DIMENSION REDUCTION AND KERNEL PRINCIPAL COMPONENT ANALYSIS

PALLE E.T. JORGENSEN, SOORAN KANG, MYUNG-SIN SONG, AND FENG TIAN

ABSTRACT. We study non-linear data-dimension reduction. We are motivated by the classical linear framework of Principal Component Analysis. In non-linear case, we introduce instead a new kernel-Principal Component Analysis, manifold and feature space transforms. Our results extend earlier work for probabilistic Karhunen-Loève transforms on compression of wavelet images. Our object is algorithms for optimization, selection of efficient bases, or components, which serve to minimize entropy and error; and hence to improve digital representation of images, and hence of optimal storage, and transmission. We prove several new theorems for data-dimension reduction. Moreover, with the use of frames in Hilbert space, and a new Hilbert-Schmidt analysis, we identify when a choice of Gaussian kernel is optimal.

CONTENTS

1

1. Introduction

Recently a number of new features of principal component analysis (PCA) have lead to exciting and new improved dimension reduction (DR). See e.g., [BN03, GGB18, JHZW19, GJ06, Bis06, Bis13, ZBB04, AFS18, VD16]. In general DR refers to the process of reducing the number of random variables under consideration in such areas as machine learning, statistics, and information theory. Within machine learning, it involves both the steps of feature selection and feature extraction. In the present paper, we shall consider linear as well as non-linear data models. The linear case arises naturally in principal component analysis (PCA). See [Son08, JS07]. Here one starts with a linear mapping of the given data into a suitable lower-dimensional space. However, this must be done in such a way that the variance of the data in the low-dimensional representation is maximized. As for the variance, we study both covariance, and the correlation matrix for the underlying data. The eigenvectors that correspond to the largest eigenvalues (the principal components) are then used in a construction of a large fraction of the initial variance, i.e., that which corresponds to the original data. The first few eigenvectors are typically interpreted in terms of the large-scale physical behavior of a particular system; and will retain the most important variance features. We begin here with an explicit model from image processing, and involving wavelet algorithms.

In nonlinear settings, principal component analysis can still be adapted, but now by means of suitable kernel tricks. In some applications, instead of starting from a fixed kernel, the optimization will instead try to learn, or adapt, for example with the use of semidefinite programs. The most prominent such a technique is known as maximum variance unfolding (MVU).

In recent years, the subject of kernel-principal component analysis, and its applications, has been extensively studied, and made progress in diverse directions. In addition to earlier papers by the co-authors [Son08, JS07], we include here a partial list of other relevant and current citations; see e.g., [LHN18, Raj18, WGLP19, RS99, GGB18, DWGC18, THH19, LLY+19, MMP19, JHZW19, CLS+19, GK19].

The goal of this paper is to extend our previous results on Karhunen-Loève transform to a nonlinear setting by means of kernel-principal component analysis (KPCA). This paper is organized as follows: In section 2.1, we illustrate Karhunen-Loève transform, which is very similar to PCA, applied to digital image compression. Then in section 2.3, we explain how PCA is used on linear data dimension reduction.

In section 3, we show our main results using KPCA on nonlinear data. The setting for our main theme begins with an example, and with results for PCA in case of classical covariance kernels; see section 2.5, and Lemma 3.9. The latter is for rank-1 projections, but is then extended to PCA selection, and algorithms, in subsequent results. Our focus is the non-linear case. Indeed, the focus in section 3 is nonlinear data dimension reduction, and the corresponding kernels, the core of our paper. In particular, our Theorem 3.11 deals with kernel PCA for nonlinear data dimension reduction (see Examples 3.15 and 3.16). In the remaining part of section 3, and in section 4, we turn to the case of Gaussian kernels and the corresponding stochastic processes. Using a new transform, we identify when a choice of Gaussian kernel is optimal for KPCA and nonlinear data dimension reduction.

Both tools, PCA and KPCA, are known. The focus of our current paper is to present a framework of operator in suitable Hilbert space, and an associated

spectral theory. Even though the applications we include here are presented in finite dimensional setting, most of our results extend to infinite dimensional spaces as well. Nonetheless, for use in recursion schemes, the finite-dimensional case is most relevant.

Our main results deal with algorithms for optimization in maximal variance, and dimension reduction-problems, from PCA. They go beyond earlier such approaches in the literature. Our man results are stated in section 3, and they include Lemma 3.9, Theorem 3.11, and Corollary 3.14 (also see Theorem 2.12); each of which are formulated and proved in a general setting of kernel analysis; hence in a non-linear framework of feature selection. In section 4, we prove a number of applied transform-results for positive definite kernels, their Hilbert spaces, and their associated Gaussian processes. This in turn extends earlier work on Monte Carlo, and Karhunen–Loève analysis, also known as the Kosambi–Karhunen–Loève approach. These new transform tools are motivated by our results in the earlier two sections in the paper. Our Examples included there are also novel, and serve as key points.

2. Karhunen-Loève transform or Principal Component Analysis

A Karhunen-Loève (KL) expansion typically refers to a rather general tool of stochastic analysis. Our present KL expansion are representations for certain stochastic processes. Given a stochastic processes, the starting point is the associated covariance function. From the latter we then build KL-representations as infinite series, each term in the expansion is a product of a term from an associated deterministic basis, and a random term from system of independent identically distributed standard Gaussians; the latter arising for example from a Monte Carlo generator; see section 3 below. We note that, for a given stochastic process, there will typically be many choices of KL expansions. We also refer to them as Karhunen-Loève transforms; often involving sampling. As outlined below, we stress that these KL-transforms are closely related to what is also known as principal component analysis (PCA); a key point for our present paper; applications to images. The importance of the KL expansions, when chosen with care, is that they yield optimal bases representations, for example with respect to mean-square error.

Principal component analysis (PCA) is a statistical tool based on certain orthogonal transforms. PCA has recently found a rich variety of applications (see e.g., [BN03, GGB18, JHZW19, GJ06, Bis06, Bis13, ZBB04, AFS18, VD16]). It serves to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of uncorrelated variables called principal components. PCA serves as an important tool in exploratory data analysis and for making predictive models. It is often used to visualize genetic distance and relatedness between populations. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after a normalization step of the initial data.

In more detail. If a particular application involves a certain number n of observations involving, say p variables, then the number of distinct principal components is $\min(n-1, p)$. The corresponding PCA transformation will then be defined such that the first principal component has the largest possible variance (accounting for as much of the variability in the data as possible), and the succeeding components in turn include the highest variance possible under orthogonality constraints; so

orthogonality to the preceding components. The resulting system of vectors, containing n observations, then forms an uncorrelated orthogonal basis set. It is known that PCA is sensitive to the relative scaling of the original variables.

Historically PCA originates with pioneering work by Karl Pearson, and it was later developed by Harold Hotelling in the 1930s. It is also closely related to what is known as the discrete Karhunen-Loève transform in signal processing, the Hotelling transform in multivariate quality control, proper orthogonal decomposition (POD) in mechanical engineering, to singular value decomposition (SVD), even to the eigenvalue decomposition from linear algebra, factor analysis (for a discussion of the differences between PCA and factor analysis or empirical orthogonal functions (EOF) in meteorological science, empirical eigenfunction decomposition, empirical component analysis, spectral decomposition in noise and vibration, and empirical modal analysis in structural dynamics.)

In general, one refers to a KL transform as an expansion in Hilbert space with respect to an ONB resulting from an application of the Spectral Theorem. A similar version is Principal Component Analysis which is used in various engineering applications, specifically dimension reduction of data. So we bring up the image PCA transform case which is 2-dimensional to serve as an illustration on how it changes the components or axes, and further results in dimension reduction in data which is multi-dimensional, thus eventually lead to our main result in KPCA on nonlinear data dimension reduction.

There are various image compression schemes apart from PCA such as discrete wavelet transform. See e.g., [MP05, KKS16, dES12], and [GJ06, DF07, GK15] for recent developments.

PCA allows an image to be compressed in the means of principal component. The method is explained in the following algorithm.

2.1. **The Algorithm for a Digital Image or Data Application.** Our aim is to reduce the number of bits needed to represent an image by removing redundancies as much as possible. Karhunen-Loève transform or PCA is a transform of m vectors with the length n formed into m-dimensional vector $X = [X_1, \cdots, X_m]$ into a vector Y according to

$$Y = A(X - m_X), \tag{2.1}$$

where matrix A is obtained by eigenvectors of the covariance matrix C as in (2.3) below. For further explanations and details please see section 2.3 and section 2.4.

The algorithm for Karhunen-Loève transform or PCA can be described as follows:

1. Take an image or data matrix X, and compute the mean of the column vectors of X

$$m_X = E(X) = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{2.2}$$

2. Subtract the mean: Subtract the mean, m_X in equation (2.2) from each column vector of X. This produces a data set matrix B whose mean is zero, and it is called centering the data.

3. Compute the covariance matrix from the matrix in the previous step

$$C = cov(X) = E\left((X - m_X)(X - m_X)^T\right) \tag{2.3}$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T - m_X m_X^T.$$

Here $X - m_X$ can be interpreted as subtracting m_X from each column of X. $C\,(i,i)$ lying in the main diagonal are the variances of

$$C(i,i) = E((X_i - m_{X_i})^2). \tag{2.4}$$

Also, $C(i,j) = E\left((X_i - m_{X_i})\left(X_j - m_{X_j}\right)\right)$ is the covariance between X_i and X_j.

4. Compute the eigenvectors and eigenvalues, λ_i of the covariance matrix.
5. Choose components and form a feature vector (matrix of vectors),

$$A = (eig_1, ..., eig_n). \tag{2.5}$$

List the eigenvectors in decreasing order of the magnitude of their eigenvalues. This matrix A is called the row feature matrix. By normalizing the column vectors of matrix A, this new matrix P becomes an orthogonal matrix. Eigenvalues found in step 4 are different in values. The eigenvector with highest eigenvalue is the principle component of the data set. Here, the eigenvectors of eigenvalues that are not up to certain specific values can be dropped thus creating a data matrix with less dimension value.

6. Derive the new data set.

Final Data $=$ Row Feature Matrix $\times$ Row Data Adjust.

The rows of the feature matrix A are orthogonal so the inversion of PCA can be done on equation (2.1) by

$$X = A^T Y + m_X. \tag{2.6}$$

With the l largest eigenvalues with more variance are used instead of $n$eigenvalues, the matrix A_l is formed using the l corresponding eigenvectors. This yields the newly constructed data or image X' as follows:

$$X' = A_l^T Y + m_X. \tag{2.7}$$

Row Feature Matrix is the matrix that has the eigenvectors in its rows with the most significant eigenvector (i.e., with the greatest eigenvalue) at the top row of the matrix. Row Data Adjust is the matrix with mean-adjusted data transposed. That is, the matrix contains the data items in each column with each row having a separate dimension (see e.g., [MP05, AW12, Mar14]).

This algorithm can be used for linear data dimension reduction and this is illustrated in section 2.3.

2.2. **Principal Component Analysis in a Digital Image.** We would like to use a color digital image PCA to illustrate dimension change in this section, so we introduce a color digital image. A color digital image is read into a matrix of pixels. We would like to use Karhunen-Loève transform or PCA applied to a digital image data illustrate dimension reduction. Here, an image is represented as a matrix of functions where the entries are pixel values. The following is an example of a matrix representation of a digital image:

$$\mathbf{f(x,y)} = \begin{pmatrix} f(0,0) & f(0,1) & \cdots & f(0,N-1) \\ f(1,0) & f(1,1) & \cdots & f(1,N-1) \\ \vdots & \vdots & \vdots & \vdots \\ f(M-1,0) & f(M-1,1) & \cdots & f(M-1,N-1) \end{pmatrix}. \tag{2.8}$$

A color image has three components. Thus a color image matrix has three of above image pixel matrices for red, green and blue components and they all appear black and white when viewed "individually." We begin with the following duality principle, (i) *spatial* vs (ii) *spectral*, and we illustrate its role for the redundancy, and for correlation of variables, in the resolution-refinement algorithm for *images*. Specifically:

(i) *Spatial Redundancy:* correlation between neighboring pixel values.
(ii) *Spectral Redundancy*: correlation between different color planes or spectral bands.

We are interested in removing these redundancies using correlations.

Starting with a matrix representation for a particular image, we then compute the covariance matrix using the steps from (3) and (4) in algorithm above. We then compute the Karhunen-Loève eigenvalues. Next, the eigenvalues are arranged in decreasing order. The corresponding eigenvectors are arranged to match the eigenvalues with multiplicity. The eigenvalues mentioned here are the same eigenvalues λ_i in step 4 above, thus yielding smallest error and smallest entropy in the computation (see e.g., [Son08]).

The following figure shows the principal components of an image in increasing eigenvalues where the original image is a color png file.

The original file is in red, green and blue color image which had three R, G, B color components. So if I is the original image it can be represented as

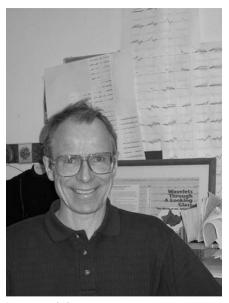$$I = w_1 R + w_2 G + w_3 B = f_R + f_G + f_B \tag{2.9}$$

where w_1, w_2 and w_3 are weights which are determined for different light intensity for color, and f_R, f_G and f_B are the three R, G, B components of the form equation (2.8). Each matrix appears black and white when viewed individually. Now, when PCA is performed on the image I, it gives alternative components. Here the original image I is Figure 2.1(A). The original image used for Figure 2.1(A), is in red, green and blue color components which are f_R, f_G, and f_B.

Here, after PCA transformation, instead of RGB components a new three components are used and this is shown in the Figures 2.1(B) , 2.1(C) and 2.1(D). The principal components of the image are in the order of increasing eigenvalues. A simpler version of this is shown in section 2.3 (Figure 2.2) where the dimension change occurs with PCA in rotation form.

This can show the dimension reduction in section 2.3 where an image data is decomposed into different components by PCA into different 'dimensions'. So instead of RGB components of 3D dimension, we have three principal components for each color value matrix according to covariance matrix. The pixel values are now projected on the new dimensions for each matrix respectively. That is, the original pixel values are represented in the new dimension according to the principal components. There are several ways to compress an image using PCA, but one way to compress a digital image using PCA is by keeping the 'significant' components, i.e., the eigenvectors with large magnitudes. The components that have only small portion of the variation in data for the effect of an image are discarded. This elimination leads to dimension reduction in final image by reducing the quantity of eigenvectors. Thus the dimension of the image matrix is reduced. (Also see e.g., [dES12].)

