \quad \text{and} \quad \widehat{\mu}_Y(\mathbb{Q}) = \int_{\mathcal{X}} k(\cdot, y)\widehat{p}(y)dy. \tag{5}$$

Now, by substituting the expressions from Equation (5) into (3) and assuming a characteristic kernel $k(x, y; \ell)$, where ℓ is known as the *bandwidth*, the RKHS-based distance between the distributions $\mathbb{P}$ and $\mathbb{Q}$ is obtained, as given by

$$\begin{aligned}
\widehat{\gamma_k^2}(\mathbb{P}, \mathbb{Q}) = & \sum_{i,j=1}^{N_{\mathbb{P}}} \sum_{k,l=1}^{M_{\mathbb{P}}} \pi_{s,i}^{\mathbb{P}} \pi_{s,j}^{\mathbb{P}} \alpha_{i,k} \alpha_{j,l} \widehat{k}(\mu_{i,k}, \mu_{j,l}; \Sigma_{i,k}, \Sigma_{j,l}, \ell) \\
& + \sum_{i,j=1}^{N_{\mathbb{Q}}} \sum_{k,l=1}^{M_{\mathbb{Q}}} \pi_{s,i}^{\mathbb{Q}} \pi_{s,j}^{\mathbb{Q}} \beta_{i,k} \beta_{j,l} \widehat{k}(\nu_{i,k}, \nu_{j,l}; \Lambda_{i,k}, \Lambda_{j,l}, \ell) \\
& - 2 \sum_{i,j=1}^{N_{\mathbb{P}}, N_{\mathbb{Q}}} \sum_{k,l=1}^{M_{\mathbb{P}}, M_{\mathbb{Q}}} \pi_{s,i}^{\mathbb{P}} \pi_{s,j}^{\mathbb{Q}} \alpha_{i,k} \beta_{j,l} \widehat{k}(\mu_{i,k}, \nu_{j,l}; \Sigma_{i,k}, \Lambda_{j,l}, \ell).
\end{aligned} \tag{6}$$

If the kernel $k(x, y; \ell) = \exp\left(-\ell \|x - y\|_2^2\right)$ is Gaussian, then

$$\widehat{k}(x, y; \Sigma, \Lambda, \ell) = \widehat{k}_G(x, y; \Sigma, \Lambda, \ell) = \frac{\sqrt{\ell}}{\sqrt{\Sigma + \Lambda + \ell}} \exp\left(-\frac{(x-y)^2}{2(\Sigma + \Lambda + \ell)}\right).$$

In this case, the distance measure of (6) is $\widehat{\gamma_{k_G}^2}(\mathbb{P}, \mathbb{Q})$. On the other hand, if it is assumed that the kernel $k(x, y; \ell) = \exp(-\ell \|x - y\|_1)$ is Laplacian, then

$$\begin{aligned}
\widehat{k}(x, y; \Sigma, \Lambda, \ell) = & \frac{\Sigma\sqrt{\pi}}{\Lambda} f_1(x, y; \Sigma, \Lambda, \ell) \times \left(1 - erf\left(\frac{d_1(x, y; \Sigma, \Lambda, \ell)}{2\Lambda\Sigma^{-1}\sqrt{\Lambda^2 + \Sigma^2}}\right)\right) \\
+ & \frac{\Lambda\sqrt{\pi}}{\Sigma} f_2(x, y; \Sigma, \Lambda, \ell) \times \left(1 - erf\left(\frac{d_2(x, y; \Sigma, \Lambda, \ell)}{2\Sigma\Lambda^{-1}\sqrt{\Lambda^2 + \Sigma^2}}\right)\right),
\end{aligned}$$

where $f_1(x, y; \Sigma, \Lambda, \ell)$ is given by

$$\begin{aligned}
& \Lambda^2\sqrt{\pi} \exp\left(\frac{\left(2\ell^2 x + \Lambda^2\right)^2}{8\ell^4\Lambda^2}\left(1 - \frac{\Lambda^2}{\Sigma^2}\right)\right) \times \exp\left(-\frac{\left(2\ell^2 y + \Lambda^2\right)}{2\ell^2}\left(\frac{1}{2\ell^2} - \frac{x}{\Sigma^2}\right)\right) \\
& \times \left(-2^{-1}\left(\left(\frac{y}{\Lambda}\right)^2 + \left(\frac{x}{\Sigma}\right)^2\right) + \frac{\Sigma^2 d_1^2(x, y; \Sigma, \Lambda, \ell)}{4\Lambda^2}\right),
\end{aligned}$$

