DATA DIMENSION REDUCTION AND KERNEL PRINCIPAL
COMPONENT ANALYSIS: INFINITE DIMENSIONS

PALLE E.T. JORGENSEN, SOORAN KANG, MYUNG-SIN SONG, AND JAMES 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 com-
ponents, which serve to minimize entropy and error; and hence to improve
digital representation of images, and hence of optimal storage, and transmis-
sion. 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

Data and digital image illustrations 2
1. Introduction 2
2. Karhunen-Loève transform or Principal Component Analysis 6
  2.1. The Algorithm for a Digital Image or Data Application 7
  2.2. Principal Component Analysis in a Digital Image 8
  2.3. Dimension Reduction and Principal Component Analysis 11
  2.4. KL Transform/PCA 13
  2.5. Principal Component Analysis and Maximal Variance 17
3. Kernel PCA 21
  3.1. Application to Optimization 25
  3.2. The Case of Gaussian Fields 28
  3.3. Optimization and Frames 29
  3.4. The Dual Problem 33
  3.5. Feature Selection 34
  3.6. Dynamic PCA 36
4. Reproducing kernels, transforms, and Gaussian processes 40
  4.1. Generalized Fourier transform 42
References 45

2000 Mathematics Subject Classification. 62H25, 34L16, 65K10, 65T60, 42C15, 47B06, 47B32,
65D15, 47A70, 37M25, 42C40.

Key words and phrases. dimension reduction, principal component analysis, eigenvalues, opti-
mization, Karhunen-Loève transform, optimal storage, transmission, wavelet image compression,
frames, entropy encoding, algorithms, Hilbert space.

arXiv:1906.06451v2  [math.FA]  26 Mar 2022
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 [WGY10, BM03]. The most prominent such a technique is known as maximum variance unfolding (MVU) [ACP13, SBXD06].

The aim of this paper is to set the stage for further developments of the linking of the mathematics of kernel-PCA (KPCA) with real-life non-linear data-sets, coming from a host of new developments in the current applied literature. A key focus for our work is a study of two complementary directions: on one side, the mathematics of kernel theory, and on the other, empirical (here, numerical experiments starting with databases from real-world situations). For the purpose, we develop (Sections 3 and 4) particular kernel tools with view to KPCA via statistical simulation, as well as corresponding analyses of kernel based choices of feature spaces and feature maps.
In recent years, the subject of kernel-principal component analysis (KPCA), and its applications, has been extensively studied, and made progress in diverse directions. There is a variety of themes covered in the literature dealing with applications of kernel-PCA tools is vast, each of some relevance to our present theme. A short user guide: the paper [GGB18] deals with sensitivity issues, [JHZW19] direct integration algorithms, [GJ06] image processing, [Bis06, Bis13, AFS18] pattern recognition and machine learning, [ZBB04] statistical properties, and finally [VD16] on selection of efficient parameterizations. In addition to earlier papers by the co-authors [Son08, JS07], we also include here a partial list of other relevant and current citations; see e.g., the papers mentioned above, [LHN18, Raj18, WGLP19, RS99, GGB18, DWGC18, THH19, LLV+19, MMP19, CLS+19, GK19].

In our present paper we stress, and motivate, the most general framework for choices of positive definite kernels, their corresponding RKHSs, and families of compatible measure spaces. These considerations encompass more general contexts for Principal Component Analysis, going beyond classical cases of kernel-based PCA in the literature. In particular, this very general and non-linear framework for PCA data-dimension reduction goes beyond existing, more classical, considerations admitting easy realizations in $\mathbb{R}^n$, classical kernels, and classical Fourier tools. Indeed, there are existing parallel considerations, in the literature involving only kernel analyses for such special cases.

However, existing non-linear realizations of big and un-structured data motivate choices from a much wider family of RKHSs, and measure spaces. This further entails infinite-dimensional stochastic analysis frameworks which do not admit realizations in $\mathbb{R}^n$, nor classical PDE/Fourier techniques. Examples of uses, in the literature, of general RKHSs in stochastic approaches include online learning with Markov sampling, data mining (e.g., in finance, intelligence, hidden data, tele-communication, energy, unstructured information integrated with traditional structured data), as well as a diverse host of different investigations, see e.g., [ACP13, AHR+20, AJS14, Bis06, BM03, BTA04, CS02, CWG19, JT16a, JT18a, JT18b, Mok07, SZ09a, SZ09b, ZBB04, JT17, Agg18, Hog20]. Moreover “big data” further dictate diverse, large, infinite-dimensional, and extremely non-linear datasets which do not lend themselves to feature analysis with choices of anyone of the simple kernels with easy realizations in $\mathbb{R}^n$. Examples of this includes many public available datasets on Global Health Observatory, U.S. Census Bureau, NIH Data Sharing Repositories, Kaggle, etc. Such cases of course dictate new and diverse choices from a much wider family of kernels and RKHSs. Encoding of meaningful programs for corresponding non-linear data-dimension reduction algorithms, while feasible, is a gigantic task, going far beyond the scope of our present paper. It is the subject of future projects involving the present authors, as well as others, see e.g., the papers cited above.

In addition, we call attention to the 2018 paper [YTLM+18] by Yamada et al. It is motivated by existing and new biological data. For the purpose, the authors introduce a novel Hilbert-Schmidt independence criterion. They apply it to specific high dimensional, and large-scale datasets, as arising in ultra-high dimensional biological data. The authors obtain classification into phenotypes, module expression in human prostate cancer patients, and detection of enzymes in protein structures. The 2019 paper [CLW+19] by Chen et al deals with a new class of kernel based nonlinear clustering algorithms. They study locality structures in
machine learning. Also motivated by specific large-scale datasets is the 2020 paper [UTK20]. This paper introduces new hyper-parameter initialization tools for kernel-based regression analysis. Yet different but related kernel based technology, and specific data sets, are further covered in the papers [FZ05], [WSS04], [MMR+01], and [DSR+21].

Motivated by related applications, we focus on non-traditional positive definite kernels and the associated RKHSs, in particular, when the family of feature spaces may be chosen in the form $L^2(\mu)$. This raises the question of which measures $\mu$ are right for a particular kernel $K$ and its RKHS.

In the use of probabilistic tools in big data, such as Monte Carlo simulation (see e.g., (1.1) below), a main issue is reduction of the required sample size; and still maintain (i) acceptable small noise in output, (ii) small error term estimates, as well as (iii) reduction of computation time. These issues are especially relevant for modern image processing. As a consequence, more traditional kernel tools must therefore be adapted. Both our present paper, and an earlier one [JT22], serve to do this by shifting the focus in the design of optimization algorithms to that of making choices of measurable partitions, as opposed to working directly with points.

For instance, when $K = K(\mu)$ is defined on a $\sigma$-finite measure space $(V, \mathcal{B}, \mu)$, where $K(\mu)(A, B) = \mu(A \cap B)$, for all $A, B$ in $\mathcal{B}_{\text{fin}} = \{ A \in \mathcal{B} : \mu(A) < \infty \}$, the associated RKHS $\mathcal{H}(K(\mu))$ is shown to be a Hilbert space of signed finite measures $F$ on $(V, \mathcal{B})$, such that $F \ll \mu$ and $\| F \|_{\mathcal{H}(K(\mu))} = \| dF/d\mu \|_{L^2(\mu)}$; see [JT22, Thms 3.2, 3.4]. Moreover, in this measurable setting, the classical Karhunen-Loève decomposition takes the following form instead:

$$X_A = \sum_{n=0}^{\infty} \left( \int_A f_n d\mu \right) Z_n, \quad A \in \mathcal{B}_{\text{fin}}$$  \hspace{1cm} (1.1)

where $X_A$ is a centered Gaussian process indexed by $\mathcal{B}_{\text{fin}}$, $\{ f_n \}_{n \in \mathbb{N}_0}$ is an orthonormal basis (ONB) in $L^2(\mu)$, and $\{ Z_n \}_{n \in \mathbb{N}_0}$ is an i.i.d. $N(0,1)$ system. These new kernels encompass realistic cases when random measurements (random variables) might not be performed at points, but instead at sample sets that are selected from suitable choices of sigma-algebras. For more details on the more general context of Parseval frames in the measure category, see Section 3.2 below, and [JT22, Cor 3.13 and sect. 5].

**Main results.** The presentation of our main results is organized as follows: In Section 2 we outline and motivate our principal-component analysis (PCA) tools to be discussed in detail inside the paper. This entails Hilbert space techniques, as well as new transforms (see especially Section 4). Specific Hilbert space techniques are use in our design of expansions of data and images (e.g., analysis/synthesis). Our expansions are realized in suitable choices of Hilbert spaces (e.g., $L^2$-spaces, and kernel Hilbert spaces), and with choices of orthonormal bases (ONBs), and generalized bases, going by the name of frames. We further introduce the important notion of maximal variance. And we discuss algorithms to be used. We include illustration of the methods making use of real-life digital images, Figure 2.1. They serve to illustrate (in particular) the effect of PCA-choices, reflected directly in the visual display of these digital images. Section 3 covers multiple interrelated tools, playing a key role in our main results. The first of these tools is our use of kernels (also called reproducing kernels) and their associated Hilbert spaces, called reproducing kernel Hilbert spaces (RKHSs.) The use of these RKHSs yields powerful
tools, and they allow explicit realizations of a variety of embeddings of non-linear structures into linear ones (Hilbert spaces). This is useful in turn for our solution to optimization questions, see Proposition 3.16 and Corollary 3.18. Our results are tested on datasets, see Figures 3.1 and 3.2. The RKHSs are also powerful tools for analysis of stochastic processes (Gaussian), powerful tools for simulation, referred to as Monte Carlo simulation, and for the statistics approach to PCA. They are presented in detail in Section 4, and they are used in our design of new transforms, see e.g., Corollary 4.5.

Organization. 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 3.10. 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 3.12 deals with kernel PCA for nonlinear data dimension reduction (see Examples 3.19 and 3.20). In the remaining part of Section 3, and in Section 4, we turn to the case of Gaussian kernels and the corresponding stochastic processes.

Both tools, PCA and KPCA, have found wide applications. 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. The points where our results go beyond the earlier literature includes the following list of five closely interrelated items: (i) Our design and use of kernel tricks (specific reproducing kernel Hilbert spaces (RKHSs) integrated into new designs of PCA-tools for image-analysis, thus serving as practical tools in dimension-reduction algorithms; (ii) Our combination of RKHS-tools with spectral theory and Hilbert space frame-estimates; — this serves in turn to make more precise both specific numerical PCA-algorithms, and their error estimates; (iii) Our use of RKHSs in creating explicit and practical embeddings of non-linear data into suitable linear spaces (feature spaces); — applications to optimization (Section 3.1); (iv) Stochastic analysis: our combination of kernel-tools with analysis of Gaussian processes/fields, and our use of this in designs of new stochastic transforms (Section 4), — applications to Monte Carlo simulation; and (v) A new dynamical PCA-analysis (Section 3.6)

Our results in Section 3 include Lemma 3.10, Theorem 3.12, and Corollary 3.18 (also see Theorem 2.13); 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 expansions are representations for certain stochastic processes. The list of results inside our paper of special relevance to KL expansions/transforms are Theorem 2.13, Proposition 3.21, and Corollary 4.5. 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), where in this section is illustrated with image principal components to let the readers better understand the concept of dimension. 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. For the benefit of readers not familiar with this theory, the following references will serve to outline the general framework for KL-expansions [Pyc21, DM08, Mok07].