(A) Original Jorgensen image (in color with RGB components.)

(B) The first component of the image after PCA.

(C) The second component of the image after PCA.

(D) The third component of the image after PCA.

FIGURE 2.1. The three components use the new dimensions to represent the pixel values instead of RGB dimensions. The three components are in the order of increasing eigenvalues.

In order for readers to understand better, in case of wavelet decomposition of the digital image, in this step, PCA is performed on the horizontal, vertical and diagonal details (matrices). In the horizontal detail matrix, the pixel values are correlated. According to the correlation, the pixels that are more frequent and less frequent are determined, according to the magnitude size of the eigenvalues. Then corresponding components are determined. Now we can just keep the first component with the largest magnitude eigenvalue then discard the rest component(s) to reconstruct and image. The idea is this, if one of the details has frequent blue pixels but less frequent light blue and dark blue pixels, the first component will pick up all the blue pixels but not light blue and dark blue pixels. Discarding the light blue and dark blue pixels that are less frequent on the image may not be detectable by human eyes. See e.g., [MP05, KKS16, dES12] for PCA image compression, and [GJ06, DF07, GK15] for discrete wavelet transform and PCA image compression.

Now, on the other hand, PCA can be used for object rotation and this will be discussed further with dimension reduction in data in section 2.3.

2.3. **Dimension Reduction and Principal Component Analysis.** The sections 2.2, 2, and 2.3, are to give background information on dimension reduction building from 2-D image case to linear data case to help readers grasp the understanding of what dimension reduction does to prepare readers for section 3 where dimension reduction for nonlinear data is discussed with results analogous to section 2.5. Also, Figure 2.1 shows how an image data could be decomposed into different principal (dimension) components instead of RGB components..

In the previous section, we discussed how principal component analysis (PCA) was used in image compression where PCA step gave different components of the image matrix. The idea for data dimension reduction is analogous. After all, a digital image is data, too. In this section we give some background illustration on how dimension reduction or dimensionality reduction is done using linear data. Although, PCA doesn't guarantee optimal dimension reduction for all types of linear data, it captures the maximum variability in the data that makes it a popular dimension reduction algorithm for engineers.

In data dimension reduction, we need to determine how to choose the right axes. One method used is principal component analysis (PCA). What PCA does is that it gives a linear subspace of dimension that is lower than the dimension of the original data in such a way that the data points lie mainly in the linear subspace with the lower dimension. Within this subspace with reduced dimension, most variability of the data is maintained. That is, PCA creates a new feature-space (subspace) that captures as much variance in the original data set as possible. The linear subspace is spanned by the orthogonal vectors that form a basis. These orthogonal vectors give principal axes, i.e., directions in the data with the largest variations. This is illustrated in Figure 2.2. As in section 2.1, the PCA algorithm performs the centering of the data by subtracting off the mean, and then determines the direction with the largest variation of the data and chooses an axis in that direction, and then further explores the remaining variation and locates another axis that is orthogonal to the first and explores as much of the remaining variation as possible. This iteration is performed until all possible axes are exhausted. Once we have a principal axis, we subtract the variance along this principal axis to obtain the remaining variance. Then the same procedure is applied again to obtain the next principal axis from the residual variance. In addition to being the direction

of maximum variance, the next principal axis must be orthogonal to the other principal axes. When all the principal axes are obtained, the data set is projected onto these axes. These new orthogonal coordinate axes are also called principal components.

The outcome is all the variation along the axes of the coordinate set, and this makes the covariance matrix diagonal which means each new variable is uncorrelated with the rest of the variables except itself. As for some of the axes that are obtained towards last have very little variation. So they don't contribute much, thus, can be discarded without affecting the variability in the data, hence reducing the dimension (see e.g., [Mar14]).

In machine learning, when dealing with data and plotting results, instead of three dimensions in the data, one usually find two dimensions easier to interpret. More training of the data is necessary as the dimension is higher. In fact, for many algorithms, the dimensionality is an explicit factor for the computational cost. The two different ways of dimensionality reduction are:

(i) *Feature selection*: Goes through the available features and select useful features such as variables or predictors, i.e., correlation to the output variables.

(ii) *Feature extraction*/derivation: Derives new features from the existing ones. This is generally done by dataset transforms that changes the axes of the dimension.

What PCA does is feature extraction/derivation by finding new axes in smaller subspace of the data set space. Unlike in image processing, in machine learning the input data is different, where an image is 2-dimensional while the data is multi-dimensional, and with more complexity (more information to be considered in algorithm to determine reduction of dimension). In this section, we explore PCA on linear data case. The nonlinear data set will be discussed in section 3.
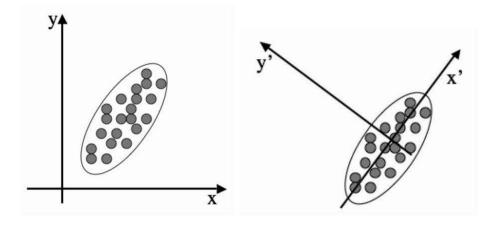

FIGURE 2.2. Two different sets of coordinate axes. The second consists of a rotation and translation of the first and was found using Principal Components Analysis (see [Mar14]).

PCA aims to remove redundancies and describe the data with less properties in a way that it performs a linear transformation moving the original set of data
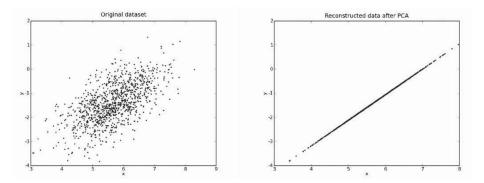
FIGURE 2.3. Computing the principal components of the 2-D data set. Right: Using on the first one to reconstruct it produces the line of data which is along the principal axis of the ellipse that the data was sampled down (see [Mar14]).

to a new space spanned by principal component. This done by constructing a new set of properties based on combination of the old properties. The properties that present low variance are considered not useful. PCA looks for properties that has maximal variation across the data to make the principal component space. The eigenvectors insection 2.1 are the new set of axes of the principal component. Dimension reduction occurs when the eigenvectors with more variance are chosen but those with less variance are discarded.

In Figure 2.3, first look at the following two-dimensional feature space with PCA transform the maximal variance across the data is found and the new axes or principal components are found along the direction with maximum variance. After this step, the data is still in 2-D, we then project the remaining data that is not on the principal axis into the the principal axis. Then every data lies on the principal axis just like in the second image with reconstructed data after PCA on the right. Thus, dimension is reduced.

Now, the cumulative distance between our original data and the projected data is considered measure of information loss. Thus it is crucial that the axis is oriented in a way that minimizes that cumulative distance and this is achieved through the variance.

2.4. **KL Transform/PCA.** We would like to study the orthonormal bases of Karhunen-Loève transform or PCA to see how captures the maximum variability in the data to make linear data dimensionality reduction effective. The discussion below is motivated by [Son08, JS07] among other sources. In computing probabilities and variance, Hilbert space serves as a helpful tool. For example, take a unit vector f in some fixed Hilbert space $\mathcal{H}$, and an orthonormal basis (ONB) ψ_i with i running over an index set I. We now introduce two families of probability measures, one family $P_f(\cdot)$ indexed by $f \in \mathcal{H}$, and a second family P_T indexed by a class of operators $T : \mathcal{H} \to \mathcal{H}$.

Definition 2.1. Let $\mathcal{H}$ be a Hilbert space. Let (ψ_i) and (ϕ_i) be orthonormal bases (ONB), with index set I. Usually

$$I = \mathbb{N} = \{1, 2, ...\}. \tag{2.10}$$

If $(\psi_i)_{i \in I}$ is an ONB, we set $Q_n :=$ the orthogonal projection onto $span\{\psi_1, ..., \psi_n\}$.

We recall Dirac's terminology [Dir47] for rank-one operators in Hilbert space. While there are alternative notation available, Dirac's bra-ket terminology is used for efficiency for our present considerations.

Definition 2.2. Let vectors u, $v \in \mathcal{H}$. Then

$$\langle u, v \rangle = \text{inner product} \in \mathbb{C}, \tag{2.11}$$

$$|u\rangle\langle v| = \text{rank-one operator, } \mathcal{H} \to \mathcal{H}, \tag{2.12}$$

where the operator $|u\rangle\langle v|$ acts as follows

$$|u\rangle\langle v| \, w = |u\rangle \, \langle v, w \rangle = \langle v, w \rangle \, u, \quad \text{for all } w \in \mathcal{H}. \tag{2.13}$$

Proposition 2.3. *Consider an ensemble of a large number N of objects of similar type such as a set of data, of which Nw^α, $\alpha = 1, 2, ..., \nu$ where the relative frequency w^α satisfies the probability axioms:*

$$w^\alpha \geq 0, \quad \sum_{\alpha=1}^{\nu} w^\alpha = 1.$$

Assume that each type specified by a value of the index α is represented by $f^\alpha(\xi)$ in a real domain $[a, b]$, which we can normalize as

$$\int_a^b |f^\alpha(\xi)|^2 d\xi = 1.$$

Let $\{\psi_i(\xi)\}$, $i = 1, 2, ...,$ be a complete set of orthonormal base functions defined on $[a, b]$. Then any function (or data) $f^\alpha(\xi)$ can be expanded as

$$f^\alpha(\xi) = \sum_{i=1}^{\infty} x_i^{(\alpha)} \psi_i(\xi) \tag{2.14}$$

with

$$x_i^\alpha = \int_a^b \psi_i^*(\xi) f^\alpha(\xi) d\xi. \tag{2.15}$$

Here, x_i^α is the component of f^α in ψ_i coordinate system. With the normalization of f^α we have

$$\sum_{i=1}^{\infty} |x_i^\alpha|^2 = 1.$$

Proof. If we substitute (2.15) into (2.14) we have

$$f^\alpha(\xi) = \int_a^b f^\alpha(\xi) \left[\sum_{i=1}^{\infty} \psi_i^*(\xi) \psi_i(\xi) \right] d\xi$$

$$= \sum_{i=1}^{\infty} \langle \psi_i, f^\alpha \rangle \psi_i(\xi)$$

by definition of ONB. Note this involves orthogonal projection. $\qquad \square$

We here give mathematical background of PCA.

Let $\mathcal{H} = L^2(a, b)$. $\psi_i : \mathcal{H} \to l^2(\mathbb{Z})$ and $U : l^2(\mathbb{Z}) \to l^2(\mathbb{Z})$ where U is a unitary operator.

Notice that the distance is invariant under a unitary transformation. Thus, using another coordinate system (principal axis) $\{\phi_j\}$ in place of $\{\psi_i\}$, would preserve the distance. The idea is that when PCA transform is applied on a set of data, the

set of data $\{x_i^\alpha\}$ in the feature space represented in $\{\psi_i\}$ basis are now represented in another coordinate system $\{\phi_j\}$.

Let $\{\phi_j\}$, $j = 1, 2, ...$, be another set of orthonormal basis (ONB) functions instead of $\{\psi_i(\xi)\}$, $i = 1, 2, ...,$. Let y_j^α be the component of f^α in $\{\phi_j\}$ where it can be expressed in terms of x_i^α by a linear relation

$$y_j^\alpha = \sum_{i=1}^\infty \langle \phi_j, \psi_i \rangle x_i^\alpha = \sum_{i=1}^\infty U_{i,j} x_i^\alpha$$

where $U : l^2(\mathbb{Z}) \to l^2(\mathbb{Z})$ is the unitary operator

$$U_{i,j} = \langle \phi_j, \psi_i \rangle = \int_a^b \phi_j^*(\xi)\, \psi_i(\xi)\, d\xi.$$

Also, x_i^α can be written in terms of y_j^α under the following relation

$$x_i^\alpha = \sum_{j=1}^\infty \langle \psi_i, \phi_j \rangle y_j^\alpha = \sum_{j=1}^\infty U_{i,j}^{-1} y_j^\alpha$$

where $U_{i,j}^{-1} = \overline{U_{i,j}}$ and $\overline{U_{i,j}} = U_{j,i}^*$. Thus,

$$f^\alpha(\xi) = \sum_{i=1}^\infty x_i^\alpha(\xi)\, \psi_i(\xi) = \sum y_i^\alpha(\xi)\, \phi_i(\xi).$$

So $U(x_i) = (y_i)$ which is coordinate change, and $\sum_{i=1}^\infty x_i^\alpha \psi_i(\xi) = \sum_{j=1}^\infty y_j^\alpha \phi_j(\xi)$, and

$$x_i^\alpha = \langle \psi_i, f^\alpha \rangle = \int_a^b \psi_i^*(\xi)\, f^{(\alpha)}(\xi)\, d\xi.$$

The squared magnitude $|x_i^{(\alpha)}|^2$ of the coefficient for ψ_i in the expansion of $f^{(\alpha)}$ can be considered as a good measure of the average in the ensemble

$$Q_i = \sum_{\alpha=1}^n w^{(\alpha)} |x_i^{(\alpha)}|^2,$$

and as a measure of importance of $\{\psi_i\}$. Notice,

$$Q_i \geq 0, \quad \sum_i Q_i = 1.$$

See also [Wat67].

Let $G(\xi, \xi') = \sum_\alpha w^\alpha f^\alpha(\xi) f^{\alpha*}(\xi')$. Then G is a Hermitian matrix that is the covariance matrix and $Q_i = G(i, i) = \sum_\alpha w^\alpha x_i^\alpha x_i^{\alpha*}$. Here, $Q_i = G(i, i)$ is the variance and $G(i, j)$ determines the covariance between x_i and x_j. The normalization $\sum Q_i = 1$ gives us trace $G = 1$, where the trace means the diagonal sum.

Then define a special function system $\{\Theta_k(\xi)\}$ as the set of eigenfunctions of G, i.e.,

$$\int_a^b G(\xi, \xi')\, \Theta_k(\xi')\, d\xi' = \lambda_k \Theta_k(\xi). \tag{2.16}$$

So $G\Theta_k(\xi) = \lambda_k \Theta_k(\xi)$. $Also, U : l^2(\mathbb{Z}) \to l^2(\mathbb{Z})$ is the unitary operator consisting of eigenfunctions of G in its columns. These eigenfunctions represent the directions of the largest variance of the data and the corresponding eigenvalues represent the magnitude of the variance in the directions. PCA allows us to choose the principal

components so that the covariance matrix G of the projected data is as large as possible. The largest eigenfunction of the covariance matrix points to the direction of the largest variance of the data and the magnitude of this function is equal to the corresponding eigenvalue. The subsequent eigenfunctions are always orthogonal to the largest eigenfunctions.