with $d_1(x,y;\Sigma,\Lambda,\ell) = \frac{\Lambda}{\sqrt{2l^2\Sigma^2}}\left(2\ell^2(x-y)+\Sigma^2+\Lambda^2\right)$. Note that

$$f_2(x,y;\Sigma,\Lambda,\ell) = f_1(y,x;\Lambda,\Sigma,\ell),$$

and $d_2(x,y;\Sigma,\Lambda,\ell) = d_1(x,y;\Lambda,\Sigma,\ell)$. $erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z \exp\left(-u^2\right)du$ is the Gauss error function. Here, the distance measure of (6) is $\widehat{\gamma^2_{k_L}}(\mathbb{P},\mathbb{Q})$.

The RKHS-based distance measures from Equation (6) have a closed form that depends on the bandwidth ℓ, the hidden states $N_{\mathbb{P}}$ and $N_{\mathbb{Q}}$, and the mixture components $M_{\mathbb{P}}$ and $M_{\mathbb{Q}}$. The parameters $\{\pi^{\mathbb{P}}_{s,i}, \pi^{\mathbb{Q}}_{s,i}, \alpha_{i,j}, \beta_{i,j}, \mu_{i,j}, \nu_{i,j}, \Sigma_{i,j}, \Lambda_{i,j}\}$ are estimated using the EM algorithm [3].

2.5. Advantages and disadvantages of RKHS-based metrics

Listed below are some strengths and limitations of the RKHS-based metrics for HMMs introduced in this paper.

Advantages. The proposed RKHS-based metrics between HMMs can be calculated analytically, allowing for greater precision in time series classification problems. Moreover, since the mean operator is a consistent estimator [36], the metrics are also consistent, ensuring that the results are reliable and unaffected by minor data variations. Another trait of the proposed metrics is their applicability to multivariate data, which makes them very useful for data with multiple dimensions. This is not the case for the HSD metric. Finally, our metrics are designed to capture nonlinear relationships within the data. By applying a characteristic kernel, the data are mapped to a higher-dimensional RKHS, transforming relationships that are nonlinear in the original space into linear ones. This transformation allows the identification of subtle similarities that might not be apparent in the original space, thereby facilitating the analysis.

Disadvantages. RKHS-based metrics between HMMs come at a high computational cost, as evaluating reproducing kernels on large datasets can be slow. Additionally, a shortcoming of these metrics is the difficulty in estimating both the kernel and the HMM parameters, which can affect performance if the estimates do not accurately capture the similarity between the HMMs. Furthermore, interpreting RKHS-based metrics can be less intuitive than those defined in the input space, making it more challenging to understand and visualize the differences between HMMs in an RKHS.

2.6. The HSD and KL measures

The proposed measures between HMMs were compared against the HSD and KL metrics, following the definition in [11], *i.e.*, if ρ_1 and ρ_2 are two HMMs and $\mathbf{X} = (X_1, X_2, \cdots, X_T)$ is a sequence of length T generated by ρ_1, then

$$D(\rho_1,\rho_2) = \frac{1}{T}\left(\sum_{i=1}^{T}\log p(x_i) - \log q(x_i)\right),$$

is the KL measure where p and q are the probability functions of ρ_1 and ρ_2, respectively, and they are estimated as given by Equation (4). To obtain a more precise KL measure in the experiments, the KL value is calculated using the symmetric version:

$$KL(\rho_1,\rho_2) = \frac{1}{2}\left(D(\rho_1,\rho_2) + D(\rho_2,\rho_1)\right).$$

The HSD measure is defined as in [11], *i.e.*, if ρ_1 and ρ_2 are two HMMs, then

$$HSD(\rho_1,\rho_2) = \int_{\mathcal{X}} |F_1(x) - F_2(x)|\,dx,$$

where $F_1(x) = \int_{\mathcal{X}} p(x)dx$ and $F_2(x) = \int_{\mathcal{X}} q(x)dx$, with p and q being stationary distributions of ρ_1 and ρ_2, respectively. They are estimated as given by Equation (4). The HSD metric is calculated numerically.

2.7. DTW-based classification

DTW is a generalization of ED for time series [37]. In this type of application, ED is perhaps the simplest and most straightforward metric to implement. However, since it maps each data point from two sequences, it becomes very fragile if two sequences are out of phase. DTW identifies the optimal path by creating nonlinear alignments between two time series, which is why it has been widely used as a metric for time series classification and clustering.