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 [SF83, Jac74]. 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 [F.R01], and it was later developed by Harold Hotelling in the 1930s [Hot33]. 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 matrix $X = [X_1, \cdots, X_m]$ into a matrix $Y$ according to

$$Y = A (X - m_X),$$

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 $X$

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

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

$$C = cov(X) = E \left( (X - m_X)(X - m_X)^T \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T - m_X m_X^T. \quad (2.3)$$

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

$$C(i, i) = E((X_i - m_{X_i})^2). \quad (2.4)$$
Also, $C(i,j) = E \left( (X_i - m_{X_i}) (X_j - m_{X_j}) \right)$ is the covariance between $X_i$ and $X_j$.

4. Compute the eigenvectors and eigenvalues, $\lambda_i$ 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 lower 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 is versatile. It can be used for a diverse set of linear data dimension-reduction problems, and it is illustrated below in Figures 2.1, 2.2, 2.3, and 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:
\[
f(x, y) = \begin{pmatrix}
f(0, 0) & f(0, 1) & \cdots & f(0, N - 1) \\
f(1, 0) & f(1, 1) & \cdots & f(1, N - 1) \\
\vdots & \vdots & \ddots & \vdots \\
f(M - 1, 0) & f(M - 1, 1) & \cdots & f(M - 1, N - 1)
\end{pmatrix}. \tag{2.8}
\]

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 \( \lambda_i \) in step 4 above, thus yielding smallest error and smallest entropy in the computation (see e.g., [Son08]).

Figure 2.1 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

\[
I = w_1 R + w_2 G + w_3 B = f_R + f_G + f_B
\]  

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 equation \((2.1)\), each \( f_R, f_G \) and \( f_B \) are taken as \( Y \) with the pixel values as matrix entries in equation \((2.8)\), resulting in principal components for each \( f_R, f_G \) and \( f_B \) which are Figures 2.1b, 2.1c and 2.1d. Here the original image \( I \) is Figure 2.1a. The original image used for Figure 2.1a, is in red, green and blue color components which are \( f_R, f_G, \) and \( f_B \). But PCA transform on image \( I \), that is, each \( f_R, f_G \) and \( f_B \) will result in principal components of each \( f_R, f_G \) and \( f_B \) as shown in Figure 2.1.

Here, after PCA transformation, instead of RGB components a new three components are used and this is shown in the Figures 2.1b, 2.1c and 2.1d. Note that there are more principal components. 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 the direction of principal components.

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 in Figure 2.1. 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...
Figure 2.1. The three principal components of the original Jorgensen image obtained by performing PCA on each red, green and blue color components of the original image, $f_R$, $f_G$ and $f_B$. These correspond to the largest eigenvalues.
the quantity of eigenvectors. Thus the dimension of the image matrix is reduced. (Also see e.g., [dES12].)

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.

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 principal (dimension) 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.

Remark. The issues of PCA as a tool for feature selection is subtle and multifaceted, and we will not be able to cover them. To help readers appreciate some of these difficulties, we list three known limitations of PCA-feature selection, with brief mention of ways they are usually addressed. The use PCA as a tool for feature selection entails selection of variables according to the magnitude (from largest to smallest in absolute values) of their coefficients (loadings). The list of limitations of these methods includes the following three issues (and more): (1) measurements from all of the original variables are used in a projection onto lower dimensional space, (2) PCA favors linear relationships, and (3) PCA does not directly take into account the potential multivariate nature of the data structures under consideration (such as higher order interaction between variables).

For dealing with (1), elaborate screening methods have been proposed. A popular tool for addressing (2) is use kernel PCAs (i.e., the kernel trick, referring to choices of reproducing kernels) leading in turn to suitable embeddings of nonlinear relationships into a lower dimensional linear spaces; see Section 3 below. Item (3) in the list is dealt with in a variety of different ways, e.g., with use of permutation-based measures. We shall not discuss these issues in detail, but we refer instead to the papers cited above, covering these diverse PCA-issues.
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.

A further detailed discussion of feature maps, feature spaces, and feature selection, in the context of positive definite kernels, and Hilbert space spectral theory, is included below in a subsection below Definition 3.2, titled “Feature maps,” as well as in subsection 3.5.

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.

PCA aims to remove redundancies and describe the data with fewer properties in a way that it performs a linear transformation moving the original set of data to a new space spanned by principal component. This done by constructing a new set of properties based on a combination of the old properties. The properties that present low variance are considered not useful. PCA looks for properties that have maximal variation across the data to make the principal component space. The eigenvectors in Section 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 a 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 they capture 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 \( H \), and an orthonormal basis (ONB) \( \psi_i \) 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 $(\psi_i)$ and $(\phi_i)$ be orthonormal bases (ONB), with index set $I$. Usually

$$I = \mathbb{N} = \{1, 2, \ldots\}.$$  \hspace{1cm} (2.10)

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

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

$$\langle u,v \rangle = \text{inner product} \in \mathbb{C},$$ \hspace{1cm} (2.11)

$$|u\rangle\langle v| = \text{rank-one operator} , \mathcal{H} \to \mathcal{H},$$ \hspace{1cm} (2.12)

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

$$|u\rangle\langle v| w = |u\rangle \langle v,w \rangle = \langle v,w \rangle u, \text{ for all } w \in \mathcal{H}.$$ \hspace{1cm} (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 $w_\alpha, \alpha = 1, 2, \ldots, n$ where the relative frequency $w_\alpha$ satisfies the probability axioms:

$$w_\alpha \geq 0, \sum_{\alpha=1}^{n} w_\alpha = 1.$$  \hspace{1cm} (2.14)

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

$$\int_{a}^{b} |f_\alpha(\xi)|^2 d\xi = 1.$$  \hspace{1cm} (2.15)
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

\[
f_\alpha(\xi) = \sum_{i=1}^{\infty} x_\alpha(i) \psi_i(\xi) \tag{2.14}
\]

with

\[
x_\alpha(i) = \int_a^b \psi_i^*(\xi) f_\alpha(\xi) \, d\xi. \tag{2.15}
\]

Here, \( x_\alpha(i) \) is the component of \( f_\alpha \) in \( \psi_i \) coordinate system. With the normalization of \( f_\alpha \) we have

\[
\sum_{i=1}^{\infty} |x_\alpha(i)|^2 = 1.
\]

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

\[
f_\alpha(\xi) = \int_a^b f_\alpha(\eta) \left[ \sum_{i=1}^{\infty} \psi_i^*(\eta) \psi_i(\xi) \right] d\eta
= \sum_{i=1}^{\infty} \langle \psi_i, f_\alpha \rangle \psi_i(\xi)
\]

by definition of ONB. Note this involves orthogonal projection. \( \blacksquare \)

We here give mathematical background of PCA.

Let \( H = L^2(a, b) \). \( \psi_i : H \to l^2(\mathbb{Z}) \) 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 \( \{ x_\alpha(i) \} \) 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 \( y_{\alpha(j)} \) be the component of \( f_\alpha \) in \( \{ \phi_j \} \) where it can be expressed in terms of \( x_\alpha(i) \) by a linear relation

\[
y_{\alpha(j)} = \sum_{i=1}^{\infty} \langle \phi_j, \psi_i \rangle x_\alpha(i) = \sum_{i=1}^{\infty} U_{i,j} x_\alpha(i)
\]

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, \( x_\alpha(i) \) can be written in terms of \( y_{\alpha(j)} \) under the following relation

\[
x_\alpha(i) = \sum_{j=1}^{\infty} \langle \psi_i, \phi_j \rangle y_{\alpha(j)} = \sum_{j=1}^{\infty} U_{i,j}^{-1} y_{\alpha(j)}
\]

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

\[
f_\alpha(\xi) = \sum_{i=1}^{\infty} x_\alpha(i) \langle \psi_i, \xi \rangle = \sum_{j=1}^{\infty} y_{\alpha(j)} \langle \phi_j, \xi \rangle.
\]
So \( U(x_{\alpha(i)}) = (y_{\alpha(j)}) \) which is coordinate change, and \( \sum_{i=1}^{\infty} x_{\alpha(i)} \psi_i(\xi) = \sum_{j=1}^{\infty} y_{\alpha(j)} \phi_j(\xi) \), and

\[
x_{\alpha(i)} = \langle \psi_i, f_\alpha \rangle = \int_a^b \psi_i^*(\xi) f_\alpha(\xi) \, d\xi.
\]

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

\[
Q_i = \sum_{\alpha=1}^{n} w_\alpha |x_{\alpha(i)}|^2,
\]

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

\[
Q_i \geq 0, \quad \sum_i Q_i = 1.
\]

See also \[Wat67\].

Let \( Q_i = G(i, i) = \sum_{\alpha=1}^{n} w_\alpha x_{\alpha(i)} x_{\alpha(i)}^* \). Here, \( Q_i = G(i, i) \) is the variance, and \( G(i, j) \) determines the covariance between \( x_{\alpha(i)} \) and \( x_{\alpha(j)} \). 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.,

\[
\int_a^b G(\xi, \xi') \Theta_k(\xi') \, d\xi' = \lambda_k \Theta_k(\xi).
\]  

(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 \( v_\alpha \)'s whose components are \( x_{\alpha(i)} \) in the \( \psi_i \) coordinate system, we have

\[
\sum_{i'} G(i, i') t_{k(i')} = \lambda_k t_{k(i)}
\]  

(2.17)

where \( t_{k(i)} \) is the \( i^{th} \) component of the vector \( \Theta_k \) in the coordinate system \( \{\psi_i\} \). So we get \( \psi : \mathcal{H} \to (x_{\alpha(i)}) \) and also \( \Theta : \mathcal{H} \to (t_{k(i)}) \). The two ONBs result in

\[
x_{\alpha(i)} = \sum_k c_{\alpha(k)} t_{k(i)} \text{ for all } i, \quad c_{\alpha(k)} = \sum_i t_{k(i)}^* x_{\alpha(i)},
\]

which is the Karhunen-Loève expansion of \( f_\alpha(\xi) \) or vector \( v_\alpha \). Hence \( \{\Theta_k(\xi)\} \) is the K-L coordinate system dependent on \( \{u_\alpha\} \) 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 \( A_{ik} = t_{k(i)}^* t_{k(i)} \) which is a double stochastic matrix. Then we have the following eigendecomposition of the
covariance matrix (operator), $G$

$$G = U \begin{pmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & \lambda_{k+1}
\end{pmatrix} U^{-1}.$$  \hspace{1cm} (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 2.13. 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 3.12.