When the data are not functions but vectors v^αs whose components are $x_i^{(\alpha)}$ in the ψ_i coordinate system, we have

$$\sum_{i'} G\left(i, i'\right) t_{i'}^k = \lambda_k t_i^k \tag{2.17}$$

where t_i^k is the ith component of the vector Θ_k in the coordinate system $\{\psi_i\}$. So we get $\psi : \mathcal{H} \to (x_i)$ and also $\Theta : \mathcal{H} \to (t_i)$. The two ONBs result in

$$x_i^\alpha = \sum_k c_k^\alpha t_i^k \text{ for all } i, \quad c_k^\alpha = \sum_i t_i^{k*} x_i^\alpha,$$

which is the Karhunen-Loève expansion of $f^\alpha(\xi)$ or vector v^α. Hence $\{\Theta_k(\xi)\}$ is the K-L coordinate system dependent on $\{w^\alpha\}$ and $\{f^\alpha(\xi)\}$. Then we arrange the corresponding eigenfunctions or eigenvectors in the order of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{k-1} \geq \lambda_k \geq \ldots$ in the columns of U.

Now, $Q_i = G_{i,i} = \langle \psi_i, G\psi_i \rangle = \sum_k A_{ik} \lambda_k$ where $A_{ik} = t_i^k t_i^{k*}$ which is a double stochastic matrix. Then we have the following eigendecomposition of the covariance matrix (operator), G

$$G = U \begin{pmatrix} \lambda_1 & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & \lambda_k \end{pmatrix} U^{-1}. \tag{2.18}$$

2.5. **Principal Component Analysis and Maximal Variance.** In the present section, we would like to study the orthonormal bases of Karhunen-Loève transform or PCA to see how it captures the maximal variance in the linear data to effectively perform dimensionality reduction. We want to show how PCA captures maximal variability in the data.

For later use, and for the benefit of the reader, we shall recall here some definitions and results from [BJ02, JS07]. With the lemmas below in the present section, we are aiming at the results Theorem 2.12. Here we are adapting these results to our Principal Component Analysis (PCA) on data, and dimension reduction algorithms for both linear data set; see details in the subsequent sections for nonlinear data set, especially section 3 and our Theorem 3.11.

Let $\mathcal{H}$ be a Hilbert space which realizes trace class G as a self-adjoint operator.

Definition 2.4. $T \in B(\mathcal{H})$ is said to be trace class if and only if the series $\sum \langle \psi_i, |T| \psi_i \rangle$, with $|T| = \sqrt{T^*T}$, is convergent for some ONB (ψ_i). In this case, set

$$tr(T) := \sum \langle \psi_i, T\psi_i \rangle. \tag{2.19}$$

Definition 2.5. A sequence $(h_\alpha)_{\alpha \in A}$ in $\mathcal{H}$ is called a *frame* if there are constants $0 < c_1 \leq c_2 < \infty$ such that

$$c_1 \|f\|^2 \leq \sum_{\alpha \in A} |\langle h_\alpha, f \rangle|^2 \leq c_2 \|f\|^2 \text{ for all } f \in \mathcal{H}. \tag{2.20}$$

Also see [BJ02, HKLW07, HWW05, BH19, CCEL15, CCK13].

Lemma 2.6. *Let* $(h_\alpha)_{\alpha \in A}$ *be a frame in* $\mathcal{H}$. *Set* $L : \mathcal{H} \to l^2$,

$$L : f \mapsto (\langle h_\alpha, f \rangle)_{\alpha \in A}. \tag{2.21}$$

Then $L^* : l^2 \to \mathcal{H}$ *is given by*

$$L^*((c_\alpha)) = \sum_{\alpha \in A} c_\alpha h_\alpha \tag{2.22}$$

where $(c_\alpha) \in l^2$; *and*

$$L^*L = \sum_{\alpha \in A} |h_\alpha \rangle \langle h_\alpha|. \tag{2.23}$$

Definition 2.7. Suppose we are given $(f_\alpha)_{\alpha \in A}$, a frame, non-negative numbers $\{w_\alpha\}_{\alpha \in A}$, where A is an index set, with $\|f_\alpha\| = 1$, for all $\alpha \in A$.

$$G := \sum_{\alpha \in A} w_\alpha |f_\alpha \rangle \langle f_\alpha| \tag{2.24}$$

is called a **frame** operator associated to (f_α).

Remark 2.8. If we take vectors (f_α) from a frame (h_α) and normalize them such that $h_\alpha = \|h_\alpha\| f_\alpha$, and $w_\alpha := \|h_\alpha\|^2$, then L^*L has the form (2.24) and it becomes to covariance matrix G in section 2.4. Thus $G = L^*L : \mathcal{H} \to \mathcal{H}$.

Lemma 2.9. *Let* G *be as in (2.24). Then* G *is trace class if and only if* $\sum_\alpha w_\alpha < \infty$; *and then*

$$tr(G) = \sum_{\alpha \in A} w_\alpha. \tag{2.25}$$

Definition 2.10. Suppose we are given a frame operator

$$G = \sum_{\alpha \in A} w_\alpha |f_\alpha \rangle \langle f_\alpha| \tag{2.26}$$

and an ONB (ψ_i). Then for each n, the numbers

$$E_n^\psi = \sum_{\alpha \in A} w_\alpha \|f_\alpha - \sum_{i=1}^n \langle \psi_i, f_\alpha \rangle \psi_i\|^2 \tag{2.27}$$

are called the *error* or the *residual* of the projection.

Lemma 2.11. *When* (ψ_i) *is given, set* $Q_n := \sum_{i=1}^n |\psi_i \rangle \langle \psi_i|$ *and* $Q_n^\perp = I - Q_n$ *where* I *is the identity operator in* $\mathcal{H}$. *Then (see (2.27))*

$$E_n^\psi = tr(GQ_n^\perp). \tag{2.28}$$

The more general frame operators are as follows: Let

$$G = \sum_{\alpha \in A} w_\alpha P_\alpha \tag{2.29}$$

where (P_α) is an indexed family of projections in $\mathcal{H}$, i.e., $P_\alpha = P_\alpha^* = P_\alpha^2$, for all $\alpha \in A$. Note that P_α is trace class if and only if it is finite-dimensional, i.e., if and only if the subspace $P_\alpha \mathcal{H} = \{x \in \mathcal{H} \mid P_\alpha x = x\}$ is finite-dimensional. See also Lemma 2.6 and Definition 2.7.

As indicated by the name PCA, the scheme involves a choice of "principal components," often realized as a finite-dimensional subspace of a global (called latent) data set. We now briefly outline two views of PCA. We begin with the simplest case

of consideration of covariance operators, and in the next section we turn to kernel PCA, referring to a class of reproducing kernels, as used in learning theory. In the latter case, one identifies principal features for the machine learning algorithm.

As outlined, the simplest way to identify a PCA subspace is to turn to a covariance operator, say G, acting on the global data; see Theorem 2.12 below. With the use of a suitable Karhunen-Loève transform or PCA, and via a system of i.i.d. standard Gaussians, one may often arrive at a covariance operator which is of trace class. An application of the spectral theorem to this associated operator G (see (2.30) below), we then get an algorithm for computing eigenspaces corresponding to the top of the spectrum of G, i.e., the subspace spanned by the eigenvectors for the top n eigenvalues; see (2.31). These subspaces will then be principal components of order n since the contribution from the span of the remaining eigenspaces will be negligible.

A second approach to PCA is based on an analogous identification of principal component subspaces, but with the optimization involving maximum likelihood, or minimization of "cost."

Now, although PCA is used popularly in linear data dimension reductions as PCA decorrelates data, it is noted that the decorrelation only corresponds to statistical independence in the Gaussian case. So PCA is not generally the optimal choice for linear data dimension reduction. However, PCA captures maximal variability in the data.

PCA enables finding projections which maximize the variance: The first principal component is the direction in the feature space along which gives projections with the largest variance. The second principal component is the variance maximizing direction to all directions orthogonal to the first principal component. The ith component is the direction which maximizes variance orthogonal to the $i-1$ previous components. Thus, PCA captures maximal variability then projects a set of data in higher dimensional feature space to a lower dimensional feature space orthogonally and this will be proved in Theorem 2.12. In fact, it can be observed from the proof of Theorem 4.13 in [JS07].

Below, we formulate the iterative algorithm (see eq. (2.31)) of producing principal components in the context of trace class operators.

Theorem 2.12. *The Karhunen-Loève ONB with respect to the frame operator $G = L^*L$ gives the smallest error in the approximation to a frame operator and the covariance operator G gives maximum variance.*

Proof. Given the covariance operator G which is trace class and positive semidefinite, applying the spectral theorem to G results is a discrete spectrum, with the natural order $\lambda_1 \geq \lambda_2 \geq ...$ and a corresponding ONB (ϕ_k) consisting of eigenvectors, i.e.,

$$G\phi_k = \lambda_k \phi_k, \ k \in \mathbb{N}, \tag{2.30}$$

called the Karhunen-Loève data or principal components. The spectral data may be constructed recursively starting with

$$\lambda_1 = \sup_{\phi \in \mathcal{H}, \, \|\phi\|=1} \langle \phi, G\phi \rangle = \langle \phi_1, G\phi_1 \rangle, \text{ and}$$
$$\lambda_{k+1} = \sup_{\substack{\phi \in \mathcal{H}, \, \|\phi\|=1 \\ \phi \perp \phi_1, \phi_2, ..., \phi_k}} \langle \phi, G\phi \rangle = \langle \phi_{k+1}, G\phi_{k+1} \rangle. \tag{2.31}$$

This way, the maximal variance is achieved. Now by applying [AK06, Thm 4.1] we have

$$\sum_{k=1}^{n} \lambda_k \geq tr\left(Q_n^{\psi} G\right) = \sum_{k=1}^{n} \langle \psi_k, G\psi_k \rangle \quad \text{for all } n, \qquad (2.32)$$

where Q_n^{ψ} is the sequence of projections, deriving from some ONB (ψ_i) and are arranged such that the following holds:

$$\langle \psi_1, G\psi_1 \rangle \geq \langle \psi_2, G\psi_2 \rangle \geq ... \quad .$$

Hence we are comparing ordered sequences of eigenvalues with sequences of diagonal matrix entries.

Lastly, we have

$$tr\left(G\right) = \sum_{k=1}^{\infty} \lambda_k = \sum_{k=1}^{\infty} \langle \psi_k, G\psi_k \rangle < \infty.$$

The assertion in Theorem 2.12 is the validity of

$$E_n^{\phi} \leq E_n^{\psi} \qquad (2.33)$$

for all $(\psi_i) \in ONB(\mathcal{H})$, and all $n = 1, 2, ...$; and moreover, that the infimum on the RHS in (2.33) is attained for the KL-ONB (ϕ_k). But in view of our lemma for E_n^{ψ} (2.11), we see that (2.33) is equivalent to the system (2.32) in the Arveson-Kadison theorem. □

The Arveson-Kadison theorem is the assertion (2.32) for trace class operators, see e.g., refs [Arv07] and [AK06]. That (2.33) is equivalent to (2.32) follows from the definitions.

3. Kernel PCA

The sections 2.2 and 2.3 gave background on dimension reduction building from 2-D image case to linear data case. In this section we discuss and show results on dimension reduction for nonlinear data is discussed with results analogous to [JS07]. We shall focus here on their use in both principal component analysis (PCA), and kernel PCA (KPCA). In the applications below, we stress an important finite-dimensional setting.

In previous section, PCA was used in data dimension reduction on linear case. However, this cannot be done on nonlinear case and thus kernel principal component analysis (KPCA) is used for nonlinear dimension reduction. See, e.g., [SZ09a, SZ09b, SZ07, SY06, PS03, CS02], and Example 3.16.

Standard PCA is effective at identifying linear subspaces carrying the greatest variance in a data set. However, this method is not able to detect nonlinear submanifolds. A popular technique to tackle the latter case is kernel PCA. It first maps data into a higher dimensional space $\mathcal{H}(K)$, and performs PCA there. Here $\mathcal{H}(K)$ is the reproducing kernel Hilbert space (RKHS) associated with a given positive definite kernel K. The mapping in this context presumably sends a nonlinear submanifold in the input space to a linear subspace in $\mathcal{H}(K)$. For example, in classification problems, a kernel is usually chosen so that the mapped data can be separated by a linear decision boundary in $\mathcal{H}(K)$ (see Figure 3.1).

Remark 3.1. It would be intriguing to compare Smale's Dimension Reduction algorithm from [SZ09a] with ours. The two approaches are along a different lines of development.

The approach in Belkin's paper [BN03] is popular in current Machine Learning research. Both our results and those of Belkin et al aim for dimension reduction algorithms. Other methods exist, which constitute variants of KPCA, but with different choices of kernels, and with the Laplacian eigenmap (LE) as one of them. (See also [CWG19, VVQCR$^+$19, TF19, SGS$^+$19].) For recent developments on graph Laplacians, and Perron-Frobenius eigenfunctions as principal components, we refer to e.g., [BJ02].

Definition 3.2. Let S be a set. A positive definite (p.d.) kernel on S is a function $K : S \times S \to \mathbb{C}$, such that

$$\sum_{i,j=1}^{N} \overline{c_i} c_j K\left(v_i, v_j\right) \geq 0 \tag{3.1}$$

for all $\{x_i\}_{i=1}^{N} \subset S$, $\{c_i\}_{i=1}^{N} \subset \mathbb{C}$, and $N \in \mathbb{N}$.

Given a p.d. kernel as in (3.1), there exists a reproducing kernel Hilbert space (RKHS) $\mathcal{H}\left(K\right)$ and a mapping $\Phi : S \to \mathcal{H}\left(K\right)$ such that

$$K\left(x,y\right) = \left\langle \Phi\left(x\right), \Phi\left(y\right)\right\rangle_{\mathcal{H}(K)}. \tag{3.2}$$

The function Φ in (3.2) is called a *feature map* for the problem.