In the DTW-based classification methodology, a sequence $\mathbf{X}$ of length T_1 and a sequence $\mathbf{Y}$ of length T_2 are considered. A T_1-by-T_2 path matrix is computed, where the i-th, j-th element contains the distance between the two points x_i and y_j, such that $d(x_i, y_j) = \|(x_i - y_j)\|_p$, where $\|\cdot\|_p$ represents the L_p norm [38]. The warping path is subject to constraints such as the endpoint, continuity, and monotonicity [39]. The best match between two sequences is determined by the alignment that yields the shortest path. Consequently, the optimal warping path can be found using the recursive expression provided by

$$DTW_p(\mathbf{X}, \mathbf{Y}) = \sqrt[p]{\zeta(i,j)}, \tag{7}$$

where $\zeta(i,j)$ is the cumulative distance, as defined by

$$\zeta(i,j) = |x_i - y_i|^p + \min\{\zeta(i-1, j-1), \zeta(i-1, j), \zeta(i, j-1)\}. \tag{8}$$

3. Experimental results

This section presents the experimental results obtained from the classification of data available in the UCR Time Series Classification Archive [31] and the Massachusetts Eye and Ear Infirmary Disordered Voice Database from Kay Elemetrics [32]. According to [40], there are several advantages to using the KNN classifier for $K = 1$ (1NN). For example, it is known that the distance metric is fundamental to the performance of the 1NN classifier. Hence, the accuracy of the 1NN classifier directly reflects the effectiveness of the similarity measure. The 1NN classifier is straightforward to implement and is non-parametric, which makes it easy for anyone to reproduce our results. Finally, it has been proven that the error of the 1NN classifier is lower than that of other classifiers.

3.1. UCR data

The performance of the RKHS-based measures was evaluated using different databases from the UCR Time Series Classification Archive. These databases are described in in Table 1, which shows the size of the training and test data, the number of classes, and the time series length for the 36 most representative datasets.

A concept that will be used in the analysis of the results obtained from the classification of the binary databases is the *multivariate coefficient of variation* (CV) [41]. The multivariate CV is a measure of the variability of a set of time series [42]. In addition, it can be used to compare the variability of two such sets. According to [42], the definition of the multivariate CV is as follows: consider a set of n time series of length p, with a mean vector $\mu \neq 0$ and a covariance matrix Σ. The multivariate CV is given by

$$CV = \left(\text{Trace}(\Sigma)\left(\mu^\top \mu\right)^{-1}\right)^{1/2}.$$

In all the experiments, it was assumed that $M_{\mathbb{P}} = M_{\mathbb{Q}} = 1$, and, to select the number of hidden states for each one of databases, the BIC was implemented. The fundamental objective of the method was to maximize the probability of the data while penalizing large models, *i.e.*, to reduce the size of the model parameters. In the BIC method, the optimal number of states N is the value that maximizes the function $BIC(N) = \log(\mathbf{X}|\rho_b^N) - (R_N/2)\log(T)$, where $\mathbf{X}$ is the set of observed data, T is the total number of observations in $\mathbf{X}$, ρ_b^N denotes the maximum likelihood estimation of the HMM model with N, and R_N is the total number of free parameters in ρ_b^N [43].

The values of ℓ for the Gaussian and Laplacian kernels are estimated using cross-validation with the training data. 40% of the training data was randomly selected to train the KNN classifier and evaluate its performance, using the remaining 60% to choose the best ℓ value, which provides the highest accuracy in the classification step. The accuracy was determined by summing the number of successful cases and dividing the result by the total number of test samples. This procedure was repeated ten times, and the mean and the standard deviation $(\mu \pm \sigma)$ were computed [43].