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 $\{\psi_i\}$. In this case, set $\text{tr}(T) := \sum \langle \psi_i, T \psi_i \rangle$. \hspace{1cm} (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

$$c_1 \|f\|^2 \leq \sum_{\alpha \in A} |\langle h_\alpha, f \rangle|^2 \leq c_2 \|f\|^2$$

for all $f \in \mathcal{H}$. \hspace{1cm} (2.20)

In the case when $c_1 = c_2 = 1$, then $(h_\alpha)_{\alpha \in A}$ is said to be a **Parseval frame**.

Given a frame $(h_\alpha)_{\alpha \in A}$ in $\mathcal{H}$, then $L : \mathcal{H} \to l^2$, defined as $Lf = (\langle h_\alpha, f \rangle)_\alpha$, is called the **analysis operator**. The adjoint $L^* : l^2 \to \mathcal{H}$ is called the **synthesis operator**. (See Lemma 2.6 below.)

For more details about frame theory, we refer to [BJ02, HKLW07, HWW05, BH19, CCEL15, CCK13].

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

$$L : f \mapsto (\langle h_\alpha, f \rangle)_{\alpha \in A}.$$  \hspace{1cm} (2.21)

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

$$L^*(c_\alpha) = \sum_{\alpha \in A} c_\alpha h_\alpha$$  \hspace{1cm} (2.22)

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

$$L^* L = \sum_{\alpha \in A} |h_\alpha \rangle \langle h_\alpha|.$$  \hspace{1cm} (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 \(\alpha \in A\).

\[
G := \sum_{\alpha \in A} w_\alpha |f_\alpha \rangle \langle f_\alpha|
\]  

(2.24)
is called a frame operator associated to \((f_\alpha)\).

Remark 2.8. If we take vectors \((f_\alpha)\) from a frame \((h_\alpha)\) 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

\[
\text{tr}(G) = \sum_{\alpha \in A} w_\alpha.
\]  

(2.25)

Definition 2.10. Suppose we are given a frame operator

\[
G = \sum_{\alpha \in A} w_\alpha |f_\alpha \rangle \langle f_\alpha|
\]  

(2.26)

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

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

(2.27)

are called the error or the residual of the projection.

If \((\psi_i)_{i \in \mathbb{N}}\) is an ONB, then the projection \(Q_n := \text{proj span} \{\psi_1, \ldots, \psi_n\}\) that is given by \(Q_n := \sum_{i=1}^{n} |\psi_i \rangle \langle \psi_i|\), and for each \(i\), \(|\psi_i \rangle \langle \psi_i|\) is the projection onto the one-dimensional subspace, \(\mathbb{C} \psi_i \subset \mathcal{H}\).

In our present calculations we have found useful Dirac’s terminology of brackets and ket-bras. Both refer to operations on pairs of vectors in Hilbert space. In Dirac’s notation, the bracket of two vectors is simply the inner product. By contrast, for a pair of unit-norm vectors, the ket-bra is the corresponding rank-one projection. See 2.2, (2.11)–(2.13), Lemma 2.6, and Definition 2.7.

Lemma 2.11. When \((\psi_i)\) 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))

\[
E_n^\psi = \text{tr}(GQ_n^\perp).
\]  

(2.28)

Proof. The proof follows from the previous facts, using that

\[
\|f_\alpha - Q_nf_\alpha\|^2 = \|f_\alpha\|^2 - \|Q_nf_\alpha\|^2
\]

for all \(\alpha \in A\) and \(n \in \mathbb{N}\). The expression 2.27 for the error term is motivated as follows. The vector components \(f_\alpha\) in Definition 2.24 are indexed by \(\alpha \in A\) and are assigned weights \(w_\alpha\). Using these, we replace the vectors \(f_\alpha\) with finite approximations \(Q_nf_\alpha\), then the error term 2.27 measures how well the approximation fits the data.

Lemma 2.12. \(\text{tr}(GQ_n) = \sum_{\alpha \in A} w_\alpha \|Q_nf_\alpha\|^2\).
Proof.

\[ \text{tr}(GQ_n) = \sum_i \langle \psi_i | GQ_n \psi_i \rangle = \sum_i \langle Q_n \psi_i | GQ_n \psi_i \rangle = \sum_{i=1}^n \langle \psi_i | G \psi_i \rangle \]
\[ = \sum_{\alpha \in A} w_{\alpha} \sum_{i=1}^n |\langle \psi_i | f_{\alpha} \rangle|^2 = \sum_{\alpha \in A} w_{\alpha} \| Q_n f_{\alpha} \|^2, \]
as claimed.

Proof. Proof of Lemma 2.11 continued. The relative error is represented by the difference:
\[ \sum_{\alpha \in A} w_{\alpha} - \sum_{\alpha \in A} w_{\alpha} \| Q_n f_{\alpha} \|^2 = \sum_{\alpha \in A} w_{\alpha} \| f_{\alpha} \|^2 - \sum_{\alpha \in A} w_{\alpha} \| Q_n f_{\alpha} \|^2 \]
\[ = \sum_{\alpha \in A} w_{\alpha} (\| f_{\alpha} \|^2 - \| Q_n f_{\alpha} \|^2) = \sum_{\alpha \in A} w_{\alpha} - Q_n f_{\alpha} \|^2 \]
\[ = \sum_{\alpha \in A} w_{\alpha} \| \mathcal{Q}_n f_{\alpha} \|^2 = \text{tr}(G \mathcal{Q}_n^{'}). \]

The more general frame operators are as follows: Let
\[ G = \sum_{\alpha \in A} w_{\alpha} P_{\alpha} \]
where \( (P_{\alpha}) \) 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 \( P_{\alpha} \) 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 2.13 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 \( i \)th 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 2.13. 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. Here, the maximum variance is the largest spectral variability in \( (\phi_k) \). These orthogonal vectors give principal axes, i.e., directions in \( (\phi_k) \) with the largest variations. The PCA transformation with the covariance operator \( G \) will then be defined such that the first principal component has the largest possible variance (accounting for as much of the variability in \( (\phi_k) \) 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 ONB.

The development of obtaining ONB in the spectral theorem can be found at section III of [JS07].

**Theorem 2.13.** 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 \((\phi_k)\) consisting of eigenvectors, i.e.,

\[
G\phi_k = \lambda_k \phi_k, \quad k \in \mathbb{N},
\]

(2.30)
called the Karhunen-Loève data or principal components. The spectral data may be constructed recursively starting with

\[
\lambda_1 = \sup_{\phi \in \mathcal{H}, \|\phi\| = 1} \langle \phi, G\phi \rangle = \langle \phi_1, G\phi_1 \rangle, \quad \text{and}
\]

\[
\lambda_{k+1} = \sup_{\phi \perp \phi_1, \phi_2, ..., \phi_k} \langle \phi, G\phi \rangle = \langle \phi_{k+1}, G\phi_{k+1} \rangle.
\]

(2.31)

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

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

(2.32)

where \( Q_n^\psi \) is the sequence of projections, deriving from some ONB \((\psi_i)\) and are arranged such that the following holds:

\[
\langle \psi_1, G\psi_1 \rangle \geq \langle \psi_2, G\psi_2 \rangle \geq ...\).

Hence we are comparing ordered sequences of eigenvalues with sequences of diagonal matrix entries.
Lastly, we have
\[ \text{tr} (G) = \sum_{k=1}^{\infty} \lambda_k = \sum_{k=1}^{\infty} \langle \psi_k, G \psi_k \rangle < \infty. \]

The assertion in Theorem 2.13 is the validity of
\[ E_{\phi_n}^\psi \leq E_{\psi_n}^\psi \]
for all \((\psi_i) \in ONB(\mathcal{H})\), and all \(n = 1, 2, \ldots\); and moreover, that the infimum on the RHS in (2.33) is attained for the KL-ONB \((\phi_k)\). But in view of our lemma for \(E_{\psi_n}^\psi\) (2.11), we see that (2.33) is equivalent to the system (2.32) in the Arveson-Kadison theorem. \(\square\)

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

**Comparison of KPCA to standard PCA** (brief sketch.) The standard PCA always finds linear principal components which serve to represent a given large data set in suitable lower dimension. But PCA will fail to find good representative directions when applied to most non-linear data of interest. For non-linear and more tricky data sets, one turns instead to Kernel PCA (KPCA). It has the following advantages and features (not a complete list): (i) KPCA will perform PCA but in a new space. (ii) It uses diverse kernel tricks in order to still find principal components, but they will be in a different space (typically in a higher dimensional space.) (iii) KPCA will use Hilbert space geometry in order to still find new directions based on families of kernel-matrices. Via geometric algorithms, KPCA will then extract corresponding eigenvalues (corresponding to a suitable number of observations.) However, (iv) the computational complexity for KPCA, dictated by extraction of principal components, will take more time, as compared to standard PCA. The list of papers on KPCA and applications to diverse sets of non-linear data is long; some recent ones are [HKF+20, ZSS19, Law05, AH11, HLMS04, Nai17, FCB].

In this section, we make precise, and show that the choice of Gaussian kernel is “optimal” for KPCA in capturing maximal variance (i.e., maximal variability of data). There is an extensive literature on PCA and KPCA. Our approach and emphasis are different. For instance, the paper [RT01] deals with certain Least Squares Regression problems with the use of a particular (small) class of Reproducing tools involving Kernel Hilbert Space. While there are connections to our present themes, our present results are much more general, offering in particular specific spectral theoretic, and stochastic tools, penalty terms, especially relevant for digital images, and for dynamical PCA. The focus of [SSM98] is the design of certain Nonlinear Component Analysis as a particular class of Kernel Eigenvalue Problems. By comparison, our present results and applications have a different focus, use different tools both as regards to infinite-dimensional analysis, optimization, and Gaussian fields; as we outline below. This paper [KW71] deals with a particular class of splines (numerical analysis). This in turn entails a class of kernels. But the focus of [KW71] is quite different from ours. The paper [WMd97] deals with matrix algorithms close in spirit to our present paper, but our focus and applications are global in nature, and are very different as we outline below. The
book [CST00] deals with kernel methods and some of their applications to support
vector machines. This is an exciting application of feature selections, and learning
algorithms, but the aim and the tools developed in this book are quite different
from our present focus.

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 focus on a number of technical
results. They are applied in turn to an analysis of dimension reduction for nonlinear
data. This allows us to obtain extensions of some PCA-results from [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 3.20.