Moreover, the following reproducing property holds:

$$f\left(x\right) = \left\langle K_x, f\right\rangle_{\mathcal{H}(K)}, \tag{3.3}$$

for all $f \in \mathcal{H}\left(K\right)$, and $x \in S$.

Remark 3.3. $\mathcal{H}\left(K\right)$ may be chosen as the Hilbert completion of

$$span\left\{K_x := K\left(\cdot, x\right)\right\} \tag{3.4}$$

with respect to the $\mathcal{H}\left(K\right)$-inner product

$$\left\langle \sum c_i K_{x_i}, \sum d_j K_{x_j} \right\rangle_{\mathcal{H}(K)} := \sum \overline{c_i} d_j K\left(x_i, x_j\right). \tag{3.5}$$

Initially the LHS in formula (3.5) only refers to finite linear combinations. Hence, the vector space (3.4) becomes a pre-Hilbert space. The RKHS $\mathcal{H}(K)$ itself then results from the standard Hilbert completion. It is this Hilbert space we will use in our subsequent study of optimization problems, and in our KPCA-dimension reduction. Sections 3.1–3.3 deal with separate issues of kernel-optimization. Before turning to these, however, we will first introduce a setting of Hilbert-Schmidt operators. This will play a crucial role in the formulation of our main result, Theorem 3.11 in section 3.3.

Recall that a data set $(x_j)_{j=1}^{n}$, $x_j \in \mathbb{C}^m$, may be viewed as an $m \times n$ matrix X, where x_j is the jth column vector. Here, m is the number of features, and n the number of sample points. The total variance is

$$\|X\|_{HS}^2 = \sum |x_{ij}|^2 = tr\left(X^*X\right), \tag{3.6}$$

and $\|\cdot\|_{HS}$ in (3.6) denotes the Hilbert-Schmidt norm.

Remark 3.4. Let $\mathcal{H}$ be a Hilbert space, and let $HS(\mathcal{H})$ be the Hilbert-Schmidt operators $\mathcal{H} \xrightarrow{X} \mathcal{H}$ with inner product

$$\langle X, Y \rangle = tr(X^*Y). \tag{3.7}$$

Then the two Hilbert spaces $HS(\mathcal{H})$, and $\mathcal{H} \otimes \overline{\mathcal{H}}$ (tensor-product), are naturally isometrically isomorphic via

$$HS(\mathcal{H}) \ni |u\rangle\langle v| \longrightarrow u \otimes \overline{v} \in \mathcal{H} \otimes \overline{\mathcal{H}}, \tag{3.8}$$

see 2.2, the ket-bra notation (2.12). Indeed,

$$\||u\rangle\langle v|\|_{HS}^2 = \|u\|_{\mathcal{H}}^2 \|v\|_{\mathcal{H}}^2,$$

and the assertion follows from isometric extension of (3.8).

3.1. **Application to Optimization.** One of the more recent applications of kernels and the associated reproducing kernel Hilbert spaces (RKHS) is to optimization, also called kernel-optimization. See [YLTL18, LLL11]. In the context of machine learning, it refers to training-data and feature spaces. In the context of numerical analysis, a popular version of the method is used to produce splines from sample points; and to create best spline-fits. In statistics, there are analogous optimization problems going by the names "least-square fitting," and "maximum-likelihood" estimation. In the latter instance, the object to be determined is a suitable probability distribution which makes "most likely" the occurrence of some data which arises from experiments, or from testing.

What these methods have in common is a minimization (or a max problem) involving a "quadratic" expression Q with two terms. The first in Q measures a suitable $L^2(\mu)$-square applied to a difference of a measurement and a "best fit." The latter will then to be chosen from anyone of a number of suitable reproducing kernel Hilbert spaces (RKHS). The choice of kernel and RKHS will serve to select desirable features. So we will minimize a quantity Q which is the sum of two terms as follows: (i) a L^2-square applied to a difference, and (ii) a penalty term which is a RKHS norm-squared. (See eq. (3.10).) In the application to determination of splines, the penalty term may be a suitable Sobolev normed-square; i.e., L^2 norm-squared applied to a chosen number of derivatives. Hence non-differentiable choices will be "penalized."

In all of the cases, discussed above, there will be a good choice of (i) and (ii), and we show that there is then an explicit formula for the optimal solution; see eq (3.13) in Theorem 3.5 below.

Let X be a set, and let $K : X \times X \longrightarrow \mathbb{C}$ be a positive definite (p.d.) kernel. Let $\mathcal{H}(K)$ be the corresponding reproducing kernel Hilbert space (RKHS). Let $\mathscr{B}$ be a sigma-algebra of subsets of X, and let μ be a positive measure on the corresponding measure space $(X, \mathscr{B})$. We assume that μ is sigma-finite. We shall further assume that the associated operator T given by

$$\mathcal{H}(K) \ni f \xrightarrow{T} (f(x))_{x \in X} \in L^2(\mu) \tag{3.9}$$

is densely defined and closable.

Fix $\beta > 0$, and $\psi \in L^2(\mu)$, and set

$$Q_{\psi,\beta}(f) = \|\psi - Tf\|_{L^2(\mu)}^2 + \beta \|f\|_{\mathcal{H}(K)}^2 \tag{3.10}$$

defined for $f \in \mathcal{H}(K)$, or in the dense subspace $dom\,(T)$ where T is the operator in (3.9). Let

$$L^2\left(\mu\right) \xrightarrow{\;T^*\;} \mathcal{H}\left(K\right) \tag{3.11}$$

be the corresponding adjoint operator, i.e.,

$$\langle F, T^*\psi\rangle_{\mathcal{H}(K)} = \langle Tf, \psi\rangle_{L^2(\mu)} = \int_X \overline{f\left(s\right)}\psi\left(s\right) d\mu\left(s\right). \tag{3.12}$$

Theorem 3.5. *Let K, μ, ψ, β be as specified above; then the optimization problem*

$$\inf_{f \in \mathcal{H}(K)} Q_{\psi,\beta}\left(f\right)$$

has a unique solution F in $\mathcal{H}\left(K\right)$, it is

$$F = \left(\beta I + T^*T\right)^{-1} T^*\psi \tag{3.13}$$

where the operator T and T^* are as specified in (3.9)-(3.12).

Proof. (Sketch) We fix F, and assign $f_\varepsilon := F + \varepsilon h$ where h varies in the dense domain $dom\,(T)$ from (3.9). For the derivative $\frac{d}{d\varepsilon}\big|_{\varepsilon=0}$ we then have:

$$\frac{d}{d\varepsilon}\big|_{\varepsilon=0} Q_{\psi,\beta}\left(f_\varepsilon\right) = 2\Re\left\langle h, \left(\beta I + T^*T\right) F - T^*\psi\right\rangle_{\mathcal{H}(K)} = 0$$

for all h in a dense subspace in $\mathcal{H}\left(K\right)$. The desired conclusion follows. $\qquad\square$

Least-square Optimization

We now specialize the optimization formula from Theorem 3.5 to the problem of minimize a "quadratic" quantity Q. It is still the sum of two individual terms: (i) a L^2-square applied to a difference, and (ii) a penalty term which is the RKHS norm-squared. But the least-square term in (i) will simply be a sum of a finite number of squares of differences; hence "least-squares." As an application, we then get an easy formula (Theorem 3.6) for the optimal solution.

Let K be a positive definite kernel on $X \times X$ where X is an arbitrary set, and let $\mathcal{H}\left(K\right)$ be the corresponding reproducing kernel Hilbert space (RKHS). Let $m \in \mathbb{N}$, and consider sample points:

$\{t_j\}_{j=1}^m$ as a finite subset in X, and

$\{y_i\}_{i=1}^m$ as a finite subset in $\mathbb{R}$, or equivalently, a point in $\mathbb{R}^m$.

Fix $\beta > 0$, and consider $Q = Q_{(\beta,t,y)}$, defined by

$$Q\left(f\right) = \sum_{i=1}^m \underbrace{|f\left(t_i\right) - y_i|^2}_{\text{least square}} + \underbrace{\beta\,\|f\|_{\mathcal{H}(K)}^2}_{\text{penalty form}}, \quad f \in \mathcal{H}\left(K\right). \tag{3.14}$$

We introduce the associated dual pair of operators as follows:

$$T : \mathcal{H}\left(K\right) \longrightarrow \mathbb{R}^m \simeq l_m^2, \text{ and} \tag{3.15}$$
$$T^* : l_m^2 \longrightarrow \mathcal{H}\left(K\right)$$

where

$$Tf = \left(f\left(t_i\right)\right)_{i=1}^m, \quad f \in \mathcal{H}\left(K\right); \text{ and} \tag{3.16}$$

$$T^*y = \sum_{i=1}^m y_i K\left(\cdot, t_i\right) \in \mathcal{H}\left(K\right), \tag{3.17}$$

for all $\vec{y} = \left(y_i\right) \in \mathbb{R}^m$.

Note that the duality then takes the following form:

$$\langle T^*y, f \rangle_{\mathcal{H}(K)} = \langle y, Tf \rangle_{l_m^2}, \quad \forall f \in \mathcal{H}(K), \, \forall y \in l_m^2; \tag{3.18}$$

consistent with (3.12).

Applying Theorem 3.5 to the counting measure

$$\mu = \sum_{i=1}^m \delta_{t_i} = \delta_{\{t_i\}}$$

for the set of sample points $\{t_i\}_{i=1}^m$, we get the two formulas:

$$T^*Tf = \sum_{i=1}^m f(t_i) K(\cdot, t_i) = \sum_{i=1}^m f(t_i) K_{t_i}, \text{ and} \tag{3.19}$$

$$TT^*y = K_m \vec{y} \tag{3.20}$$

where K_m denotes the $m \times m$ matrix

$$K_m = (K(t_i, t_j))_{i,j=1}^m = \begin{pmatrix} K(t_1, t_1) & \cdots & \cdots & K(t_1, t_m) \\ K(t_2, t_1) & \cdots & \cdots & K(t_2, t_m) \\ \vdots & & & \\ K(t_m, t_1) & & & K(t_m, t_m) \end{pmatrix}. \tag{3.21}$$

Theorem 3.6. *Let K, X, $\{t_i\}_{i=1}^m$, and $\{y_i\}_{i=1}^m$ be as above, and let K_m be the induced sample matrix (3.21).*

Fix $\beta > 0$; consider the optimization problem with

$$Q_{\beta, \{t_i\}, \{y_i\}}(f) = \sum_{i=1}^m |y_i - f(t_i)|^2 + \beta \|f\|_{\mathcal{H}(K)}^2, \quad f \in \mathcal{H}(K). \tag{3.22}$$

Then the unique solution to (3.22) is given by

$$F(\cdot) = \sum_{i=1}^m (K_m + \beta I_m)_i^{-1} K(\cdot, t_i) \text{ on } X; \tag{3.23}$$

i.e., $F = \arg\min Q$ on $\mathcal{H}(K)$.

Proof. From Theorem 3.5, we get that the unique solution $F \in \mathcal{H}(K)$ is given by:

$$\beta F + T^*TF = T^*y,$$

and by (3.19)-(3.20), we further get

$$\beta F(\cdot) = \sum_{i=1}^m (y_i - F(t_i)) K(\cdot, t_i) \tag{3.24}$$

where the dot $\cdot$ refers to a free variable in X. An evaluation of (3.24) on the sample points yields:

$$\beta \vec{F} = K_m \left(\vec{y} - \vec{F} \right) \tag{3.25}$$

where $\vec{F} := (F(t_i))_{i=1}^m$, and $\vec{y} = (y_i)_{i=1}^m$. Hence

$$\vec{F} = (\beta I_m + K_m)^{-1} K_m \vec{y}. \tag{3.26}$$

Now substitute (3.26) into (3.25), and the desired conclusion in the theorem follows. We used the matrix identity

$$I_m - (\beta I_m + K_m)^{-1} K_m = \beta (\beta I_m + K_m)^{-1}.$$

$\square$

3.2. **The Case of Gaussian Fields.** For a number of applications, it will be convenient to consider general stochastic processes (X_s) indexed by $s \in S$, where S is merely a *set*; so not *a priori* equipped with any additional structure. Consideration of stochastic processes will always assume some fixed probability space, $(\Omega, \mathscr{A}, \mathbb{P})$ where Ω is a set of sample points; $\mathscr{A}$ is a σ-algebra of events, fixed at the outset; and $\mathbb{P}$ is a probability measure defined on $\mathscr{A}$. A given process $(X_s)_{s \in S}$ is then said to be *Gaussian* and centered iff (Def.) for all choice of finite subsets of S $(s_1, s_2, \cdots, s_N)$, then the system of random variables $\{X_{s_i}\}_{i=1}^{N}$ is jointly Gaussian, i.e., the joint distribution of $\{X_{s_i}\}_{i=1}^{N}$ on $\mathbb{R}^N$ is the Gaussian $g_N(x_1, \cdots, x_N)$ which has mean zero, and covariance matrix

$$K_{ij}^{(N)} := \left(\mathbb{E}\left(X_{s_i} X_{s_j} \right) \right)_{i,j=1}^{N}; \tag{3.27}$$

so for $x = (x_1, \cdots, x_N) \in \mathbb{R}^N$,

$$g_N(x) = (2\pi)^{-N/2} \det\left(K^{(N)} \right)^{-1/2} e^{-\frac{1}{2} x^T K^{(N)^{-1}} x}. \tag{3.28}$$

If $A_N \subset \mathbb{R}^N$ is a Borel set, then

$$\mathbb{P}\left((X_{s_1}, \cdots, X_{s_N}) \in A_N \right) = \int_A g_N(x) \, d^N x \tag{3.29}$$

holds. Note we consider the joint distributions for all finite subsets of S.

Let S be a set, and let $\{X_s\}_{s \in S}$ be a Gaussian process with $\mathbb{E}(X_s) = 0$, $\forall s \in S$; and with

$$\mathbb{E}\left(\overline{X}_s X_t \right) := K_X(s, t) \tag{3.30}$$

as its covariance kernel. Finally, let $\mathcal{H}(K)$ be the corresponding RKHS.

Then the following general results hold (see e.g., [AJL11, AJ12, AJS14, AJ15, JT15, JT16a, JT16b, AJL17, JT18a, JT18b]):

(i) Every positive definite kernel $S \times S \xrightarrow{K} \mathbb{C}$ arises as in (3.30) from some Gaussian process $\{X_s\}_{s \in S}$.