Table 1. The 36 databases, including the size of their training and test sets, used to compare the performance of the RKHS-based metrics using the Laplacian kernel (KEL) and the Gaussian kernel (KEG) against that of the KL measure using the KNN algorithm for $K = 1$

Database	Number of classes	Train size	Test size	Time series length	HSD	KL	KEG	KEL
Adiac	37	390	391	176	**0.6957**	0.4808	0.6650	0.6957
BeetleFly	2	20	20	512	0.6560	**0.8500**	0.7000	0.6500
BirdChicken	2	20	20	512	**0.8500**	0.7500	**0.8500**	**0.8500**
Car	4	60	60	577	0.4667	**0.6333**	0.4667	0.4000
CBF	3	30	900	128	0.6467	0.6533	**0.6600**	0.6456
Coffee	2	28	28	286	**0.7500**	0.6429	0.5714	**0.7500**
Computers	2	250	250	720	0.6640	0.5240	0.6280	**0.6760**
DiatomSizeReduction	4	16	306	345	0.9444	0.9281	0.9575	**0.9706**
DistalPhalanxOutlineCorrect	2	276	600	80	**0.6650**	0.5960	0.6267	0.6367
DistalPhalanxTW	6	139	400	80	0.6850	0.4150	**0.6875**	0.6750
Earthquakes	2	139	322	512	0.6957	0.6957	0.7019	**0.8199**
GunPoint	2	50	150	150	0.8333	**0.9133**	0.7467	0.7933
Ham	2	109	105	431	0.5142	0.4857	0.4857	**0.5238**
Herring	2	64	64	512	0.4531	0.5156	0.4062	**0.5938**
ItalyPowerDemand	2	67	1029	24	**0.7512**	0.5063	0.7123	0.6929
LargeKitchenAppliances	3	375	375	720	0.6773	0.5840	0.6480	**0.6960**
Lightning-7	7	70	73	319	0.4521	0.5479	0.5890	**0.6027**
Meat	3	60	60	448	0.3667	**0.5667**	0.2667	0.3333
MedicalImages	10	381	760	99	**0.5118**	0.3789	0.4763	0.4934
MiddlePhalanxOutlineCorrect	2	291	600	80	0.6033	0.3750	0.5783	**0.6133**

According to Table 1, the 1-NN classifier based on the HSD metric exhibits a lower performance than that based on the KEL and KEG metrics. All this, when the size of the training set is small compared to the size of the test data, *e.g.*, the CBF, DiatomSizeReduction, MoteStrain, and TwoLeadECG databases. This probably has to do with the fact that the HSD metric has not shown consistency regarding execution times and is difficult to

Database	Number of classes	Train size	Test size	Time series length	HSD	KL	KEG	KEL
MiddlePhalanxTW	6	154	399	80	0.5288	0.4962	0.4461	**0.5388**
MoteStrain	2	20	1252	84	0.7492	**0.7524**	0.7276	0.7324
PhalangesOutlinesCorrect	2	1800	858	80	0.6468	0.4009	0.5874	**0.6585**
Plane	7	105	105	144	0.9810	0.9143	0.8090	**0.9905**
ProximalPhalanxOutlineAgeGro	3	400	205	80	0.7561	0.3659	**0.7756**	**0.7756**
ProximalPhalanxTW	6	205	400	80	0.6850	0.2925	0.6725	**0.6950**
SmallKitchenAppliances	3	375	375	720	0.5840	**0.6187**	0.5413	0.5493
SonyAIBORobotSurface	2	20	601	70	0.7486	**0.8136**	0.6639	0.4925
Strawberry	2	370	613	235	**0.7684**	0.4845	0.6933	0.7635
SyntheticControl	6	300	300	60	**0.7300**	0.1433	0.5863	0.6900
Trace	4	100	100	275	**0.9800**	0.1300	0.7200	0.9000
TwoLeadECG	2	23	1139	82	0.6927	0.5909	**0.7032**	0.6883
Wafer	2	1000	6164	152	**0.9644**	0.7583	0.9283	0.9517
Wine	2	57	54	152	**0.6111**	0.5185	**0.6111**	0.5926
Worms	5	77	181	900	**0.4475**	0.2707	0.4144	0.4199
Yoga	2	300	3000	426	**0.7523**	0.6780	0.6770	0.6653
Wins					13/36	7/36	6/36	**16/36**

solve when dealing with a large number of hidden states [44]. Therefore, in [44], an algorithm based on a numeric calculation for Gaussian HMMs is proposed. Quite the opposite happened with the RKHS-based metrics, which are consistent [28] and have a closed form.

The aforementioned table shows the databases where the KEL and KEG measures report a better classification accuracy than the KL metric: Coffee, DistalPhalanxOutlineCorrect, ItalyPowerDemand, MiddlePhalanxOutlineCorrect, TwoLeadECG, and Wine. In these databases, when the CV of the training and test data are low (less than 1), then at least one of the KEL or KEG measures performs noticeably better than the KL metric. The CV for these databases is presented in Table2. The fundamental reason for this competitive performance is that the KEL and KEG measures are based on smooth kernel functions and are trained with a set of smooth time series. Therefore, the performance regarding their classification is better when the test set also contains smooth time series.