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 sub-
manifolds. A popular technique to tackle the latter case is kernel PCA. It first maps
data into a higher dimensional space $H(K)$, and performs PCA there. Here $H(K)$
is the reproducing kernel Hilbert space (RKHS) associated with a given positive
definite kernel $K$. (Recall that an RKHS is a Hilbert space $H$ of functions defined
on a set $X \neq \emptyset$, such that the evaluation functionals $H \ni f \to f(x) \in \mathbb{C}$ are
bounded for all $x \in X$. See also Definition 3.2 below.) The mapping in this context
presumably sends a nonlinear submanifold in the input space to a linear subspace
in $H(K)$. For example, in classification problems, a kernel is usually chosen so that
the mapped data can be separated by a linear decision boundary in $H(K)$ (see
Figure 3.1).

Remark 3.1. It would be intriguing to compare Smale’s Dimension Reduction al-
gorithm 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].

As applications of the theory presented in the last two sections, 3 and 4, we include
a discussion of new non-linear, and real-life data/simulated examples. Moreover,
inside the sections, we offer explanations for our use of RKHS-theory, as applied
in new KPCA based Gaussian process simulations. In particular, we outline
how they serve to produce optimal choices for detection of maximal variance.

The optimization issues in the present section refer to details in subsections 3.3
and 3.4; see especially Lemma 3.10, Remark 3.11 and Theorem 3.12, all adapted
to the PCA issues at hand. Our notation for solutions to optimization problems
is “argmax” (standard terminology in the optimization literature.) The role of
kernels enters via associated choices of feature maps $\Phi$, see (3.45), (3.46), (3.47),
By a \textit{feature} we mean an individual measurable property or characteristic. One chooses informative, discriminating and \textit{independent} features to be used in algorithmic constructions in machine learning, in pattern recognition, and regression. The concept of “feature” is also related to that of explanatory variables used in statistical techniques, see Section 4 below. Vector spaces associated with particular features are called the \textbf{feature spaces}. In KPCA, the choices of \textit{feature spaces} come about from specified \textit{kernels} (assumed positive definite, p.d.), see Definition 3.2 below. If $K$ is a fixed p.d. kernel defined on general non-linear sets of data, a natural choice of feature space is then in the form of a \textit{Hilbert space} called the \textit{reproducing kernel Hilbert space} (RKHS), written $\mathcal{H}(K)$, see Remark 3.3 below.

\textbf{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

$$\sum_{i,j=1}^{N} c_i c_j K(x_i, x_j) \geq 0$$

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) $\mathcal{H}(K)$ and a mapping $\Phi : S \to \mathcal{H}(K)$ such that

$$K(x, y) = \langle \Phi(x) , \Phi(y) \rangle_{\mathcal{H}(K)}.$$  \hspace{1cm} (3.2)

The function $\Phi$ in (3.2) is called a \textit{feature map} for the problem.

Moreover, the following reproducing property holds:

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

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

While the theory of \textit{positive definite kernels} and their RKHSs have served for a long time as powerful tools in diverse areas of pure and applied mathematics, the particular applications of them we need here are of a more recent vintage; for background literature, see e.g., [AJ15, BTA04, JT15, JT18a, JT18b]. The theory of \textit{frames} (non-orthogonal expansions, Definition 2.5) also plays an essential role in our present approach to PCA, and the list of relevant papers for background includes [HKLW07, HWW05, CCEL15, CCK13].

\textbf{Feature maps.} In the subsequent discussion, reference to \textit{kernel} will always mean positive definite kernel, see (3.1). For each choice of kernel, defined on data sets (typically non-linear configurations), there will then be associated \textit{feature maps}. Generally, a feature map $\Phi$ is simply a \textit{function} which maps sets of data configurations (non-linear) into some choice of \textit{feature space}, a Hilbert space, say $\mathcal{L}$. Typically the \textit{inner product} in $\mathcal{L}$ will serve to model correlation, or regression numbers, or other measurements corresponding to other features of interest. Given a kernel $K$, a natural choice of feature space $\mathcal{L}$ is the RKHS defined directly from $K$, see (3.2), but there are other possibilities. In a different choice for $\Phi$, the range of $\Phi$, and the inner product on the right-hand side in (3.2) might be different. The main logic in the use of kernels in \textit{machine learning} is that it yields representations of learning algorithm for data sets that will then become more amenable to
regression analysis, detection, learning, and to classification. Analysis is aided with
the use of powerful tools from the theory of linear operators in Hilbert space; see
e.g., [HKLW07, JT18b, SZ07]. We stress that the phrase feature map is broad, and
that a wide variety of functions $\Phi$ may serve as feature maps. But the main use
of them relates to suitable choices of kernels. Support Vector Machines (and other
kernel-based methods) make use both implicit and explicit feature maps. This leads
to remapping of data and will allow non-linearly separable data sets to get linear
representations. The feature space leads to effective separation of data via suitable
choices of hyperplanes in higher dimension. But reaching these dimensions might be
computationally expensive because bad choices of feature mappings might require
many computations.

Remark 3.3. $H(K)$ may be chosen as the Hilbert completion of

$$\text{span}\{K_x := K(\cdot, x)\}$$

with respect to the $H(K)$-inner product

$$\langle \sum c_i K_{x_i}, \sum d_j K_{x_j} \rangle_{H(K)} := \sum c_i d_j K(x_i, x_j).$$

Initially the LHS in formula (3.5) only refers to finite linear combinations. Hence,
the vector space (3.4) becomes a pre-Hilbert space. (By pre-Hilbert space we mean
an inner product space that is not complete.) The RKHS $H(K)$ 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 3.12 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 $j^{th}$ column vector. Here, $m$ is the number of features, and $n$ the
number of sample points. Set

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

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

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

$$\langle X, Y \rangle = tr(X^*Y).$$

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

$$HS(H) \ni |u\rangle \langle v| \longrightarrow u \otimes v \in H \otimes H,$$

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

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

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 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:

1. a $L^2(\mu)$-square applied to a difference, and
2. 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 (3.14) in Theorem 3.5 below. The following makes precise choices of optimal kernels.

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

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

is densely defined and closable.

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

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

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

$$L^2(\mu) \xrightarrow{T^*} \mathcal{H}(K) \quad (3.11)$$

be the corresponding adjoint operator, i.e.,

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

Our present RKHS/$L^2(\mu)$ framework is close to that of [SS16]. But we have included the results we need in our present framework, and for our purpose.

**Theorem 3.5.** Let $K$, $\mu$, $\psi$, $\beta$ be as specified above; then the optimization problem

$$\inf_{f \in \mathcal{H}(K)} Q_{\psi,\beta}(f) \quad (3.13)$$

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

$$F = (\beta I + T^*T)^{-1} T^* \psi$$  \hfill (3.14)

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

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

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

for all $h$ in a dense subspace in $\mathcal{H}(K)$. The desired conclusion follows. Note that convexity of the function $Q_{\psi,\beta}$ makes these conditions sufficient as well. \qed

As already mentioned, the general conclusion in the present Theorem 3.5 connects to both many uses of kernel techniques in classical analysis, as well as to new and exciting applications. In the second group, we mention the following approach to Neural Networks [YO21], leading in turn to encoding of deep regularization into training of inner layers which make up specific Neural Network (NN) constructions, with the use of kernel flows. For details, we refer to [YO21] and the cited literature for details. But here we only stress that, as a step in particular NN-designs, one makes use of an recursive iteration of substitution into a fixed kernel (the kernel resulting from the substitution iterations is called a warped kernel); and then, at each step, there will be an RKHS-optimization of a term that arises as a special case of (3.13), see [YO21, eq (1.2)].

**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 $\mathcal{H}(K)$ 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

$$Q(f) = \sum_{i=1}^m \left[ f(t_i) - y_i \right]^2 + \beta \| f \|^2_{\mathcal{H}(K)} , \quad f \in \mathcal{H}(K).$$  \hfill (3.15)

We introduce the associated dual pair of operators as follows:

$$T : \mathcal{H}(K) \rightarrow \mathbb{R}^m \simeq l^2_m, \text{ and}$$

$$T^* : l^2_m \rightarrow \mathcal{H}(K)$$  \hfill (3.16)

where

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

$$T^*y = \sum_{i=1}^m y_i K(\cdot,t_i) \in \mathcal{H}(K).$$  \hfill (3.17)
for all $\vec{y} = (y_i) \in \mathbb{R}^m$.

Note that the duality then takes the following form:

$$
\langle T^* y, f \rangle_{\mathcal{H}(K)} = (y, T f)_{t_m^2}, \quad \forall f \in \mathcal{H}(K), \forall y \in t_m^2;
$$

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:

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

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

where $K_m$ denotes the $m \times m$ matrix

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

The following makes precise choices of optimal kernels.

**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 (3.22).

**Proposition 3.7.** Fix $\beta > 0$; consider the optimization problem with

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

Then the unique solution to (3.23) is given by

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

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:

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

and by (3.20)-(3.21), we further get

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

where the dot $\cdot$ refers to a free variable in $X$. An evaluation of (3.25) on the sample points yields:

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

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

$$
\vec{F} = (\beta I_m + K_m)^{-1} K_m \vec{y}.
$$
Now substitute (3.27) into (3.26), 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 \textit{a priori} equipped with any additional structure. Consideration of stochastic processes will always assume some fixed probability space, \((\Omega, \mathcal{A}, \mathbb{P})\) where \(\Omega\) is a set of sample points; \(\mathcal{A}\) is a \(\sigma\)-algebra of events, fixed at the outset; and \(\mathbb{P}\) is a probability measure defined on \(\mathcal{A}\). A given process \((X_s)_{s \in S}\) is then said to be \textit{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

\[ K^{(N)}_{ij} := \left( \mathbb{E} (X_{s_i} X_{s_j}) \right)_{i,j=1}^N; \]

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

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

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

\[ \mathbb{P} ((X_{s_1}, \cdots, X_{s_N}) \in A_N) = \int_{A_N} g_N (x) \, dN x \]

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

\[ \mathbb{E} (\overline{X}_s X_t) := K_X (s,t) \]

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

Then the following general results hold (see e.g., [AJL11, AJL12, AJS14, AJ15, JT15, JT16a, JT16b, AJL17, JT18a, JT18b]):

(i) Every positive definite kernel \(S \times S \xrightarrow{K} \mathbb{C}\) arises as in (3.31) 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 \(f_n : S \to \mathbb{C}, n \in \mathbb{N}\),

\[ K (s, t) = \sum_{n \in \mathbb{N}} f_n (s) f_n (t), \]

absolutely convergent on \(S \times S\).

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

(iv) Given (3.32), 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

\[ X_s (\cdot) := \sum_{n \in \mathbb{N}} f_n (s) Z_n (\cdot) \]
is valid in $L^2$ of the underlying probability space, and $\mathbb{E}(X_s X_t) = K(s, t)$ on $S \times S$.

The following makes precise choices of optimal kernels.

**Note 3.8.** When a fixed Gaussian process $(X_s)$ is given, then the associated decomposition (3.33) 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 2.13 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 3.9.** 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 (3.32) and (3.33). 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 3.12 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:

$$\arg\max \left\{ ||QX||^2_{HS} : ||Q||^2_{HS} = k \right\}, \quad (3.34)$$

where $Q$ is a selfadjoint projection.