(ii) Assume $\mathcal{H}(K)$ is separable; then we have a representation $\{f_n\}_{n \in \mathbb{N}}$ for a system of functions $f_n : S \to \mathbb{C}$, $n \in \mathbb{N}$,

$$K(s, t) = \sum_{n \in \mathbb{N}} \overline{f_n(s)} f_n(t), \tag{3.31}$$

absolutely convergent on $S \times S$.

(iii) A system $\{f_n\}_{n \in \mathbb{N}}$ satisfies (ii) if and only if it forms a *Parseval frame* in $\mathcal{H}(K)$.

(iv) Given (3.31), then, for every sequence of independent identically distributed (i.i.d.) Gaussian system $\{Z_n\}_{n \in \mathbb{N}}$, $Z_n \sim N(0, 1)$, i.e., each Z_n is a standard Gaussian random variable, $\mathbb{E}(Z_n) = 0$, $\mathbb{E}(Z_n Z_m) = \delta_{n,m}$; the representation

$$X_s(\cdot) = \sum_{n \in \mathbb{N}} f_n(s) Z_n(\cdot) \tag{3.32}$$

is valid in L^2 of the underlying probability space.

Important Note 3.7. When a fixed Gaussian process (X_s) is given, then the associated decomposition (3.32) is called a Karhunen-Loève (KL) transform for X_s. The conclusion from (i)–(iv) above is that there is a direct connection between the two KL transforms; (a) the relatively better known KL-transforms for positive definite kernels (Theorem 2.12 above and [JS07] Theorem 4.15), on the one hand; and (b) the corresponding KL transform for Gaussian processes (see (i)), on the other. This correspondence will be further studied in section 4 below, see especially Corollary 4.5.

The object in principal component analysis (PCA) is to find optimal representations; and to select from them the "leading terms", the principal components.

Remark 3.8. The present general kernel framework (RKHSs and Gaussian processes) encompasses the special case we outlined in [JS07] Example 3.1. By way of comparison, note that the particular positive definite kernel in the latter example is only a special case of the present ones, see (3.31) and (3.32). These types of kernels are often referred to as the case of Mercer kernels; see also [SY06, SZ07, SZ09a, SZ09b]. A Mercer kernel is continuous, and it defines a trace class operator, as illustrated in the example. This latter feature in turn leads to a well defined "top part of the spectrum." And this then allows us to select the principal components; i.e., the maximally correlated variables. We shall show, in Theorem 3.11 below, that there is an alternative approach to principal components which applies to the general class of positive definite kernels, and so goes far beyond the case of Mercer kernels.

3.3. **Optimization and Frames.** Fix a p.d. kernel K on $S := \mathbb{C}^m$, i.e., a functional $K : \mathbb{C}^m \times \mathbb{C}^m \to \mathbb{C}$ satisfying (3.1); and let $\mathcal{H}(K)$ be the associated RKHS. In PCA, one solves the quadratic optimization problem:

$$argmax \left\{ \|QX\|_{HS}^2 : \|Q\|_{HS}^2 = k \right\}, \tag{3.33}$$

where Q runs through all rank-k (self-adjoint) projections in the input space $\mathbb{C}^m$.

Kernel PCA, by contrast, solves a similar problem in $\mathcal{H}(K)$:

$$argmax \left\{ \|Q\Phi(X)\|_{HS}^2 : \|Q\|_{HS}^2 = k \right\}, \tag{3.34}$$

where

$$\Phi(X) = \begin{bmatrix} \Phi(x_1) & \cdots & \Phi(x_n) \end{bmatrix}. \tag{3.35}$$

It is understood that $\|\cdot\|_{HS}$ as in (3.34) refers to the Hilbert-Schmidt class in $B(\mathcal{H}(K))$.

Indeed, both (3.33) and (3.34) are finite dimensional instances of Theorem 2.12. See section 2.5 and details below.

A Finite Frame in $\mathcal{H}(K)$

Let X, K, and Φ be as above. Then $(\Phi(x_j))_{j=1}^n$ is a finite frame whose span $\mathcal{H}_\Phi$ is a closed subspace in $\mathcal{H}(K)$.

Set $L : \mathcal{H}(K) \to \mathbb{C}^n$ by

$$Lf = \sum_{j=1}^n \langle \Phi(x_j), f \rangle_{\mathcal{H}(K)} \delta_j, \ f \in \mathcal{H}(K), \tag{3.36}$$

where δ_j is the standard ONB in $\mathbb{C}^m$. The adjoint $L^* : \mathbb{C}^n \to \mathcal{H}(K)$ is given by

$$L^* c = \sum_{j=1}^{n} \Phi(x_j) c_j, \ c = (c_j) \in \mathbb{C}^n. \qquad (3.37)$$

It follows that

$$L^* L = \sum_{j=1}^{n} |\Phi(x_j)\rangle\langle\Phi(x_j)| = \Phi(X)\Phi(X)^*. \qquad (3.38)$$

(See (3.35), and the frame operator in (2.26).)

Lemma 3.9. *Let L^*L be as in (3.38), and let*

$$G := L^* L = \sum \lambda_j^2 |\phi_j\rangle\langle\phi_j|$$

be the corresponding spectral representation, with $\lambda_1^2 \geq \lambda_2^2 \geq \cdots$. Then

$$|\phi_1\rangle\langle\phi_1| = argmax \left\{ \|Q_1 \Phi(X)\|_{HS}^2 : \|Q_1\|_{HS}^2 = 1 \right\} \qquad (3.39)$$
$$= argmax \left\{ tr(Q_1 G) : \|Q_1\|_{HS}^2 = 1 \right\};$$

where Q_1 runs through all rank-1 projections.

Equivalently, the best rank-1 approximation to G is

$$\lambda_1^2 |\phi_1\rangle\langle\phi_1|.$$

Remark 3.10. Note that the conclusion of the lemma yields a solution the optimization problem we introduced above. Indeed, in the statement of the lemma (see (3.39)) we use the standard notation *argmax* for the data which realizes a particular optimization. In the present case, we are maximizing a certain quadratic expression over the unit-ball in the Hilbert-Schmidt operators. The Hilbert-Schmidt norm is designated with the subscript HS. Part of the conclusion of the lemma asserts that the maximum, as specified in (3.39), is attained for a definite rank-one operator.

Proof of Lemma 3.9. Note that

$$\|Q_1 \Phi(X)\|_{HS}^2 = tr(Q_1 \Phi(X)\Phi(X)^*) = tr(Q_1 G).$$

Let w be a unit vector in $\mathcal{H}(K)$, and set $Q_1 = |w\rangle\langle w|$; then

$$tr(Q_1 G) = \sum_j \lambda_j^2 |\langle w, \phi_j\rangle|^2. \qquad (3.40)$$

Since $\sum_j |\langle w, \phi_j\rangle|^2 = \|w\|^2 = 1$, the r.h.s. of (3.40) is a convex combination of λ_j^2's; therefore,

$$\sum_j \lambda_j^2 |\langle w, \phi_j\rangle|^2 \leq \lambda_1^2$$

and equality holds if and only if $|\langle w, \phi_1\rangle| = 1$, and $\langle w, \phi_j\rangle = 0$, for $j > 1$. $\qquad \square$

Lemma 3.9 can be applied inductively which yields the best rank-1 approximation at each iteration. In fact, the result holds more generally; see Theorem 3.11 below.

Theorem 3.11. *Let* $T : \mathcal{H} \to \mathcal{H}$ *be a compact operator, and let*

$$TT^* = \sum \lambda_j^2 \, |\phi_j \rangle\langle \phi_j|$$

with $\lambda_1^2 \geq \lambda_2^2 \geq \cdots$, $\lambda_j \to 0$. *Then*

$$\sum_{j=1}^{n} |\phi_j \rangle\langle \phi_j| = argmax \left\{ \|Q_n T\|_{HS}^2 : \|Q_n\|_{HS}^2 = n \right\};$$

where Q_n *runs over all s.a. projections of rank* n. *(Recall that* $\|Q_n T\|_{HS}^2 = tr(Q_n G)$, *where* $G := TT^*$. *See section* 2.5.)

Proof. Let $Q_n = \sum_{j=1}^{n} P_j$, where each P_j is rank-1, and $P_i P_j = 0$, for $i \neq j$. Then,

$$\|Q_n T\|_{HS}^2 = \sum_{j=1}^{n} \|P_j T\|_{HS}^2, \quad \|P_j\|_{HS}^2 = 1.$$

Let $\mathcal{L}$ be the corresponding Lagrangian, i.e.,

$$\mathcal{L} = \sum_{j=1}^{n} \|P_j T\|_{HS}^2 - \sum_{j=1}^{n} \mu_j \left(\|P_j\|_{HS}^2 - 1 \right)$$

$$= \sum_{j=1}^{n} \langle P_j, TT^* P_j \rangle_{HS} - \sum_{j=1}^{n} \mu_j \left(\langle P_j, P_j \rangle_{HS} - 1 \right).$$

It follows that

$$\frac{\partial \mathcal{L}}{\partial P_k} = 2 \left(TT^* P_k - \mu_k P_k \right) = 0$$

$$\Updownarrow$$

$$TT^* P_k = \mu_k P_k, \ \mu_k = \lambda_k^2.$$

Hence P_k is a spectral projection of TT^*.

The conclusion of the theorem follows from this. $\qquad\square$

3.4. **The Dual Problem.** Fix a data set $X = (x_j)_{j=1}^n$, $x_j \in \mathbb{C}^m$. Let $\Phi : X \to \mathcal{H}(K)$ be the feature map in (3.35), i.e.,

$$\Phi(X) = \begin{bmatrix} \Phi(x_1) & \cdots & \Phi(x_n) \end{bmatrix}. \tag{3.41}$$

Let L, L^* be the analysis and synthesis operators from (3.36)-(3.37), and

$$L^* L : \mathcal{H}(K) \longrightarrow \mathcal{H}(K), \quad L^* L f = \sum_{j=1}^{n} \langle \Phi(x_j), f \rangle \, \Phi(x_j) \tag{3.42}$$

be the frame operator in (3.38). In particular,

$$L^* L = \Phi(X) \Phi(X)^*.$$

In view of Theorems 2.12 and 3.11, the KL basis for $L^* L$ contains the principal directions carrying the greatest variance in $\Phi(X)$. In applications, it is more convenient to first find the KL basis of LL^* instead, where

$$LL^* = \Phi(X)^* \Phi(X) = (K(x_i, x_j))_{ij=1}^{n}, \tag{3.43}$$

as an $n \times n$ matrix in $\mathbb{C}^n$; see (3.2). (By general theory, if $A : \mathcal{H} \to \mathcal{H}$ is a linear operator in a Hilbert space with dense domain, then $\sigma(A^* A) \setminus \{0\} = \sigma(AA^*) \setminus \{0\}$.)

Proposition 3.12. *Set $A : \mathbb{C}^n \to \mathcal{H}(K)$ by*

$$A\delta_j = \Phi(x_j), \tag{3.44}$$

and extend linearly, where $(\delta_j)_{j=1}^n$ denotes the standard basis in $\mathbb{C}^n$. Then the adjoint operator $A^* : \mathcal{H}(K) \to \mathbb{C}^n$ is

$$A^*h = (\langle \Phi(x_j), h \rangle)_{j=1}^n \in \mathbb{C}^n. \tag{3.45}$$

That is, $A = L^*$ and $A^* = L$.

Proof. Let $v \in \mathbb{C}^n$, and $h \in \mathcal{H}$, then

$$\langle Av, h \rangle_{\mathcal{H}(K)} = \left\langle \sum_{j=1}^n v_j \Phi(x_j), h \right\rangle_{\mathcal{H}(K)}$$

$$= \sum_{j=1}^n \overline{v}_j \langle \Phi(x_j), h \rangle_{\mathcal{H}(K)} = \left\langle v, \left(\langle \Phi(x_j), h \rangle_{\mathcal{H}(K)} \right) \right\rangle_{\mathbb{C}^n},$$

and the assertions follows. $\square$

Hence $LL^* : \mathbb{C}^n \to \mathbb{C}^n$ is the Gramian matrix in $\mathbb{C}^n$ given by

$$
\begin{aligned}
LL^* &= \Phi(X)^* \Phi(X) \\
&= \left(\langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}(K)} \right)_{i,j=1}^n \\
&\underset{\text{by (3.2)}}{=} (K(x_i, x_j))_{i,j=1}^n;
\end{aligned} \tag{3.46}
$$

see (3.43).

By the singular value decomposition, $L^* = WDU^*$, so that

$$LL^* = UD^2U^*, \tag{3.47}$$

$$L^*L = WD^2W^*, \tag{3.48}$$

where $D = diag(\lambda_j)$ consists of the non-negative eigenvalues of $\sqrt{LL^*}$. Therefore,

$$L^* = \sum \lambda_j |w_j\rangle\langle u_j|.$$

Note that $W = (w_j)$ is the KL basis that diagonalizes L^*L as in (3.42), i.e.,

$$L^*L = \sum_{j=1}^n \lambda_j^2 |w_j\rangle\langle w_j|. \tag{3.49}$$

It also follows from (3.47)–(3.48), that

$$W = L^*UD^{-1}. \tag{3.50}$$

Remark 3.13. In the above discussion, $\Phi(X)$ may be centered by removing its mean. Specifically, let

$$J = 1 - \frac{1}{n} |\mathbb{1}\rangle\langle\mathbb{1}|$$

be the projection onto $span\{\mathbb{1}\}^\perp$, where $\mathbb{1}$ denotes the constant vector $\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}$. Then

$$\widetilde{\Phi}(X) := \Phi(X) - \frac{1}{n} \sum_{j=1}^n \Phi(x_j) = \Phi(X)J, \tag{3.51}$$

and so

$$LL^* = \widetilde{\Phi}(X)^* \widetilde{\Phi}(X) = J\Phi(X)^* \Phi(X) J. \tag{3.52}$$

The effect of J in (3.52) is to exclude the eigenspace of the Gramian $(K\left(x_i, x_j\right))_{i,j=1}^n$ spanned by the constant eigenvector.

In what follows, we shall always assume $\Phi\left(X\right)$ is centered as in (3.51)-(3.52).