Table 1 shows the performance of the RKHS-based distance measures in the datasets available from the UCR Time Series Classification Archive. Note that the KEL measure performs better than the HSD, KL, and KEG metrics in 16 databases when using a KNN classifier for $k = 1$, and the HSD performs better than the KL measure in 25. Nevertheless, the KEL and KEG metrics reported a better classification performance than the HSD in 16 databases, and the KEL metric surpasses the KL one in 26. In addition, the KEG measure outperforms the KL measure in 23 databases, and the KEL metric surpasses the KEG one in 26. In Table 2, note that the KEL and KEG

Table 2. Databases with a low training and test set CV

Database	Training data CV	Testing data CV
Coffee	0.0955	0.1052
DistalPhalanxOutlineCorrect	0.2820	0.2722
ItalyPowerDemand	0.5911	0.5303
MiddlePhalanxOutlineCorrect	0.1533	0.1620
TwoLeadECG	0.4240	0.4186
Wine	0.2000	0.0252

measures are outperformed by the KL measure when at least one of the CVs of the training and test sets is greater than 1, *i.e.*, when the time series of these databases have a high variability. For example, the CV of the training and test sets of BeetleFly are 2.8510 and 2.1903, respectively, while MoteStrain shows 1.1358 and 1.0404. For SonyAIBORobotSurface, these values are 13.5086 and 4.9382, and Yoga reports 1.7160 and 1.6958. This may be due to the fact that the KEL and KEG measures are based on smooth kernel functions, and, when they are trained with highly variable time series, their classification performance decreases. This may explain the KL measure's better performance in comparison with the KEL and KEG metrics for these databases. Finally, it was observed that the KEL metric performs better than the KEG one, given that the slow decay of the eigenvalue sequence of the Gram matrix associated with the Laplacian kernel enhances the detection of dependencies encoded in higher-frequency components of the probability distribution [28].

One of the best-known metrics in the literature for classifying time series is the DTW measure. This measure, together with the KNN classifier for $k = 1$, has a better classification performance than our proposal in most of the 36 databases presented in Table 1 – see [31] and Table 1. However, the purpose of our research is to develop metrics between HMMs by embedding probability distributions in an RKHS. Among the advantages of the proposed metrics in comparison with DTW and the HSD are their ease of use and their better performance in hypothesis testing problems in high-dimensional Hilbert spaces, just like the MMD metric presented in [24]. This advantage is due to the fact that KEL and KEG are true metrics induced by the norm of RKHS. This is not the case for the HSD metric, since it is not induced by a norm from a Hilbert space; and even less so for DTW, which is not a metric at all. Finally, the metric based on the Laplacian kernel can also be calculated for multivariate data, albeit involving the estimation of many parameters. Figure **??** shows a graphical comparison of our proposal (KEG and KEL) against the KL and HSD metrics according to the results presented in Table 1. Note that KEL and KEG outperformed KL (see Figures **??**a and Figure **??**c). Figure **??**b shows that the performance of HSD is similar to that of KEL.

3.2. Automatic assessment of voice quality using metrics between HMMs

Acoustic analysis has been improved to detect conditions affecting the quality of a person's voice. This technique is impartial, relies on the digital processing of voice signals, and does not require instruments that puncture the skin or physically enter the body (it is non-invasive). [45, 46]. This subsection compares the performance of the KEG, KL, and DTW metrics on a voice database. The database used in this work is described by [32].

Description of the voice database. The dataset features the analysis of 226 voice samples characterized by cepstral coefficients in the Mel frequency scale. Each voice sample is segmented into 10 ms intervals, with each interval described by 12 cepstral coefficients and an energy term. Consequently, each voice data segment is detailed with 39 distinct features. Each signal is categorized as either a *pathological voice* or a *healthy voice*. This study employed a 1-NN classifier as well as the KEG, KL, and DTW distance metrics. The HMMs was configured as shown in subsection 3.1. For the DTW-based classifier, the original speech signal were used.