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

$$\arg\max \left\{ ||Q\Phi(X)||^2_{HS} : ||Q||^2_{HS} = k \right\}, \quad (3.35)$$

where

$$\Phi(X) = \left[ \Phi(x_1) \cdots \Phi(x_n) \right]. \quad (3.36)$$

It is understood that $||\cdot||_{HS}$ as in (3.35) refers to the Hilbert-Schmidt class in $B(\mathcal{H}(K))$.

Indeed, both (3.34) and (3.35) are finite dimensional instances of Theorem 2.13. See Section 2.5 and details below.

**A Finite Frame**

Let $X$, $K$, and $\Phi$ 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
\[
Lf = \sum_{j=1}^n \langle \Phi(x_j), f \rangle_{\mathcal{H}(K)} \delta_j, \; f \in \mathcal{H}(K),
\]
where $\delta_j$ is the standard ONB in $\mathbb{C}^n$. The adjoint $L^* : \mathbb{C}^n \to \mathcal{H}(K)$ is given by
\[
L^*c = \sum_{j=1}^n \Phi(x_j) c_j, \; c = (c_j) \in \mathbb{C}^n.
\]
(3.38)

It follows that
\[
L^*L = \sum_{j=1}^n |\Phi(x_j)\rangle \langle \Phi(x_j)| = \Phi(X) \Phi(X)^*.
\]
(3.39)

(Lemma 3.10. Let $L^*L$ be as in (3.39), 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
\[
|\phi_1\rangle \langle \phi_1| = \operatorname{argmax} \left\{ \|Q_1 \Phi(X)\|_{HS}^2 : \|Q_1\|_{HS}^2 = 1, \; Q_1 = Q_1^* = Q_1^2 \right\}
\]
\[
= \operatorname{argmax} \left\{ \operatorname{tr} (Q_1G) : \|Q_1\|_{HS}^2 = 1 \right\}.
\]
Equivalently, the best rank-1 approximation to $G$ is
\[
\lambda_1^2 |\phi_1\rangle \langle \phi_1|.
\]
(See (3.36), and the frame operator in (2.26).)

Remark 3.11. Note that the conclusion of the lemma yields a solution the optimization problem we introduced above. Indeed, in the statement of the lemma (see (3.40)) we use the standard notation \texttt{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 (3.40), is attained for a definite rank-one operator.

\textbf{Proof of Lemma 3.10.} Note that
\[
\|Q_1 \Phi(X)\|_{HS}^2 = \operatorname{tr} (Q_1 \Phi(X) \Phi(X)^*) = \operatorname{tr} (Q_1 G).
\]
Let $w$ be a unit vector in $\mathcal{H}(K)$, and set $Q_1 = |w\rangle \langle w|; \text{ then}
\[
\operatorname{tr} (Q_1 G) = \sum_j \lambda_j^2 |\langle w, \phi_j ||^2.
\]
(3.41)

Since $\sum_j |\langle w, \phi_j ||^2 = \|w\|^2 = 1$, the r.h.s. of (3.41) is a convex combination of $\lambda_j^2$'s; therefore,
\[
\sum_j \lambda_j^2 |\langle w, \phi_j ||^2 \leq \lambda_1^2
\]
and equality holds if and only if $|\langle w, \phi_1 || = 1$, and $\langle w, \phi_j \rangle = 0$, for $j > 1$.

Note that, given (3.41), the conclusion of the lemma follows directly from comparison of positive series.
Lemma 3.10 can be applied inductively which yields the best rank-1 approximation at each iteration. In fact, the result holds more generally; see Theorem 3.12 below.

Our present RKHS/$L^2(\mu)$ framework is close to that of [SS16]. But we have include the results we need in our present framework, and for our purpose.

**Theorem 3.12.** Let $T : \mathcal{H} \to \mathcal{H}$ be a Hilbert Schmidt operator, and let

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

(3.42)

with $\lambda_1^2 \geq \lambda_2^2 \geq \cdots$, and $\lambda_j \to 0$ as $j \to \infty$. Then

$$\sum_{j=1}^n |\phi_j\rangle\langle \phi_j| = \arg\max \left\{ \|Q_n T\|_{HS}^2 : \|Q_n\|_{HS}^2 = n, Q_n = Q_n^* = Q_n^2 \right\}.$$ 

Note that $\|Q_n T\|_{HS}^2 = tr(Q_n G)$, where $G := TT^*$. See Section 2.5.

**Proposition 3.13.** (Recall the assumption that $T$ is Hilbert Schmidt is equivalent to $T^*T$ trace class. We then apply the Spectral Theorem to $T^*T$ yielding an orthonormal basis of eigenvectors for $T^*T$, and a convergent representation for the infinite sum in (3.42).)

**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.,

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

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

It follows that

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

(3.43)

$$\Downarrow$$

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

Hence $P_k$ is a spectral projection of $TT^*$. The conclusion of the theorem follows from this.

Note that one may use the usual variational directional derivative argument applied to a sesquilinear form $(x,x) \mapsto \langle x, Tx \rangle_{\mathcal{H}'}$ in any Hilbert space $\mathcal{H}'$, and obtain

$$\frac{d}{dx} \langle x, Tx \rangle_{\mathcal{H}'} = Tx + T^* x, \quad x \in \mathcal{H}.'$$ 

In particular, if $T$ is self-adjoint, then

$$\frac{d}{dx} \langle x, Tx \rangle_{\mathcal{H}'} = 2Tx.$$
Especially, the Hilbert space in (3.43) is the space of Hilbert-Schmidt operators. (This argument is a key step in the proof of the spectral theorem, i.e., taking the variational directional derivative of a bounded bilinear form in Hilbert space.)

The precise meaning of (and the justification for) the generalized gradient assertion (3.43) is included in Lemma 3.14 below. See also the following citations of sources for the underlying operator theory.

The variational arguments in the proof of Theorem 3.12, used in the present setting of Hilbert space, and the variety of projections, can be justified with standard tools from operator theory, including the spectral theorem; see e.g., [JT17, ch.2], [JT21, ch.1], [Con90, ch.2] and also [AK06, BJ02, HKLW07, SZ09a]. We have included more details below:

**Lemma 3.14.** Suppose $T$ is a Hilbert-Schmidt (HS) operator in a separable Hilbert space $\mathcal{H}$. For the directional derivative $\nabla_T := \frac{\partial}{\partial T}$, we have

$$\frac{\partial}{\partial T} \langle x, Tx \rangle_{\mathcal{H}} = |x\rangle\langle x|, \quad x \in \mathcal{H}.$$  

(Here, $|x\rangle\langle x|$ is the rank-1 operator $\mathcal{H} \ni u \mapsto \langle x, u \rangle x$ using Dirac’s notation.)

**Proof.** Let $Q = |x\rangle\langle x|$. Note that $f_Q(T) := \langle x, Tx \rangle = \text{trace}(TQ)$. In fact, if $\{e_n\}_{n=1}^\infty$ is an orthonormal basis in $\mathcal{H}$, then

$$\text{trace}(TQ) = \sum_{n=1}^\infty \langle x, e_n \rangle_{\mathcal{H}} \langle e_n, Tx \rangle_{\mathcal{H}}$$

$$= \langle x, Tx \rangle_{\mathcal{H}}, \quad \text{by the Parseval identity.}$$

But

$$\text{trace}(TQ) = \langle Q, T \rangle_{HS}$$

where $\langle \cdot, \cdot \rangle_{HS}$ denotes the Hilbert-Schmidt inner product.

Now we use a general fact about inner products:

$$|\langle Q, T \rangle_{HS}|^2 \leq \langle Q, Q \rangle_{HS} \langle T, T \rangle_{HS}, \quad \forall Q, T \in \mathcal{HS}$$

where $\mathcal{HS}$ is the Hilbert-Schmidt operators, viewed as a Hilbert space, relative to its trace-inner product (3.44). That is, the Schwarz inequality is applied to the $\langle \cdot, \cdot \rangle_{HS}$-inner product. One obtains maximum when $T \sim Q$. In particular, if $Q = |x\rangle\langle x|$, then $Q^2 = \|x\|^2 Q$ and $\text{trace}(Q^2) = \|x\|^4_{\mathcal{H}}$.

Therefore, it follows from $f_Q(T) := \langle Q, T \rangle_{HS}$ that

$$\nabla_T f_Q = \langle Q, \cdot \rangle_{HS},$$

which is identified with $Q = |x\rangle\langle x|$, since

$$(\mathcal{HS})^* = (\mathcal{HS}).$$

**Remark 3.15.** Our use of frame analysis (see Definition 2.5 and the lemmas) serves as the technical tools for “error estimates.” More specifically, frames are designed such that, for a given problem involving finite-dimensional PCA computations, the frame estimates are designed to allow estimation of the corresponding “error terms.” (See especially 2.10, and Lemma 2.11 and Theorem 2.13.)

For our present considerations of optimization questions, it may be of interest to consider a comparison between the following two settings, one general, and the
other special: On the one hand, there is a variety of (i) general calculus of variation issues in infinite-dimensional contexts; and on the other, (ii) particular optimization questions in the restricted framework of specific choices of pairs of Hilbert spaces (still infinite-dimensional). We note that the context for (i) is wider than that of (ii). On the other hand (as outlined below), for the specialized framework of (ii) considered here, the use of standard Hilbert space geometry, and a little operator theory, offer much simplification. 

Detail: Our present focus is (ii). By contrast, in consideration of (i) one must address separately such technical issues as (a) functional-derivatives, and (b) existence of extremal (optimal) solutions. By contrast, the smaller class of optimization questions from (ii) lend themselves directly to computation in Hilbert space. Here we note that application of Hilbert space geometry will then simplify the two issues (a) and (b) in (i). The point we make for our present optimization question (i.e., a specific choice of a Hilbert space context (ii)), is that a “natural” choice of a pair of Hilbert spaces will be a combination of the two: an appropriate RKHS, and a compatible $L^2(\mu)$-Hilbert space. This then allows for relatively simple solutions.

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 (3.36), i.e.,

$$\Phi (X) = [\Phi (x_1) \cdots \Phi (x_n)]. \quad (3.45)$$

Let $L, L^*$ be the analysis and synthesis operators from (3.37)-(3.38), and

$$L^*L : \mathcal{H}(K) \to \mathcal{H}(K), \quad L^*f = \sum_{j=1}^n (\Phi (x_j), f) \Phi (x_j) \quad (3.46)$$

be the frame operator in (3.39). In particular,

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

In view of Theorem 2.13 and Theorem 3.12, 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

$$LL^* = \Phi (X)^*\Phi (X) = (K (x_i, x_j))_{ij=1}^n, \quad (3.47)$$

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

$$A\delta_j = \Phi (x_j), \quad (3.48)$$

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

$$A^*h = (\langle \Phi (x_j), h \rangle)_{j=1}^n \in \mathbb{C}^n. \quad (3.49)$$

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

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

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

and the assertions follows. \qed
Hence $LL^*: \mathbb{C}^n \rightarrow \mathbb{C}^n$ is the Gramian matrix in $\mathbb{C}^n$ given by

$$LL^* = \Phi(X)^*\Phi(X) = \left(\langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}(K)}\right)_{i,j=1}^n$$

by (3.2) $\langle K(x_i, x_j) \rangle_{i,j=1}^n$; (3.50)

see (3.47).