3.5. **Feature Selection.** Feature selection, also called variable selection, or attribute selection, is a procedure for automatic selection of those attributes in data sets which are most relevant to particular predictive modeling problems. Which features should one use in designs of predictive models? This is a difficult question that requires detailed knowledge of the problem at hand. The aim is algorithmic designs which automatically select those features from prescribed data, which are most useful, or most relevant, for the particular problem. The process is called feature selection. A central premise of feature selection is that the input data will contain features that are either redundant or irrelevant, and can therefore be removed. The use of sample correlations in the process is based in turn on the following principle: A particular relevant feature might be redundant, in the presence of some other relevant feature, with which it is strongly correlated.

Our present purpose is not a systematic treatment of feature selection, but merely to identify how our present tools suggest recursive algorithms in the general area. With this in mind we now consider the following setup:

Let $x \in \mathbb{C}^m$ be a test example. The image $\Phi\left(x\right)$ under the feature map can be projected onto the principal directions in $\mathcal{H}\left(K\right)$, via

$$\Phi\left(x\right) \longmapsto WW^*\Phi\left(x\right).$$

The mapping $x \mapsto WW^*\Phi\left(x\right)$ is in general nonlinear. See Examples 3.15 and 3.16 below.

Corollary 3.14. *Let $L^* = WDU^*$ be as above, assuming D is full rank. For all $x \in \mathbb{C}^m$, the coefficients of the projection $WW^*\Phi\left(x\right)$ are*

$$W^*\Phi\left(x\right) = \begin{bmatrix} \lambda_1^{-1} \sum_{j=1}^n \overline{u_{j1}} K\left(x_j, x\right) \\ \vdots \\ \lambda_n^{-1} \sum_{j=1}^n \overline{u_{jn}} K\left(x_j, x\right) \end{bmatrix}.$$

Proof. By (3.50), $W^* = D^{-1}U^*L$, so that

$$
\begin{aligned}
W^*\Phi\left(x\right) \quad &= \quad D^{-1}U^*L\Phi\left(x\right) \\
&\underset{(3.45)}{=} \quad D^{-1}U^* \begin{bmatrix} \langle \Phi\left(x_1\right), \Phi\left(x\right) \rangle_{\mathcal{H}(K)} \\ \vdots \\ \langle \Phi\left(x_n\right), \Phi\left(x\right) \rangle_{\mathcal{H}(K)} \end{bmatrix} \\
&\underset{(3.2)}{=} \quad D^{-1}U^* \begin{bmatrix} K\left(x_1, x\right) \\ \vdots \\ K\left(x_n, x\right) \end{bmatrix} \\
&= \quad \begin{bmatrix} \lambda_1^{-1} \sum_{j=1}^n \overline{u_{j1}} K\left(x_j, x\right) \\ \vdots \\ \lambda_n^{-1} \sum_{j=1}^n \overline{u_{jn}} K\left(x_j, x\right) \end{bmatrix}.
\end{aligned}
$$

$\square$

Example 3.15 (Spectral Clustering, see Figure 3.1)**.** This is included as an instructive example. The data set X has two classes: Class "0" consists of 867 points uniformly distributed in $\{x \in \mathbb{R}^3 \, ; \, 0.6 < \|x\|^2 < 1\}$; class "1" consists of 126 points in the open ball $\{x \in \mathbb{R}^3 \, ; \, \|x\|^2 < 0.2\}$. Hence X has dimension 3×993, and each column x_j of X corresponds to a sample point.

We choose the Gaussian kernel $K(x, y) = e^{-\|x-y\|^2/\sigma}$ with $\sigma = 0.05$, so that x_j is embedded into the associated RKHS by

$$\mathbb{R}^3 \ni x_j \longmapsto \Phi(x_j) = e^{-\|\cdot - x_j\|^2/\sigma} \in \mathcal{H}(K).$$

By projecting $\Phi(x_j)$ onto the first two principal components as in Corollary 3.14, each sample point $x_j \in \mathbb{R}^3$ has a 2D representation via the mapping

$$\mathbb{R}^3 \ni x_j \longmapsto \Phi(x_j) \longmapsto \begin{bmatrix} \lambda_1^{-1} \sum_{k=1}^n \overline{u_{k1}} K(x_k, x_j) \\ \lambda_2^{-1} \sum_{k=1}^n \overline{u_{kn}} K(x_k, x_j) \end{bmatrix} \in \mathbb{R}^2.$$

As shown in Figure 3.1(B), the two clusters are linearly separable in $\mathcal{H}(K)$.
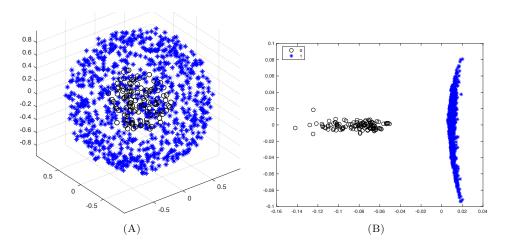

(A)   (B)

FIGURE 3.1. Spectral clustering via Gaussian kernel. (A) The two classes "$*$", and "o" are not separable by standard PCA. (B) KPCA with Gaussian kernel $K(x, y) = e^{-\|x-y\|^2/\sigma}$, $\sigma = 0.05$. Project $\Phi(X)$ onto the first two principal components, then the resulting 2D representation is separable by a linear decision boundary in $\mathcal{H}(K)$.

Example 3.16 (Dimension Reduction, see Figure 3.2)**.** Let X be a collection of 100 grayscale 256×256 images of an ellipse, rotated successively by $\pi/100$. Figure 3.2(A) shows 6 sample images corresponding to different rotation angles. The images are unrolled as column vectors, thus X has dimension 65536×100.

This data set may be viewed as 1D submanifold embedded in $\mathbb{R}^{65536}$, i.e., it has only one degree of freedom, the rotation angle. For dimension reduction, KPCA will ideally extract this information, and each image is then represented by a single projection coefficient. We choose the Gaussian kernel with $\sigma = 300$. In Figure 3.2(B), there are 4 subplots consisting of the projections onto the first, second,

third, and fourth principal directions. The rotation angle is encoded in e.g. PC 1. As a consequence, the dimension of the data set is reduced from 65536×100 to 1×100.

In particular, the projection coefficients onto the kth principal direction (see Figure 3.2(B), PC1 – PC4) are proportional to the kth eigenvector of the centered Gramian $G := JK(x_i, x_j) J$; see (3.52) and Corollary 3.14.

3.6. **Dynamic PCA.** Theorem 3.11 states that when a system of PCA eigenvalues $\{\lambda_i\}$ is given, then we have a algorithmic solution to the corresponding optimization question (see (3.33)-(3.34)). We now turn to a formula for generating PCA features as a limit of a certain iteration of operators. The family of operators T discussed below is a generalization of the operators from section 3.3 above.

Proposition 3.17. *Suppose $T : \mathcal{H} \to \mathcal{H}$ is compact, and T^*T has simple spectrum, i.e.,*

$$T^*T = \sum_{j=1}^{\infty} \lambda_j^2 |u_j\rangle\langle u_j|$$

with $\lambda_1^2 > \lambda_2^2 > \cdots > 0$; $\lambda_j^2 \to 0$. Then,

$$\lim_{n \to \infty} \frac{\left\| T^{n+1} x \right\|^2}{\left\| T^n x \right\|^2} = \lambda_1^2, \quad \forall x \notin ker\left(T^*T - \lambda_1^2\right). \tag{3.53}$$

Moreover, given λ_1, then

$$\lim_{n \to \infty} \lambda_1^{-2n} \left(T^*T\right)^n = |u_1\rangle\langle u_1|, \tag{3.54}$$

where convergence in (3.54) is w.r.t. the norm topology of $B(\mathcal{H})$.

Proof. Note that

$$
\begin{aligned}
\lim_{n \to \infty} \frac{\left\| T^{n+1} x \right\|^2}{\left\| T^n x \right\|^2} &= \lim_{n \to \infty} \frac{\left\langle x, (T^*T)^{n+1} x \right\rangle}{\left\langle x, (T^*T)^n x \right\rangle} \\
&= \lim_{n \to \infty} \frac{\sum \lambda_j^{2(n+1)} |\langle u_j, x \rangle|^2}{\sum \lambda_j^{2n} |\langle u_j, x \rangle|^2} \\
&= \lim_{n \to \infty} \lambda_1^2 \frac{|\langle u_1, x \rangle|^2 + \sum_{j \geq 2} (\lambda_j/\lambda_1)^{2(n+1)} |\langle u_j, x \rangle|^2}{|\langle u_1, x \rangle|^2 + \sum_{j \geq 2} (\lambda_j/\lambda_1)^{2n} |\langle u_j, x \rangle|^2} \\
&= \lambda_1^2.
\end{aligned}
$$

Now, given λ_1, we have

$$
\begin{aligned}
\left\| \lambda_1^{-2n} T^n - |u_1\rangle\langle u_1| \right\| &= \left\| \sum_{j \geq 2} (\lambda_j/\lambda_1)^{2n} |u_j\rangle\langle u_j| \right\| \\
&\leq \sum_{j \geq 2} (\lambda_j/\lambda_1)^{2n} \xrightarrow[n \to \infty]{} 0.
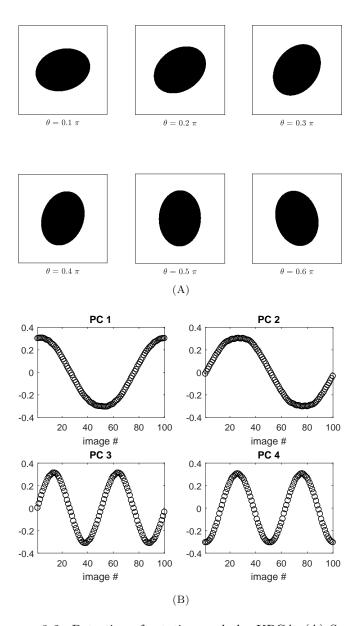\end{aligned}
$$

$\square$

FIGURE 3.2. Detection of rotation angle by KPCA. (A) Sample points from the data set X of 100 grayscale images of an ellipse, rotated successively by $\pi/100$, each has resolution 256×256. (B) Apply kernel PCA with Gaussian kernel $K(x,y) = e^{-\|x-y\|^2/\sigma}$, $\sigma = 300$, then project $\Phi(X)$ onto the first, second, third, and forth principal directions in $\mathcal{H}(K)$. The rotation angle is captured in e.g. PC 1.

Corollary 3.18. *The system of PCA eigenvalues $\{\lambda_j\}$ can be obtained inductively as follows: Set $Q_k = \sum_{j=1}^{k} |u_j\rangle\langle u_j|$, then*

$$\lambda_{k+1}^2 = \lim_{n \to \infty} \frac{\left\|\left(TQ_k^\perp\right)^{n+1} x\right\|^2}{\left\|\left(TQ_k^\perp\right)^n x\right\|}, \quad \forall x \notin ker\left(T^*T - \lambda_k^2\right)$$

and

$$|u_{k+1}\rangle\langle u_{k+1}| = \lim_{n \to \infty} \lambda_k^{-2n} \left(Q_k^\perp T^* T Q_k^\perp\right)^n.$$

Proof. One checks that

$$TQ_k^\perp = T\left(1 - Q_k\right) = \sum_{j=k+1}^{\infty} \lambda_j^2 |u_j\rangle\langle u_j|,$$

$$\left(TQ_k^\perp\right)^* \left(TQ_k^\perp\right) = Q_k^\perp T^* T Q_k^\perp;$$

and so the assertion follows from Proposition 3.17. $\qquad\square$

Example 3.19. Consider the covariance function of standard Brownian motion B_t, $t \in [0, \infty)$, i.e., a Gaussian process $\{B_t\}$ with mean zero and covariance function

$$\mathbb{E}\left(B_s B_t\right) = s \wedge t = \min\left(s, t\right). \tag{3.55}$$

Let $F_N = \{x_1, x_2, \ldots, x_N\}$ be a finite subsets of V, such that

$$0 < x_1 < x_2 < \cdots < x_N;$$

and let

$$K_N = \begin{bmatrix} x_1 & x_1 & x_1 & \cdots & x_1 \\ x_1 & x_2 & x_2 & \cdots & x_2 \\ x_1 & x_2 & x_3 & \cdots & x_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1 & x_2 & x_3 & \cdots & x_N \end{bmatrix} = \left(x_i \wedge x_j\right)_{i,j=1}^{N}. \tag{3.56}$$

Lemma 3.20. *Let K_N be as in (3.56). Then*

(i) *The determinant of K_N is given by*

$$\det\left(K_N\right) = x_1 \left(x_2 - x_1\right)\left(x_3 - x_2\right) \cdots \left(x_N - x_{N-1}\right). \tag{3.57}$$

(ii) *K_N assumes the LU decomposition*

$$K_N = A_N A_N^*, \tag{3.58}$$

where

$$A_n = \begin{bmatrix} \sqrt{x_1} & 0 & 0 & \cdots & 0 \\ \sqrt{x_1} & \sqrt{x_2 - x_1} & 0 & \cdots & \vdots \\ \sqrt{x_1} & \sqrt{x_2 - x_1} & \sqrt{x_3 - x_2} & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \sqrt{x_1} & \sqrt{x_2 - x_1} & \sqrt{x_3 - x_2} & \cdots & \sqrt{x_N - x_{N-1}} \end{bmatrix}. \tag{3.59}$$

Proof. For details, see e.g., [JT15]. $\qquad\square$

Example 3.21. Fix N, then the top eigenvalue of $K := K_N$ can be extracted by the method from Proposition 3.17. For instance, if $N = 3$ and $F_3 = \{1, 2, 3\}$, we have

$$K = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix}.$$

Let $e_1 = [1, 0, 0]$, then

$$\hat{\lambda}_1 = \frac{\left\| K^3 e_1 \right\|}{\left\| K^2 e_1 \right\|} \approx 5.0455, \quad \hat{v}_1 = \frac{\lambda_1^{-5} K^5 e_1}{\left\| \lambda_1^{-5} K^5 e_1 \right\|} \approx \begin{pmatrix} 0.3284 & 0.5913 & 0.7366 \end{pmatrix}^T.$$

Standard numerical algorithm returns

$$\lambda_1 \approx 5.0489, \quad v_1 \approx \begin{pmatrix} 0.3280 & 0.5910 & 0.7370 \end{pmatrix}^T.$$

4. Reproducing kernels, transforms, and Gaussian processes

In this section we discuss two of the positive definite kernels K used in section 3 above; see especially Figures 3.1 and 3.2, and Example 3.19.