To assess the statistical significance of the results, ten iterations of each classifier were executed using hold-out partitions. As illustrated in Table 3, the KEG metric showed a superior classification accuracy in the voice database when compared to the DTW measure. However, KEG delivered a comparable performance to KL. To assess the statistical significance of the results for the different measures, a Lilliefors test for normality was applied to the ten repetitions of each classifier. If the null hypothesis of normality was disproved, a Kruskal-Wallis test was performed

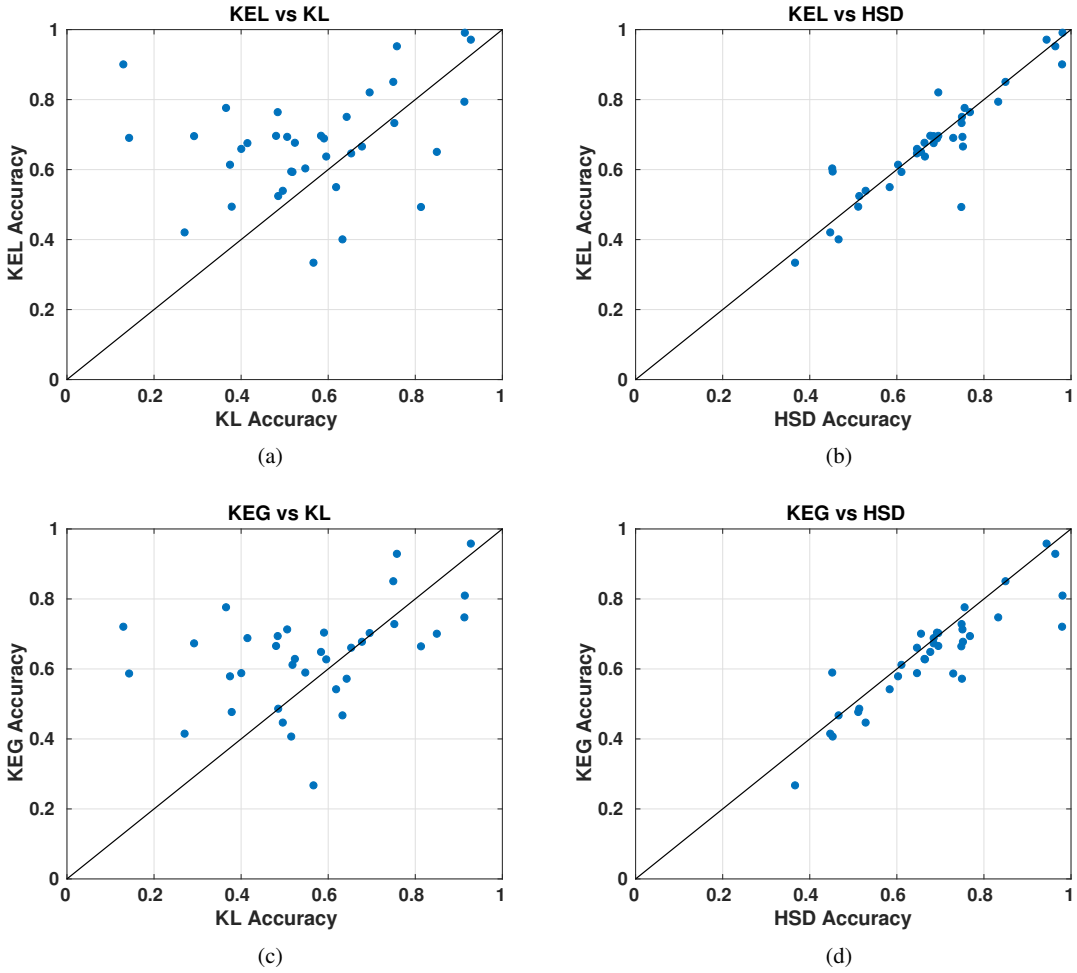

Figure 2. The KEL and HSD distance measures provide competitive classification accuracy when compared to the KEG and KL metrics using the KNN algorithm for $k = 1$ in the 36 databases of the UCR Time Series Classification Archive

Table 3. Accuracy results obtained using the KEG, KL and DTW metrics for $K = 1$. The mean μ and the standard deviation σ are shown for 10 repetitions of each experiment ($\mu \pm \sigma$).

KEG	KL	DTW
$\mathbf{0.8745 \pm 0.0100}$	0.8679 ± 0.0267	0.8481 ± 0.0196

to evaluate the average performance differences between the classifiers. If the null hypothesis of equal medians was rejected, multiple comparison tests were conducted using the Tukey-Kramer method to determine which classifiers differed from each other. All significance levels were set at 5% [47]. Our findings indicate that the accuracy results for the KEG and DTW metrics are statistically different, whereas the results for the KEG and KL metrics are equivalent. A possible reason for KEG's superior performance in comparison with the DTW measure is that the latter is not well-suited for multichannel signal speech recognition [48].