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

$$LL^* = UD^2U^*, \quad (3.51)$$

$$L^*L = WD^2W^*, \quad (3.52)$$

where $D = \text{diag}(\lambda_j)$ 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 (3.46), i.e.,

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

(3.53)

It also follows from (3.51)–(3.52), that

$$W = L^*UD^{-1}. \quad (3.54)$$

**Remark 3.17.** In the above discussion, $\Phi(X)$ may be centered by removing its mean. Specifically, let

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

be the projection onto span $\{1\}^\perp$, where $1$ denotes the constant vector $[1 \cdots 1]$. Then

$$\tilde{\Phi}(X) := \Phi(X) - \frac{1}{n} \sum_{j=1}^n \Phi(x_j) = \Phi(X) J, \quad (3.55)$$

and so

$$LL^* = \tilde{\Phi}(X)^*\tilde{\Phi}(X) = J\Phi(X)^*\Phi(X)J.$$  

(3.56)

The effect of $J$ in (3.56) is to exclude the eigenspace of the Gramian $\langle K(x_i, x_j) \rangle_{i,j=1}^n$ spanned by the constant eigenvector.

In what follows, we shall always assume $\Phi(X)$ is centered as in (3.55)-(3.56).

**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 $\Phi(x)$ under the feature map can be
projected onto the principal directions in $\mathcal{H}(K)$, via

$$\Phi(x) \mapsto WW^* \Phi(x).$$

The mapping $x \mapsto WW^* \Phi(x)$ is in general nonlinear. See Examples 3.19 and 3.20,
and Figures 3.1 and 3.2 below.

**Corollary 3.18.** Let $L^* = WDU^*$ be as above, assuming $D$ is full rank. For all $x \in \mathbb{C}^m$, the coefficients of the projection $WW^* \Phi(x)$ are

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

(3.57)$$

**Proof.** By (3.54), $W^* = D^{-1} U^* L$, so that

$$W^* \Phi(x) = D^{-1} U^* L \Phi(x)$$

$$= (3.49) D^{-1} U^* \begin{bmatrix}
(\Phi(x_1), \Phi(x))_{\mathcal{H}(K)} \\
\vdots \\
(\Phi(x_n), \Phi(x))_{\mathcal{H}(K)}
\end{bmatrix}$$

$$= (3.2) D^{-1} U^* \begin{bmatrix}
K(x_1, x) \\
\vdots \\
K(x_n, x)
\end{bmatrix}

= \begin{bmatrix}
\lambda_1^{-1} \sum_{j=1}^n u_{j1} K(x_j, x) \\
\vdots \\
\lambda_n^{-1} \sum_{j=1}^n u_{jn} K(x_j, x)
\end{bmatrix}.

□

**Example 3.19** (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 $3 \times 993$, 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

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

By projecting $\Phi(x_j)$ onto the first two principal components, 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 u_{k1} K(x_k, x_j) \\
\lambda_2^{-1} \sum_{k=1}^n u_{k2} K(x_k, x_j)
\end{bmatrix} \in \mathbb{R}^2.

See the general formula in Corollary 3.18; note that $n = 2$ in the present example.

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

This figure is an illustration of the feature selection in Corollary 3.18. For the present binary clustering problem (“∗” vs “o”), the kernel $K$ is centered using (3.55)–(3.56), and the vector $\Phi(X)$ with entries in the Gaussian RKHS $\mathcal{H}(K)$ is projected onto the first two column vectors of the matrix $W$ using (3.57). For a fixed sample point $x$, the two projection coefficients are $\lambda_1^{-1} \sum_{j=1}^{n} \overline{w_j} K(x_j, x)$ and $\lambda_2^{-1} \sum_{j=1}^{n} \overline{w_j} K(x_2, x)$.

Example 3.20 (Dimension Reduction, see Figure 3.2). Let $X$ be a collection of 100 grayscale $256 \times 256$ images of an ellipse, rotated successively by $\pi/100$. Figure 3.2a shows 6 sample images corresponding to different rotation angles. The images are unrolled as column vectors, thus $X$ has dimension $65536 \times 100$.

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 3.2b, 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 $65536 \times 100$ to $1 \times 100$.

In particular, the projection coefficients onto the $k^{th}$ principal direction (see Figure 3.2b, PC1 – PC4) are proportional to the $k^{th}$ eigenvector of the centered Gramian $G := JK(x_i, x_j) J$; see (3.56) and Corollary 3.18.

3.6. Dynamic PCA. In some applications, the tool of Dynamic PCA (DPCA), see e.g., [HA1+07] serves as a useful tool for high-dimensional and time-dependent data. The idea is that in favorable cases, with the use of DPCA, one can arrange that the input matrices can be augmented by addition of time-lagged values of the
DIMENSION REDUCTION AND KERNEL PRINCIPAL COMPONENT ANALYSIS

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 $256 \times 256$. (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.

This figure illustrates KPCA feature selection using Corollary 3.18. The Gaussian kernel is centered by (3.55)–(3.56). For each fixed sample point $x$, the vector $\Phi(x)$ in the Gaussian RKHS $\mathcal{H}(K)$ is projected onto the first 4 column vectors of the matrix $W$ as in (3.57). The resulting projection coefficients are given by $\lambda_1^{-1} \sum_{k=1}^n w_{k1} K(x_k, x)$, $\lambda_2^{-1} \sum_{k=1}^n w_{k2} K(x_k, x)$, $\lambda_3^{-1} \sum_{k=1}^n w_{k3} K(x_k, x)$ and $\lambda_4^{-1} \sum_{k=1}^n w_{k4} K(x_k, x)$.
variables under consideration. In summary, the method is based on the behavior of the eigenvalues of the lagged autocorrelation, and partial autocorrelation, matrices.

Our present Theorem 3.12 states that when a system of PCA eigenvalues \( \{ \lambda_i \} \) is given, then we have an algorithmic solution to the corresponding optimization question (see (3.34)-(3.35)). 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.

When principal component analysis (PCA) is used for problems involving stochastic (statistical) processes, for example in monitoring applications, it often relies on an implicit, but unrealistic, assumption that data involved are time independent. This however is evidently unrealistic, as work with industrial data shows: in a host of diverse applications, one is faced with serial correlations. Hence, the literature over recent years (see e.g., [AHR+20, KOL20, Sha20, GS20, Sha19]) has witnessed a host of new approaches to PCA, going by what is now called dynamic PCA (DPCA) methods. DPCA-models involve a variety of stochastic and dynamical systems-tools, each one serving as a remedy for plain-vanilla PCA applied to data as they appear in: (i) in extremely high-dimensions, (ii) in time-dependent data-sets, and (iii) in digital image processing. A main feature of DPCA is time-feedback; — hence in DPCA, the input-matrix is dynamically augmented, thus taking the form of a transfer operator;— in the simplest models, it operates by dynamic additions of time-lagged values of the variables.

**Proposition 3.21.** 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 \( \lambda_1^2 > \lambda_2^2 > \cdots > 0 \); \( \lambda_j^2 \to 0 \) as \( j \to \infty \). Setting \( S = \sqrt{T^*T} \), then,

\[
\lim_{n \to \infty} \frac{\|S^{n+1}x\|^2}{\|S^n x\|^2} = \lambda_1^2, \quad \forall x \text{ s.t. } \langle u_1, x \rangle \neq 0.
\]

Moreover, given \( \lambda_1 \), then

\[
\lim_{n \to \infty} \lambda_1^{-2n} (T^*T)^n = |u_1\rangle\langle u_1|,
\]

where convergence in (3.59) is w.r.t. the norm topology of \( B(\mathcal{H}) \).

**Proof.** Note that

\[
\lim_{n \to \infty} \frac{\|S^{n+1}x\|^2}{\|S^n x\|^2} = \lim_{n \to \infty} \frac{\langle x, S^{2n+2}x \rangle}{\langle x, S^{2n}x \rangle}
\]

\[
= \lim_{n \to \infty} \frac{\langle x, (T^*T)^{n+1} x \rangle}{\langle x, (T^*T)^n x \rangle}
\]

\[
= \lim_{n \to \infty} \frac{\sum \lambda_j^{2(n+1)} |\langle u_j, x \rangle|^2}{\sum \lambda_j^{2n} |\langle u_j, x \rangle|^2}
\]

\[
= \lim_{n \to \infty} \lambda_1^2 \frac{|\langle u_1, x \rangle|^2 + \sum_{j \geq 2} (\lambda_j/\lambda_1)^{2(n+1)} |\langle u_j, x \rangle|^2}{|\langle u_1, x \rangle|^2 + \sum_{j \geq 2} (\lambda_j/\lambda_1)^{2n} |\langle u_j, x \rangle|^2}
\]

\[
= \lambda_1^2.
\]
Now, given $\lambda_1$, we have
\[
\|\lambda_1^{-2n} (T^* T)^n - |u_1\rangle\langle u_1|\| = \left\| \sum_{j \geq 2} \left( \lambda_j / \lambda_1 \right)^{2n} |u_j\rangle\langle u_j| \right\| \\
\leq \sum_{j \geq 2} \left( \lambda_j / \lambda_1 \right)^{2n} \xrightarrow{n \to \infty} 0.
\]
\[\square\]

**Corollary 3.22.** The system of PCA eigenvalues $\{\lambda_j\}$ can be obtained inductively as follows: Let $S = \sqrt{T^* T}$ as before, and set $Q_k = \sum_{j=1}^k |u_j\rangle\langle u_j|$. Then
\[
\lambda_{k+1}^2 = \lim_{n \to \infty} \frac{\|(SQ_k^\dagger)^{n+1} x\|^2}{\|(SQ_k^\dagger)^n x\|}, \quad \forall x \text{ s.t. } \langle u_{k+1}, x \rangle \neq 0,
\]
and
\[
|u_{k+1}\rangle\langle u_{k+1}| = \lim_{n \to \infty} \lambda_k^{-2n} \left( Q_k^\dagger T^* T Q_k^\dagger \right)^n.
\]

**Proof.** One checks that
\[
SQ_k^\dagger = S (1 - Q_k) = \sum_{j=k+1}^\infty \lambda_j |u_j\rangle\langle u_j|,
\]
\[
(SQ_k^\dagger)^* (SQ_k^\dagger) = Q_k^\dagger T^* T Q_k^\dagger;
\]
and so the assertion follows from Proposition 3.21. \[\square\]

**Example 3.23.** 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
\[
E(B_s B_t) = s \land t = \min (s,t).
\]
(3.60)
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
\[
K_N = \begin{bmatrix}
  x_1 & x_1 & x_1 & \cdots & x_1 \\
  x_1 & x_2 & x_2 & \cdots & x_2 \\
  x_1 & x_2 & x_3 & \cdots & x_3 \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  x_1 & x_2 & x_3 & \cdots & x_N
\end{bmatrix} = (x_i \land x_j)_{i,j=1}^N.
\]
(3.61)

**Lemma 3.24.** Let $K_N$ be as in (3.61). Then

(i) The determinant of $K_N$ is given by
\[
\det(K_N) = x_1 (x_2 - x_1) (x_3 - x_2) \cdots (x_N - x_{N-1}).
\]
(3.62)

(ii) $K_N$ assumes the LU decomposition
\[
K_N = A_N A_N^*,
\]
(3.63)
where

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

(3.64)

Proof. For details, see e.g., [JT15].