We show that each kernel K is associated with a certain transform for its reproducing kernel Hilbert space (RKHS) $\mathcal{H}(K)$. The transform is studied in detail; – it may be viewed as an infinite-dimensional Fourier transform, see Definition 4.3. In detail, this transform $\mathcal{T}$ is defined on an L^2 path-space $L^2(\Omega, \mathbb{P})$ of Brownian motion; and $\mathcal{T}$ is shown to be an isometric isomorphism of $L^2(\Omega, \mathbb{P})$ onto the RKHS $\mathcal{H}(K)$; see Corollary 4.5. For earlier results dealing with the use of RKHSs in PCA, and related areas, see e.g., [BTA04, AW12, MRW$^+$18].

Consider the following positive definite (p.d.) kernel on $\mathbb{R} \times \mathbb{R}$;

$$K(s, t) := e^{-\frac{1}{2}|s-t|}, \ s, t \in \mathbb{R}. \tag{4.1}$$

In order to understand its PCA properties, we consider the top part of the spectrum in sampled versions of (4.1). We show below that K is the covariance kernel of the complex process $\left\{ e^{iX_t} \right\}_{t \in \mathbb{R}}$ where $\{X_t\}_{t \in \mathbb{R}}$ is the standard Gaussian process.

Let X_t be the standard Brownian motion indexed by $t \in \mathbb{R}$; i.e., X_t is realized on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, such that, for all $s, t \in \mathbb{R}$,

$$\mathbb{E}(X_s X_t) = \begin{cases} |s| \wedge |t| & \text{if } s \text{ and } t \text{ have the same sign} \\ 0 & \text{if } st \leq 0. \end{cases} \tag{4.2}$$

Here $\mathbb{E}(\cdot)$ denotes the expectation,

$$\mathbb{E}(\cdots) = \int_\Omega (\cdots) \, d\mathbb{P}. \tag{4.3}$$

Remark 4.1. The process $\{X_t\}_{t \in \mathbb{R}}$ can be realized in many different but equivalent ways.

Note that

$$\mathbb{E}\left(|X_s - X_t|^2 \right) = |s - t|, \tag{4.4}$$

so the process X_t has stationary and independent increments. In particular, if $s, t > 0$, then $|s - t| = s + t - 2 \, s \wedge t$, and

$$s \wedge t = \frac{s + t - |s - t|}{2}. \tag{4.5}$$

Proposition 4.2. *The kernel K from (4.1) is positive definite.*

Proof. Let $\{X_t\}_{t \in \mathbb{R}}$ be the standard Brownian motion and let e^{iX_t} be the corresponding complex process, then by direct calculation,

$$e^{-\frac{1}{2}|t|} = \mathbb{E}\left(e^{iX_t}\right), \ \forall t \in \mathbb{R}; \tag{4.6}$$

and

$$\begin{aligned} e^{-\frac{1}{2}|s-t|} &= \mathbb{E}\left(e^{iX_s}e^{-iX_t}\right) \\ &= \left\langle e^{iX_s}, e^{iX_t}\right\rangle_{L^2(\mathbb{P})}, \ \forall s, t \in \mathbb{R}. \end{aligned} \tag{4.7}$$

The derivation of (4.6)-(4.7) is based on power series expansion of e^{iX_t}, and the fact that

$$\mathbb{E}\left(X_t^{2n}\right) = (2n-1)!! \, |t|^n, \ \text{and} \tag{4.8}$$

$$\mathbb{E}\left(|X_t - X_s|^{2n}\right) = (2n-1)!! \, |t-s|^n, \tag{4.9}$$

where

$$(2n-1)!! = (2n-1)(2n-3)\cdots 5 \cdot 3 = \frac{(2n)!}{2^n n!}. \tag{4.10}$$

Now the p.d. property of K follows from (4.7), since the RHS of (4.7) is p.d. In details: for all $(c_j)_{j=1}^N$, $c_j \in \mathbb{R}$:

$$\begin{aligned} \sum_j \sum_k c_j c_k \, e^{-\frac{1}{2}|t_j - t_k|} &= \sum_j \sum_k c_j c_k \mathbb{E}\left(e^{iX_{t_j}}e^{-iX_{t_k}}\right) \\ &= \mathbb{E}\left(\left|\sum_j c_j e^{iX_{t_j}}\right|^2\right) \geq 0. \end{aligned}$$

$\square$

Generalized Fourier transform.

Definition 4.3. Let $\mathcal{H}(K)$ be the RKHS from the kernel K in (4.1); and define the following transform $\mathcal{T} : L^2(\mathbb{P}) \longrightarrow \mathcal{H}(K)$,

$$\begin{aligned} \mathcal{T}(F)(t) &:= \mathbb{E}\left(e^{-iX_t}F\right) \\ &= \int_\Omega e^{-iX_t(\omega)}F(\omega)\, d\mathbb{P}(\omega) \end{aligned} \tag{4.11}$$

for all $F \in L^2(\mathbb{P})$.

It is known that the standard Brownian motion, indexed by $\mathbb{R}$, has a continuous realization (see e.g., [Hid80].) Hence the transform $\mathcal{T}$ defined by (4.11) maps $L^2(\Omega, \mathbb{P})$ into the bounded continuous functions on $\mathbb{R}$. Corollary 4.5, below, is the stronger assertion that $\mathcal{T}$ maps $L^2(\Omega, \mathbb{P})$ isometrically onto the RKHS $\mathcal{H}(K)$ where K is the kernel in (4.1).

Set

$$K_t(\cdot) = e^{-\frac{1}{2}|t-\cdot|} \in \mathcal{H}(K), \tag{4.12}$$

then by the reproducing property in $\mathcal{H}(K)$, we have

$$\langle K_t, \psi \rangle_{\mathcal{H}(K)} = \psi(t), \ \forall \psi \in \mathcal{H}(K). \tag{4.13}$$

Lemma 4.4. *Let $\mathcal{T}$ be the generalized Fourier transform in (4.11), and $\mathcal{T}^*$ be the adjoint operator; see the diagram below.*

$$L^2(\mathbb{P}) \overset{\mathcal{T}}{\underset{\mathcal{T}^*}{\rightleftarrows}} \mathcal{H}(K) \qquad (4.14)$$

Then, we have

$$J\left(e^{iX_t}\right) = K_t, \ and \qquad (4.15)$$

$$J^*(K_t) = e^{iX_t}. \qquad (4.16)$$

Proof. Recall the definition $\mathcal{T}(F)(s) := \mathbb{E}\left(e^{-iX_s}F\right)$.

Proof of (4.15). Setting $F = e^{iX_t}$, then

$$\mathcal{T}\left(e^{iX_t}\right)(s) = \mathbb{E}\left(e^{-iX_s}e^{iX_t}\right) = \mathbb{E}\left(e^{i(X_t - X_s)}\right)$$

$$= e^{-\frac{1}{2}|t-s|} = K_t(s).$$

Proof of (4.16). Let $F \in L^2(\mathbb{P})$, then

$$\langle \mathcal{T}^*(K_t), F \rangle_{L^2(\mathbb{P})} = \langle K_t, \mathcal{T}(F) \rangle_{\mathcal{H}(K)} = \mathcal{T}(F)(t)$$

$$= \mathbb{E}\left(e^{-iX_t}F\right) = \left\langle e^{iX_t}, F \right\rangle_{L^2(\mathbb{P})}.$$

$\square$

Corollary 4.5. *The generalized Fourier transform in (4.14) is an isometric isomorphism from $L^2(\mathbb{P})$ onto $\mathcal{H}(K)$, with $\mathcal{T}\left(e^{iX_t}\right) = K_t$, see (4.15).*

Proof. This is a direct application of (4.15)-(4.16). Also note that $span\left\{e^{iX_t}\right\}$ is dense in $L^2(\mathbb{P})$, and $span\left\{K_t\right\}$ is dense in $\mathcal{H}(K)$. $\square$

Conclusion. $\mathcal{H}(K)$ is naturally isometrically isomorphic to $L^2(\mathbb{P})$.

Remark 4.6. To understand this isometric isomorphism $L^2(\mathbb{P}) \overset{\simeq}{\longrightarrow} \mathcal{H}(K)$, we must treat $L^2(\mathbb{P})$ as a *complex* Hilbert space, while $\mathcal{H}(K)$ is defined as a real Hilbert space; i.e., the generating functions $e^{iX_t} \in L^2(\mathbb{P})$ are complex, where the inner product in $L^2(\mathbb{P})$ is $\langle u, v \rangle_{L^2(\mathbb{P})} = \int_\Omega u\overline{v}d\mathbb{P}$; but the functions K_t, $t \in \mathbb{R}$, in $\mathcal{H}(K)$ are real valued.

Remark 4.7. Assume the normalization $X_0 = 0$, and $0 < s < t$. The two processes X_{t-s}, and $X_t - X_s$ are different, but they have the same distribution $N(0, t-s)$. Indeed, we have

$$\mathbb{E}\left(e^{iX_t}e^{-iX_s}\right) \quad = \quad \mathbb{E}\left(e^{i(X_t - X_s)}\right)$$

$$= \quad \sum_{n=0}^{\infty} \frac{i^n}{n!}\mathbb{E}\left((X_t - X_s)^n\right)$$

$$\underset{\text{by (4.8)}}{=} \quad \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}\mathbb{E}\left((X_t - X_s)^{2n}\right)$$

$$\underset{\text{by (4.9)}}{=} \quad \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}(2n-1)!!\,(t-s)^n$$

$$\underset{\text{by } (4.10)}{=} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n n!} (t-s)^n$$

$$= \sum_{n=0}^{\infty} \frac{1}{n!} \left(-\frac{1}{2} (t-s) \right)^n$$

$$= e^{-\frac{1}{2}(t-s)} = \mathbb{E}\left(e^{iX_{t-s}} \right).$$

Proposition 4.8. *For the Gaussian kernel* $K(x,y) = e^{\frac{1}{2t}(x-y)^2}$, *we have*

$$e^{-x^2/2t} = \mathbb{E}\left(e^{ixX_{1/t}} \right), \text{ and} \tag{4.17}$$

$$e^{-(x-y)^2/2t} = \mathbb{E}\left(e^{ixX_{1/t}} e^{-iyX_{1/t}} \right) = \left\langle e^{ixX_{1/t}}, e^{iyX_{1/t}} \right\rangle_{L^2(\mathbb{P})}. \tag{4.18}$$

Proof. A direct calculation yields

$$\mathbb{E}\left(e^{ixX_{1/t}} \right) = \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} \mathbb{E}\left(X_{1/t}^n \right)$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} (2n-1)!! \frac{1}{t^n}$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n n!} \frac{x^{2n}}{t^n} = \sum_{n=0}^{\infty} \frac{1}{n!} \left(-\frac{x^2}{2t} \right)^n = e^{-x^2/2t},$$

which is (4.17); and (4.18) follows from this. $\qquad\square$

Lemma 4.9. *Assume* $0 < s < t$. *If* $F \in L^2(\mathbb{P}) \cap \mathscr{B}_s$, *where* $\mathscr{B}_s = \sigma$-*algebra generated by* $\{X_u \, ; \, u \le s\}$, *then*

$$\mathcal{T}(F)(t) = e^{-\frac{t-s}{2}} \mathcal{T}(F)(s). \tag{4.19}$$

Proof.

$$\text{LHS}_{(4.19)} = \mathbb{E}\left(e^{-iX_t} F \right)$$

$$= \mathbb{E}\left(e^{-i(X_t - X_s)} e^{-iX_s} F \right)$$

$$\underset{\text{of increament}}{\overset{\text{the independence}}{=}} \mathbb{E}\left(e^{-i(X_t - X_s)} \right) \mathbb{E}\left(e^{-iX_s} F \right)$$

$$= e^{-\frac{t-s}{2}} \mathcal{T}(F)(s) = \text{RHS}_{(4.19)}.$$

$$\square$$

Proposition 4.10. *Let* $0 < s < t$, *and let* $H_n(\cdot)$, $n \in \mathbb{N}_0$, *be the Hermite polynomials; then*

$$\mathcal{T}(X_s^n)(t) = i^n e^{-\frac{t}{2}} s^{\frac{n}{2}} H_n\left(\sqrt{s} \right). \tag{4.20}$$

Proof. By Lemma 4.9, we have

$$\mathcal{T}(X_s^n)(t) = e^{-\frac{t-s}{2}} \mathcal{T}(X_s^n)(s)$$

$$= e^{-\frac{t-s}{2}} \mathbb{E}\left(e^{-iX_s} X_s^n \right)$$

$$= e^{-\frac{t-s}{2}} i^n \left(\frac{d}{d\lambda} \right)^n \Big|_{\lambda=1} \mathbb{E}\left(e^{-i\lambda X_s} \right)$$

$$= e^{-\frac{t-s}{2}} i^n \left(\frac{d}{d\lambda} \right)^n \Big|_{\lambda=1} \left(e^{-\frac{\lambda^2 s}{2}} \right)$$

$$= i^n e^{-\frac{t-s}{2}} e^{-\frac{s}{2}} s^{\frac{n}{2}} H_n \left(\sqrt{s} \right)$$
$$= i^n e^{-\frac{t}{2}} s^{\frac{n}{2}} H_n \left(\sqrt{s} \right)$$

which is the RHS in (4.20). In the calculation above, we have used the following version of the Hermite polynomials $H_n \left(\cdot \right)$, the probabilist's variant; defined by

$$\left(\frac{d}{d\xi} \right)^n e^{-\frac{\xi^2}{2}} = H_n \left(\xi \right) e^{-\frac{\xi^2}{2}}$$

with the substitution $\xi = \sqrt{s}\lambda$ for $s > 0$ fixed, and $\lambda \to 1 \Leftrightarrow \xi \to \sqrt{s}$. $\square$

References

[AFS18]   Carlos M. Alaíz, Michaël Fanuel, and Johan A. K. Suykens, *Convex formulation for kernel PCA and its use in semisupervised learning*, IEEE Trans. Neural Netw. Learn. Syst. **29** (2018), no. 8, 3863–3869. MR 3854652