4. Conclusion and future work

This paper proposed two new distance measures between HMMs (KEL and KEG) which are based on Hilbert space embeddings of probability distributions and the HMM stationary probability distribution. These metrics can be calculated analytically and can be used for multivariate data. Experiments with real data show that our measures outperformed the KL metric, while KEL performed similarly to the HSD in classifying time series from the UCR repository. Moreover, the KEG metric was tested in voice quality assessment, obtaining better results than the well-known DTW algorithm. An interesting extension of this work would be the construction of a measure between two HMMs while using RKHS and considering the transition matrix to define the distance. Future work should also focus on constructing metrics to compare random processes using path signatures in RKHS.

ACKNOWLEDGEMENTS

The authors would like to thank the Master's Program in Mathematics Teaching of Universidad Tecnológica de Pereira and the GCEM research group of Universidad Distrital Francisco José de Caldas.

REFERENCES

1. Rubayyi Alghamdi. Hidden markov models (hmms) and security applications. *International Journal of Advanced Computer Science and Applications*, 7(2), 2016.
2. Mete Özbaltan. Hidden abstract stack markov models with learning process. *Mathematics*, 12(13):1–19, 2024.
3. Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective (Adaptive Computation And Machine Learning Series)*. The MIT Press, 2012.
4. Tina Raissi, Wei Zhou, Simon Berger, Ralf Schlüter, and Hermann Ney. Hmm vs. ctc for automatic speech recognition: Comparison based on full-sum training from scratch, 2022.
5. E. Trentin and M. Gori. A survey of hybrid ANN and HMM models for automatic speech recognition. *Neurocomputing*, 37(1-4):91–126, 2001.
6. B. J. Yoon. Hidden Markov models and their applications in biological sequence analysis. *Current Genomics*, 10(6):402–415, 8 2009.
7. Y. Wang, S. Resnick, and C. Davatzikos. *Spatio-temporal analysis of brain MRI images using Hidden Markov models*, pages 160–168. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
8. O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer series in statistics. Springer, 1st edition, 2005.
9. John R. Hershey, Peder A. Olsen, and Steven J. Rennie. Variational kullback-leibler divergence for hidden markov models. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 323–328, 2007.
10. Yukun Chen, Jianbo Ye, and Jia Li. A distance for HMMs based on aggregated Wasserstein metric and state registration. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.
11. J. Zeng, j. Duan, and C. Wu. A new distance measure for hidden Markov models. *Expert Systems with Applications*, 37(2):1550–1555, 2010.
12. Atena Mojahed, Mohammad Moattar, and Hamidreza Ghaffari. Supervised kernel-based multi-modal bhattacharya distance learning for imbalanced data classification. *Knowledge and Information Systems*, pages 1–26, 09 2024.
13. Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5(Jul):819–844, 2004.
14. Benyamin Ghojogh, Fakhri Karray, and Mark Crowley. Fisher and kernel fisher discriminant analysis: Tutorial. *arXiv preprint arXiv:1906.09436*, 2019.
15. F. Perronnin and J. Rodriguez. Fisher kernels for handwritten word-spotting. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 106–110. IEEE, 2009.
16. Yukun Chen, Jianbo Ye, and Jia Li. Aggregated wasserstein distance and state registration for hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2133–2147, 2020.
17. Mahdi Hashemi and Hassan A Karimi. *Statistics, Optimization &; Information Computing*, 6(4):497–525, Nov 2018.
18. Rajan Kumar Soni, Karthick Seshadri, and Balaraman Ravindran. Metric learning for comparison of hmms using graph neural networks. In *Asian Conference on Machine Learning*, 2021.
19. A. J. Smola, A. Gretton, L. Song, and B. Schöllkopf. A Hilbert space embedding for distributions. In *18th international conference on algorithmic learning theory: Springer-Verlag, Berli, Germany*, pages 13–31, 2011.
20. S. Zhou. Which is better? regularization in rkhs vs r^m on reduced svms. *Statistics, Optimization and Information Computing*, 1(11):82–106, 2013.
21. Tian-Yi Zhou, Namjoon Suh, Guang Cheng, and Xiaoming Huo. Approximation of rkhs functionals by neural networks, 2024.
22. L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola. Hilbert space embeddings of hidden Markov. In *27th Annual International Conference on Machine Learning, Haifa-Israel*, 2010.
23. S. Liu W. Lei, Z. Ma and Y. Lin. Eeg mental recognition based on rkhs learning and source dictionary regularized rkhs subspace learning. *IEEE, Access*, 9:150545–150559, 2021.
24. A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur. A fast consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