Example 3.25. Fix \( N \), then the top eigenvalue of \( K := K_N \) can be extracted by the method from Proposition 3.21. 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

\[
\hat{\lambda}_1 = \frac{\|K^3 e_1\|}{\|K^2 e_1\|} \approx 5.0455, \quad \hat{v}_1 = \frac{\lambda_1^{-5} K^5 e_1}{\|\lambda_1^{-5} K^5 e_1\|} \approx (0.3284 \ 0.5913 \ 0.7366)^T.
\]

Standard numerical algorithm returns

\[
\lambda_1 \approx 5.0489, \quad v_1 \approx (0.3280 \ 0.5910 \ 0.7370)^T.
\]

4. REPRODUCING KERNELS, TRANSFORMS, AND GAUSSIAN PROCESSES

Our last section below deals with a number of applied transform-results. In more detail, they are infinite-dimensional and stochastic transforms. They are motivated to a large extent by our results (sect. 3.6 above) on dynamical PCA (DPCA). Moreover, our results below will combine two main themes in our paper, i.e., combining (i) adaptive kernel tricks (positive definite kernels, their Hilbert spaces, and choices of feature spaces) with (ii) a new analysis of corresponding classes of Gaussian processes/fields. This approach also builds on our parallel results on Monte Carlo simulation and use of Karhunen-Loève analysis. These new transforms, and their applications (see especially Corollary 4.5 and the subsequent results in sect. 4), are motivated directly by DPCA analysis, and they apply directly to design of DPCA algorithms.

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.

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 \( T \) is defined on an \( L^2 \) path-space \( L^2(\Omega, \mathbb{P}) \) of Brownian motion; and \( T \) 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 \( \mathbb{R} \times \mathbb{R} \);

\[
K(s, t) := e^{-\frac{1}{2}|s-t|}, \quad s, t \in \mathbb{R}.
\]  

(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 $\{e^{ix_t}\}_{t \in \mathbb{R}}$ 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 $(\Omega, \mathcal{F}, \mathbb{P})$, 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 $\mathbb{E}(|X_s - X_t|^2) = |s - t|$, \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,

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

and

$$
e^{-\frac{1}{2} |s-t|} = \mathbb{E}(e^{ix_s} e^{-ix_t}) = \langle e^{ix_s}, e^{ix_t} \rangle_{L^2(\mathbb{P})}, \forall s, t \in \mathbb{R}. \tag{4.7}
$$

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

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

where

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

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

$$
\sum_j \sum_k c_j c_k e^{-\frac{1}{2} |t_j - t_k|} = \sum_j \sum_k c_j c_k \mathbb{E}(e^{ix_{t_j} e^{-ix_{t_k}}}) = \mathbb{E}\left(|\sum_j c_j e^{ix_{t_j}}|^2\right) \geq 0. \tag{4.9}
$$
4.1. 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 $T : L^2(\mathbb{P}) \rightarrow \mathcal{H}(K)$,

$$
T(F)(t) := \mathbb{E}(e^{-iX_t}F) = \int_{\Omega} e^{-iX_t(\omega)}F(\omega)\,d\mathbb{P}(\omega)
$$

(4.11)

for all $F \in L^2(\mathbb{P})$. Here, $\{X_t\}$ is the Brownian motion.

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

Set

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

(4.12)

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

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

(4.13)

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

$$
L^2(\mathbb{P}) \xrightarrow{T} \mathcal{H}(K) \xleftarrow{T^*}
$$

(4.14)

Then, we have

$$
T(e^{iX_t}) = K_t, \text{ and } T^*(K_t) = e^{iX_t}.
$$

(4.15)

(4.16)

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

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

$$
T(e^{iX_t})(s) = \mathbb{E}(e^{-iX_s}e^{iX_t}) = \mathbb{E}(e^{i(X_t-X_s)}) = e^{-\frac{1}{2}|t-s|} = K_t(s).
$$

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

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

□

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

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

□

**Conclusion 4.6.** $\mathcal{H}(K)$ is naturally isometrically isomorphic to $L^2(\mathbb{P})$. 
Remark 4.7. To understand this isometric isomorphism \( L^2(\mathbb{P}) \xrightarrow{\sim} \mathcal{H}(K) \), 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 \( e^{ix_1} \in L^2(\mathbb{P}) \) 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 4.8. 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

\[
\mathbb{E} (e^{iX_t} e^{-iX_s}) = \mathbb{E} \left( e^{i(X_t - X_s)} \right) 
= \sum_{n=0}^{\infty} \frac{i^n}{n!} \mathbb{E} ((X_t - X_s)^n) 
= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} \mathbb{E} (X_t - X_s)^{2n} \quad \text{since the odd terms cancel}
\]

by (4.8)

\[
= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} (2n - 1)!! (t - s)^n
\]

by (4.9)

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

by (4.10)

\[
e^{-\frac{1}{2} (t - s)} = \mathbb{E} (e^{iX_{t-s}}).
\]

Proposition 4.9. For the Gaussian kernel \( K_{\text{Gauss}}(x, y) = e^{\frac{1}{2} (x - y)^2} \), we have

\[
e^{-x^2/2t} = \mathbb{E} (e^{ix_{1/t}}), \quad \text{and} \quad e^{-(x-y)^2/2t} = \mathbb{E} \left( e^{iX_{1/t} + e^{-iyX_{1/t}}} \right) = \langle e^{ix_{1/t}}, e^{iyX_{1/t}} \rangle_{L^2(\mathbb{P})}.
\]

Proof. A direct calculation yields

\[
\mathbb{E} (e^{ix_{1/t}}) = \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} \mathbb{E} (X_{1/t}^n) 
= \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} (2n - 1)!! \frac{1}{t^n}
\]

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

which is (4.17); and (4.18) follows from this.

Lemma 4.10. Assume \( 0 < s < t \). If \( F \in L^2(\mathbb{P}) \cap \mathcal{B}_s \), where \( \mathcal{B}_s = \sigma\)-algebra generated by \( \{X_u : u \leq s\} \), then

\[
\mathcal{T}(F)(t) = e^{-\frac{t-s}{2}} \mathcal{T}(F)(s).
\]

Proof.

\[
\text{LHS} (4.19) = \mathbb{E} (e^{-iX_t} F)
\]
\[ E \left( e^{-i(X_t - X_s)} e^{-iX_s} F \right) = E \left( e^{-i(X_t - X_s)} \right) E \left( e^{-iX_s} F \right) = e^{-iX_t/s^2} \mathcal{T}(F)(s) = \text{RHS}(4.19). \]

**Proposition 4.11.** Let \( 0 < s < t \), and let \( H_n(\cdot), n \in \mathbb{N}_0 \), be the Hermite polynomials; then
\[
\mathcal{T}(X^n_s)(t) = i^n e^{-\frac{t-s^2}{2}} H_n(\sqrt{s}) .
\]

**Proof.** By Lemma 4.10, we have
\[
\mathcal{T}(X^n_s)(t) = e^{-\frac{t-s^2}{2}} \mathcal{T}(X^n_s)(s)
\]
\[
= e^{-\frac{t-s^2}{2}} E(e^{-iX_s}X^n_s)
\]
\[
= e^{-\frac{t-s^2}{2}} i^n \left( \frac{d}{d\lambda} \right)^n |_{\lambda = 1} E(e^{-i\lambda X_s})
\]
\[
= e^{-\frac{t-s^2}{2}} i^n \left( \frac{d}{d\lambda} \right)^n |_{\lambda = 1} (e^{-\frac{\lambda^2}{2}})
\]
\[
= i^n e^{-\frac{t-s^2}{2}} e^{-\frac{s^2}{2}} H_n(\sqrt{s})
\]
\[
= i^n e^{-\frac{t-s^2}{2}} H_n(\sqrt{s})
\]
which is the RHS in (4.20). In the calculation above, we have used the following version of the Hermite polynomials \( H_n(\cdot) \), the probabilist’s variant; defined by
\[
\left( \frac{d}{d\xi} \right)^n e^{-\xi^2} = H_n(\xi) e^{-\xi^2}
\]
with the substitution \( \xi = \sqrt{s}\lambda \) for \( s > 0 \) fixed, and \( \lambda \to 1 \Leftrightarrow \xi \to \sqrt{s} \).

Note, the third “=” follows from standard calculation for the generating functions of Gaussian random variables (see e.g., [Hid80]):
\[
\left( \frac{d}{d\lambda} \right)^n E(e^{-i\lambda X_s}) = E((-iX_s)^n e^{-i\lambda X_s})
\]
\[
\downarrow
\]
\[
i^n \left( \frac{d}{d\lambda} \right)^n E(e^{-i\lambda X_s}) = E(X^n_s e^{-i\lambda X_s})
\]
and so
\[
i^n \left( \frac{d}{d\lambda} \right)^n |_{\lambda = 1} E(e^{-i\lambda X_s}) = E(X^n_s e^{-iX_s}).
\]

**Acknowledgement.** The co-authors thank the following colleagues for helpful and enlightening discussions: Professors Daniel Alpay, Sergii Bezuglyi, Ilwoo Cho, Wayne Polyzou, David Stewart, Eric S. Weber, and members in the Math Physics seminar at The University of Iowa. The present work was started with discussions at the NSF-CBMS Conference, Harmonic Analysis: Smooth and Non-Smooth, held at Iowa State University, June 4-8, 2018. Jorgensen was the main speaker. We are grateful to the organizers, especially to Prof Eric Weber, and to the NSF for...
financial support. Sooran Kang was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (#NRF-2017R1D1A1B03034697 and #NRF-2020R1F1A1A01076072). The authors thank an anonymous referee for making constructive suggestions.

REFERENCES

[ACP13] Ery Arias-Castro and Bruno Pelletier, On the convergence of maximum variance unfolding, J. Mach. Learn. Res. 14 (2013), 1747–1770. MR 3104494

[AFS18] Carlos M. Alai, Michael Fauvel, 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

[Agg18] C.C. Aggarwal, Machine learning for text, Springer International Publishing, 2018.