[AJ12]   Daniel Alpay and Palle E. T. Jorgensen, *Stochastic processes induced by singular operators*, Numer. Funct. Anal. Optim. **33** (2012), no. 7-9, 708–735. MR 2966130

[AJ15]   Daniel Alpay and Palle Jorgensen, *Spectral theory for Gaussian processes: reproducing kernels, boundaries, and L^2-wavelet generators with fractional scales*, Numer. Funct. Anal. Optim. **36** (2015), no. 10, 1239–1285. MR 3402823

[AJL11]   Daniel Alpay, Palle Jorgensen, and David Levanony, *A class of Gaussian processes with fractional spectral measures*, J. Funct. Anal. **261** (2011), no. 2, 507–541. MR 2793121

[AJL17]   _____, *On the equivalence of probability spaces*, J. Theoret. Probab. **30** (2017), no. 3, 813–841. MR 3687240

[AJS14]   Daniel Alpay, Palle Jorgensen, and Guy Salomon, *On free stochastic processes and their derivatives*, Stochastic Process. Appl. **124** (2014), no. 10, 3392–3411. MR 3231624

[AK06]   William Arveson and Richard V. Kadison, *Diagonals of self-adjoint operators*, Operator theory, operator algebras, and applications, Contemp. Math., vol. 414, Amer. Math. Soc., Providence, RI, 2006, pp. 247–263. MR 2277215

[Arv07]   William Arveson, *Diagonals of normal operators with finite spectrum*, Proc. Natl. Acad. Sci. USA **104** (2007), no. 4, 1152–1158 (electronic). MR 2303566

[AW12]   Arash A. Amini and Martin J. Wainwright, *Sampled forms of functional PCA in reproducing kernel Hilbert spaces*, Ann. Statist. **40** (2012), no. 5, 2483–2510. MR 3097610

[BH19]   Peter Balazs and Helmut Harbrecht, *Frames for the Solution of Operator Equations in Hilbert Spaces with Fixed Dual Pairing*, Numer. Funct. Anal. Optim. **40** (2019), no. 1, 65–84. MR 3928472

[Bis06]   Christopher M. Bishop, *Pattern recognition and machine learning*, Information Science and Statistics, Springer, New York, 2006. MR 2247587

[Bis13]   _____, *Model-based machine learning*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **371** (2013), no. 1984, 20120222, 17. MR 3005666

[BJ02]   Ola Bratteli and Palle Jorgensen, *Wavelets through a looking glass*, Applied and Numerical Harmonic Analysis, Birkhäuser Boston, Inc., Boston, MA, 2002, The world of the spectrum. MR 1913212

[BN03]   Mikhail Belkin and Partha Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Computation **15** (2003), no. 6, 1373–1396.

[BTA04]   Alain Berlinet and Christine Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*, Kluwer Academic Publishers, Boston, MA, 2004, With a preface by Persi Diaconis. MR 2239907

[CCEL15]   Jameson Cahill, Peter G. Casazza, Martin Ehler, and Shidong Li, *Tight and random nonorthogonal fusion frames*, Trends in harmonic analysis and its applications, Contemp. Math., vol. 650, Amer. Math. Soc., Providence, RI, 2015, pp. 23–36. MR 3441732

[CCK13]   Jameson Cahill, Peter G. Casazza, and Gitta Kutyniok, *Operators and frames*, J. Operator Theory **70** (2013), no. 1, 145–164. MR 3085820

[CLS⁺19]   Kayla D. Coleman, Allison Lewis, Ralph C. Smith, Brian Williams, Max Morris, and Bassam Khuwaileh, *Gradient-free construction of active subspaces for dimension reduction in complex models with applications to neutronics*, SIAM/ASA J. Uncertain. Quantif. **7** (2019), no. 1, 117–142. MR 3900802

[CS02]   Felipe Cucker and Steve Smale, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc. (N.S.) **39** (2002), no. 1, 1–49. MR 1864085

[CWG19]   Jia Chen, Gang Wang, and Georgios B. Giannakis, *Nonlinear dimensionality reduction for discriminative analytics of multiple datasets*, IEEE Trans. Signal Process. **67** (2019), no. 3, 740–752. MR 3912285

[dES12]   Rafael do EspÃrito Santo, *Principal component analysis applied to digital image compression*, Einstein (Sao Paulo) **10** (2012), no. 2, 135–139.

[DF07]   Q. Du and J. E. Fowler, *Hyperspectral image compression using jpeg2000 and principal component analysis*, IEEE Geoscience and Remote Sensing Letters **4** (2007), no. 2, 201–205.

[Dir47]   P. A. M. Dirac, *The Principles of Quantum Mechanics*, Oxford, at the Clarendon Press, 1947, 3d ed. MR 0023198 (9,319d)

[DWGC18]   Jianghu J. Dong, Liangliang Wang, Jagbir Gill, and Jiguo Cao, *Functional principal component analysis of glomerular filtration rate curves after kidney transplant*, Stat. Methods Med. Res. **27** (2018), no. 12, 3785–3796. MR 3878657

[GGB18]   Alon Gonen and Ran Gilad-Bachrach, *Smooth sensitivity based approach for differentially private principal component analysis*, Algorithmic learning theory 2018, Proc. Mach. Learn. Res. (PMLR), vol. 83, Proceedings of Machine Learning Research PMLR, [place of publication not identified], 2018, p. 13. MR 3857315

[GJ06]   M. R. Gupta and N. P. Jacobson, *Wavelet principal component analysis and its application to hyperspectral images*, 2006 International Conference on Image Processing, Oct 2006, pp. 1585–1588.

[GK15]   Kamaljeet Kaur Gurpreet Kaur, *Image compression using dwt and principal component analysis*, IOSR Journal of Electrical and Electronics Engineering **10** (2015), no. 3, 53–56.

[GK19]   Constantin Grigo and Phaedon-Stelios Koutsourelakis, *Bayesian Model and Dimension Reduction for Uncertainty Propagation: Applications in Random Media*, SIAM/ASA J. Uncertain. Quantif. **7** (2019), no. 1, 292–323. MR 3922239

[Hid80]   Takeyuki Hida, *Brownian motion*, Applications of Mathematics, vol. 11, Springer-Verlag, New York-Berlin, 1980, Translated from the Japanese by the author and T. P. Speed. MR 562914

[HKLW07]   Deguang Han, Keri Kornelson, David Larson, and Eric Weber, *Frames for undergraduates*, Student Mathematical Library, vol. 40, American Mathematical Society, Providence, RI, 2007. MR 2367342

[HWW05]   Ryan Harkins, Eric Weber, and Andrew Westmeyer, *Encryption schemes using finite frames and Hadamard arrays*, Experiment. Math. **14** (2005), no. 4, 423–433. MR 2193805

[JHZW19]   J. Jin, T. Huang, J. L. Zheng, and P. H. Wen, *Dimension reduction analysis with mapping and direct integration algorithm*, Eng. Anal. Bound. Elem. **99** (2019), 122–130. MR 3883202

[JS07]   Palle E. T. Jorgensen and Myung-Sin Song, *Entropy encoding, Hilbert space, and Karhunen-Loève transforms*, J. Math. Phys. **48** (2007), no. 10, 103503.

[JT15]   Palle Jorgensen and Feng Tian, *Discrete reproducing kernel Hilbert spaces: sampling and distribution of Dirac-masses*, J. Mach. Learn. Res. **16** (2015), 3079–3114. MR 3450534

[JT16a]   ———, *Graph Laplacians and discrete reproducing kernel Hilbert spaces from restrictions*, Stoch. Anal. Appl. **34** (2016), no. 4, 722–747. MR 3507188

[JT16b]   ———, *Positive definite kernels and boundary spaces*, Adv. Oper. Theory **1** (2016), no. 1, 123–133. MR 3721329

[JT18a]   ———, *Metric duality between positive definite kernels and boundary processes*, Int. J. Appl. Comput. Math. **4** (2018), no. 1, Art. 3, 13. MR 3736758

[JT18b]   ———, *Realizations and factorizations of positive definite kernels*, Journal of Theoretical Probability (2018).

[KKS16]   K. J. Satao Khushboo Kumar Sahu, *Image compression methods using dimension reduction and classification through pca and lda: A review*, International Journal of Science and Research **5** (2016), no. 5, 2319–7064.

[LHN18]   Yeejin Lee, Keigo Hirakawa, and Truong Q. Nguyen, *Camera-aware multiresolution analysis for raw image sensor data compression*, IEEE Trans. Image Process. **27** (2018), no. 6, 2806–2817. MR 3780557

[LLL11]   Wen Li Liu, Shu Long Lü, and Fei Bao Liang, *Kernel density discriminant method based on geodesic distance*, J. Fuzhou Univ. Nat. Sci. Ed. **39** (2011), no. 6, 807–810, 818. MR 2933765

[LLY⁺19]   Xiao Lin, Ruosha Li, Fangrong Yan, Tao Lu, and Xuelin Huang, *Quantile residual lifetime regression with functional principal component analysis of longitudinal data for dynamic prediction*, Stat. Methods Med. Res. **28** (2019), no. 4, 1216–1229. MR 3934645

[Mar14]   S. Marsland, *Machine learning: An algorithmic perspective*, Chapman and Hall/CRC, Boca Raton, FL, 2014.

[MMP19]   Shahar Mendelson, Emanuel Milman, and Grigoris Paouris, *Generalized dual Sudakov minoration via dimension-reduction—a program*, Studia Math. **244** (2019), no. 2, 159–202. MR 3850675

[MP05]   M. Mudrova and A. Prochazka, *Principal component analysis in image processing*, Proceedings of In Technical Computing Conference (2005).

[MRW⁺18]   Horia Mania, Aaditya Ramdas, Martin J. Wainwright, Michael I. Jordan, and Benjamin Recht, *On kernel methods for covariates that are rankings*, Electron. J. Stat. **12** (2018), no. 2, 2537–2577. MR 3843387

[PS03]   Tomaso Poggio and Steve Smale, *The mathematics of learning: dealing with data*, Notices Amer. Math. Soc. **50** (2003), no. 5, 537–544. MR 1968413

[Raj18]   S. P. Raja, *Secured medical image compression using DES encryption technique in Bandelet multiscale transform*, Int. J. Wavelets Multiresolut. Inf. Process. **16** (2018), no. 4, 1850028, 33. MR 3820672

[RS99]   B. Ya. Ryabko and M. P. Sharova, *Fast encoding of low-entropy sources*, Problemy Peredachi Informatsii **35** (1999), no. 1, 49–60. MR 1720704

[SGS⁺19]   C. Soize, R. Ghanem, C. Safta, X. Huan, Z. P. Vane, J. Oefelein, G. Lacaze, H. N. Najm, Q. Tang, and X. Chen, *Entropy-based closure for probabilistic learning on manifolds*, J. Comput. Phys. **388** (2019), 518–533. MR 3934524

[Son08]   Myung-Sin Song, *Entropy encoding in wavelet image compression*, Representations, wavelets, and frames, Appl. Numer. Harmon. Anal., Birkhäuser Boston, Boston, MA, 2008, pp. 293–311. MR 2459323

[SY06]   Steve Smale and Yuan Yao, *Online learning algorithms*, Found. Comput. Math. **6** (2006), no. 2, 145–170. MR 2228737

[SZ07]   Steve Smale and Ding-Xuan Zhou, *Learning theory estimates via integral operators and their approximations*, Constr. Approx. **26** (2007), no. 2, 153–172. MR 2327597

[SZ09a]   ———, *Geometry on probability spaces*, Constr. Approx. **30** (2009), no. 3, 311–323. MR 2558684

[SZ09b]   ———, *Online learning with Markov sampling*, Anal. Appl. (Singap.) **7** (2009), no. 1, 87–113. MR 2488871

[TF19]   Claudio Turchetti and Laura Falaschetti, *A manifold learning approach to dimensionality reduction for modeling data*, Inform. Sci. **491** (2019), 16–29. MR 3935948

[THH19]   I-Ping Tu, Su-Yun Huang, and Dai-Ni Hsieh, *The generalized degrees of freedom of multilinear principal component analysis*, J. Multivariate Anal. **173** (2019), 26–37. MR 3913046

[VD16]   Hai X. Vo and Louis J. Durlofsky, *Regularized kernel PCA for the efficient parameterization of complex geological models*, J. Comput. Phys. **322** (2016), 859–881. MR 3534893

[VVQCR$^+$19] Rafael Vega Vega, Héctor Quintián, José Luís Calvo-Rolle, Álvaro Herrero, and Emilio Corchado, *Gaining deep knowledge of Android malware families through dimensionality reduction techniques*, Log. J. IGPL **27** (2019), no. 2, 160–176. MR 3935831

[Wat67]   Satosi Watanabe, *Karhunen-Loève expansion and factor analysis: Theoretical remarks and applications*, Trans. Fourth Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes (Prague, 1965), Academia, Prague, 1967, pp. 635–660. MR 0234768 (38 #3084)

[WGLP19]   Ning Wang, Shuangkui Ge, Baobin Li, and Lizhong Peng, *Multiple description image compression based on multiwavelets*, Int. J. Wavelets Multiresolut. Inf. Process. **17** (2019), no. 1, 1850063, 22. MR 3911884

[YLTL18]   Yuping Ying, Yanping Lian, Shaoqiang Tang, and Wing Kam Liu, *Enriched reproducing kernel particle method for fractional advection-diffusion equation*, Acta Mech. Sin. **34** (2018), no. 3, 515–527. MR 3803845

[ZBB04]   Laurent Zwald, Olivier Bousquet, and Gilles Blanchard, *Statistical properties of kernel principal component analysis*, Learning theory, Lecture Notes in Comput. Sci., vol. 3120, Springer, Berlin, 2004, pp. 594–608. MR 2177937

(Palle E.T. Jorgensen) Department of Mathematics, The University of Iowa, Iowa City, IA 52242-1419, U.S.A.
   E-mail address: `palle-jorgensen@uiowa.edu`

(Sooran Kang) College of General Education, Choongang University, Seoul, Korea
   E-mail address: `sooran09@cau.ac.kr`

(Myung-Sin Song) Department of Mathematics and Statistics, Southern Illinois University Edwardsville, Edwardsville, IL 62026, USA
   E-mail address: `msong@siue.edu`

(Feng Tian) Department of Mathematics, Hampton University, Hampton, VA 23668, U.S.A.
   E-mail address: `feng.tian@hamptonu.edu`