25. Carlos D Zuluaga, Edgar A Valencia, Mauricio A Álvarez, and Álvaro A Orozco. A Parzen-based distance between probability measures as an alternative of summary statistics in approximate Bayesian computation. In *International Conference on Image Analysis and Processing*, pages 50–61. Springer, 2015.
26. Cristopher Salvi, Maud Lemercier, Chong Liu, Blanka Horvath, Theodoros Damoulas, and Terry Lyons. Higher order kernel mean embeddings to capture filtrations of stochastic processes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16635–16647. Curran Associates, Inc., 2021.
27. Maud Lemercier, Cristopher Salvi, Theodoros Damoulas, Edwin V. Bonilla, and Terry Lyons. Distribution regression for sequential data. In *AISTATS*, 2021.
28. Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. *Measuring Statistical Dependence with Hilbert-Schmidt Norms*, pages 63–77. Springer Berlin Heidelberg, 2005.
29. Cristhian Valencia, Andres Alvarez, Edgar Valencia, Mauricio Álvarez, and Alvaro Orozco. *Information Potential Variability for Hyperparameter Selection in the MMD Distance: 23rd Iberoamerican Congress, CIARP 2018, Madrid, Spain, November 19-22, 2018, Proceedings*, pages 279–286. 03 2019.
30. B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.
31. Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The UCR time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data.
32. Vijay Parsa and Donald G Jamieson. Identification of pathological voices using glottal noise measures. *Journal of speech, language, and hearing research*, 43(2):469–485, 2000.
33. C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.
34. B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive computation and machine learning. MIT Press, 2002.
35. Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2004.
36. Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pages 21247–21259. Curran Associates, Inc., 2020.
37. Huanhuan Li, Jingxian Liu, Zaili Yang, Ryan Wen Liu, Kefeng Wu, and Yuan Wan. Adaptively constrained dynamic time warping for time series classification and clustering. *Information Sciences*, 534:97–116, 2020.
38. Young-Seon Jeong, Myong K Jeong, and Olufemi A Omitaomu. Weighted dynamic time warping for time series classification. *Pattern Recognition*, 44(9):2231–2240, 2011.
39. Yuru Teng, Guotao Wang, Cailing He, Yaoyang Wu, and Chaoran Li. Optimization of dynamic time warping algorithm for abnormal signal detection. *International Journal of Data Science and Analytics*, pages 1–13, 09 2023.
40. Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.
41. J. M Novo, R. D. S Lima, and C. S. M Komino. The use of coefficients of variation for comparison of force–time curves from handgrip tests. In *VI International Conference on Computational Bioengineering*, 2015.
42. V. V. Leigh. The statistics of variation. *Variation: A central concept in biology*, pages 29–48, 2005.
43. Manuele Bicego, Vittorio Murino, and Mário AT Figueiredo. A sequential pruning strategy for the selection of the number of states in hidden Markov models. *Pattern Recognition Letters*, 24(9):1395–1407, 2003.
44. J. Zeng, L. Xie, U. Kruger, J. Yu, J. Sha, and X. Fu. Process monitoring based on Kullback Leibler divergence. In *Control Conference (ECC), 2013 European*, pages 416–421. IEEE, 2013.
45. Nicolás Sáenz-Lechón, Juan I Godino-Llorente, Víctor Osma-Ruiz, Manuel Blanco-Velasco, and Fernando Cruz-Roldán. Automatic assessment of voice quality according to the grbas scale. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 2478–2481. IEEE, 2006.
46. Julián Gil González, Mauricio A Álvarez, and Álvaro A Orozco. Automatic assessment of voice quality in the context of multiple annotations. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 6236–6239. IEEE, 2015.
47. Hernán Darıo Vargas Cardona, Alvaro Angel Orozco, and Mauricio A Alvarez. Multi-patient learning increases accuracy for subthalamic nucleus identification in deep brain stimulation. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 4341–4344. IEEE, 2012.
48. Dr Kriti Suneja and Malti Bansal. Hardware design of dynamic time warping algorithm based on fpga in verilog. 2015.