[AH11] Trine Julie Abrahansen and Lars Kai Hansen, A cure for variance inflation in high dimensional kernel principal component analysis, J. Mach. Learn. Res. 12 (2011), 2027–2044. MR 2819026

[AHR+20] Mohammad Reza Askari, Iman Hajizadeh, Mudassir Rashid, Nicole Hobbs, Victor M. Zavala, and Ali Cinar, Adaptive-learning model predictive control for complex physiological systems: automated insulin delivery in diabetes, Annu. Rev. Control 50 (2020), 1–12. MR 4188895

[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] Daniel Alpay, Palle Jorgensen, and Guy Salomon, 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 2305566

[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 3097010

[Bal19] 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] Christopher M. Bishop, 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

[BM03] Samuel Burer and Renato D. C. Monteiro, A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization, vol. 95, 2003, Computational semidefinite and second order cone programming: the state of the art, pp. 329–357. MR 1976484

[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

[CCE15] Jameson Cahill, Peter G. Casazza, and Martin Ehler, *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 neuromics*, SIAM/ASA J. Uncertain. Quantif. 7 (2019), no. 1, 117–142. MR 3900802

[CLW+19] Linjun Chen, Guangquan Lu, Guoqiu Wen, Jiaye Li, and Malong Tan, *A nonlinear clustering algorithm via kernel function and locality structure learning*, 2019 IEEE Symposium Series on Computational Intelligence (SSCI), 2019, pp. 2603–2607.

[Con90] John B. Conway, *A course in functional analysis*, second ed., Graduate Texts in Mathematics, vol. 96, Springer-Verlag, New York, 1990. MR 1070713

[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

[CST00] Nello Cristianini and John Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, 2000.

[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 Espírito 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)

[DM08] Paul Deheuvels and Guennadi V. Martynov, *A Karhunen-Loève decomposition of a Gaussian process generated by independent pairs of exponential random variables*, J. Funct. Anal. 255 (2008), no. 9, 2363–2394. MR 2473261

[DSR+21] Derek Driggs, Ian Selby, Michael Roberts, Effrossyni Gkrania-Klotsas, James H. F. Rudd, Guang Yang, Judith Babar, Evís Sala, and Carola-Bibiane Schönlieb, *Machine learning for covid-19 diagnosis and prognostication: Lessons for amplifying the signal while reducing the noise*, Radiology: Artificial Intelligence 3 (2021), no. 4, e210011.

[DWGC18] Jangh 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

[FB] Mathieu Fauvel, Jocelyn Chanussot, and Jon Atlie Benediktsson, *Kernel principal component analysis for feature reduction in hyperspectral images analysis*.

[F.R01] Karl Pearson F.R.S., *Liii. on lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (1901), no. 11, 559–572.

[FZ05] H. Frohlich and A. Zell, *Efficient parameter selection for support vector machines in classification and regression via model-based global optimization*, Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., vol. 3, 2005, pp. 1431–1436 vol. 3.

[GGB18] Akin 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
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.

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.

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

Bastian Gribisch and Michael Stollenwerk, *Dynamic principal component CAW models for high-dimensional realized covariance matrices*, Quant. Finance 20 (2020), no. 5, 799–821. MR 4085822

Shinpei Hara, Yuta Akira, Eisuke Ishii, Masato Inoue, and Masato Okada, *LDPC decoding dynamics from a PCA viewpoint*, Interdiscip. Inform. Sci. 13 (2007), no. 1, 43–48. MR 2314342

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

M. F. Harkat, A. Kouadri, R. Fezai, M. Mansouri, H. Nounou, and M. Nounou, *Machine learning-based reduced kernel pca model for nonlinear chemical process monitoring*, Journal of Control, Automation and Electrical Systems 31 (2020), no. 5, 1196–1209.

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

Jihun Ham, Daniel D. Lee, Sebastian Mika, and Bernhard Schölkopf, *A kernel view of the dimensionality reduction of manifolds*, Proceedings of the Twenty-First International Conference on Machine Learning (New York, NY, USA), ICML ’04, Association for Computing Machinery, 2004, p. 47.

A. Hogan, *The web of data*, Springer International Publishing, 2020.

H. Hotelling, *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology 24 (1933), 498–520.

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

Albert Jacquard, *The genetic structure of populations*, Biomathematics, Vol. 5, Springer-Verlag, New York-Heidelberg, 1974, Translated from the French by D. Charlesworth and B. Charlesworth, with additions and revisions, With a preface by R. C. Lewontin and a chapter by Daniel Courgeau. MR 0421724

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

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

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

Palle T. Jorgensen and Myung-Sin Song, *Graph Laplacians and discrete reproducing kernel Hilbert spaces from restrictions*, Stoch. Anal. Appl. 34 (2016), no. 4, 722–747. MR 3507188

Palle Jorgensen and Feng Tian, *Positive definite kernels and boundary spaces*, Adv. Oper. Theory 1 (2016), no. 1, 123–133. MR 3721329

Palle Jorgensen, *Non-commutative analysis*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2017, With a foreword by Wayne Polyzou. MR 3642406

Palle Jorgensen and Feng Tian, *Metric duality between positive definite kernels and boundary processes*, Int. J. Appl. Comput. Math. 4 (2018), no. 1, Art. 3, 13. MR 3736758

Palle Jorgensen and Feng Tian, *Realizations and factorizations of positive definite kernels*, Journal of Theoretical Probability (2018).
Palle Jorgensen and James Tian, *Infinite-dimensional analysis*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, [2021] ©2021, Operators in Hilbert space; stochastic calculus via representations, and duality theory. MR 4274591

Palle Jorgensen and James Tian, *Reproducing kernels and choices of associated feature spaces, in the form of l2-spaces*, J. Math. Anal. Appl. 505 (2022), no. 2, Paper No. 125535. MR 4295177

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.

Junhyeon Kwon, Hee-Seok Oh, and Yaeji Lim, *Dynamic principal component analysis with missing values*, J. Appl. Stat. 47 (2020), no. 11, 1957–1969. MR 4131352

George Kimeldorf and Grace Wahba, *Some results on Tchebycheff spline functions*, J. Math. Anal. Appl. 33 (1971), 82–95. MR 290013

Neil Lawrence, *Probabilistic non-linear principal component analysis with Gaussian process latent variable models*, J. Mach. Learn. Res. 6 (2005), 1783–1816. MR 2249872

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

Wen Li Liu, Shu Long Lu, 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

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

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

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

K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, *An introduction to kernel-based learning algorithms*, IEEE Transactions on Neural Networks 12 (2001), no. 2, 181–201.

Oleksandr Moklyachuk, *Simulation of random processes with known correlation function with the help of Karhunen-Loève decomposition*, Theory Stoch. Process. 13 (2007), no. 4, 163–169. MR 2482257

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

Horia Mania, Aadiyta 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

G.R. Naik, *Advances in principal component analysis: Research and development*, Springer Singapore, 2017.

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

J. R. Pycke, *On Three Families of Karhunen-Loève Expansions Associated with Classical Orthogonal Polynomials*, Results Math. 76 (2021), no. 3, 148. MR 4279486

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

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

Roman Rosipal and Leonid J Trejo, *Kernel partial least squares regression in reproducing kernel Hilbert space*, Journal of machine learning research 2 (2001), no. Dec, 97–123.
[SBXD06] Jun Sun, Stephen Boyd, Lin Xiao, and Persi Diaconis, The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem, SIAM Rev. 48 (2006), no. 4, 681–699. MR 2278445

[SF83] Stanley Sawyer and Joseph Felsenstein, Isolation by distance in a hierarchically clustered population, J. Appl. Probab. 20 (1983), no. 1–10. MR 688075

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

[Sha19] Han Lin Shang, Dynamic principal component regression: application to age-specific mortality forecasting, Astin Bull. 49 (2019), no. 3, 619–645. MR 4010088

[Sha20] Han Lin Shang, Dynamic principal component regression for forecasting functional time series in a group structure, Scand. Actuar. J. (2020), no. 4, 307–322. MR 4097183

[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

[SSM98] Bernhard Scholkopf, Alexander Smola, and Klaus-Robert Müller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, Neural Computation 10 (1998), no. 5, 1299–1319.

[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] Steve Smale and Ding-Xuan Zhou, Geometry on probability spaces, Constr. Approx. 30 (2009), no. 3, 311–323. MR 2558684

[SZ09b] Steve Smale and Ding-Xuan Zhou, 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

[UTK20] Nalika Ulapane, Karthick Thiyagarajan, and Sarath Kodagoda, Hyper-parameter initialization for squared exponential kernel-based gaussian process regression, 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2020, pp. 1154–1159.

[VdGLP19] Ning Wang, Shuangkui Ge, Baobin Li, and Lizhong Peng, Multiple description image compression based on multwavelets, Int. J. Wavelets Multiresolut. Inf. Process. 17 (2019), no. 1, 1850063, 22. MR 3911884

[VdGY10] Zaiwen Wen, Donald Goldfarb, and Wotao Yin, Alternating direction augmented Lagrangian methods for semidefinite programming, Math. Program. Comput. 2 (2010), no. 3-4, 203–230. MR 2741485

[WMd97] W. Wu, D.L. Massart, and S. de Jong, The kernel PCA algorithms for wide data. Part I: Theory and algorithms, Chemometrics and Intelligent Laboratory Systems 36 (1997), no. 2, 165–172.
[WSS04] Kilian Q. Weinberger, Fei Sha, and Lawrence K. Saul, *Learning a kernel matrix for nonlinear dimensionality reduction*, Proceedings of the Twenty-First International Conference on Machine Learning (New York, NY, USA), ICML ’04, Association for Computing Machinery, 2004, p. 106.

[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

[YO21] Gene Ryan Yoo and Houman Owhadi, *Deep regularization and direct training of the inner layers of Neural Networks with Kernel Flows*, Phys. D 426 (2021), Paper No. 132952. MR 4296249

[YTLM+18] Makoto Yamada, Jiliang Tang, Jose Lugo-Martinez, Ermin Hodzic, Raunak Shrestha, Avishek Saha, Hua Ouyang, Dawei Yin, Hiroshi Mamitsuka, Cenk Sahinalp, Predrag Radivojac, Filippo Menczer, and Yi Chang, *Ultra high-dimensional nonlinear feature selection for big biological data*, IEEE Transactions on Knowledge and Data Engineering 30 (2018), no. 7, 1352–1365.

[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

[ZSS19] Lan Zhang, Hongjun Su, and Jingwei Shen, *Hyperspectral dimensionality reduction based on multiscale superpixelwise kernel principal component analysis*, Remote Sensing 11 (2019), no. 10.

(PALLE E.T. JORGENSEN) DEPARTMENT OF MATHEMATICS, THE UNIVERSITY OF IOWA, IOWA CITY, IA 52242-1419, U.S.A.

Email address: palle-jorgensen@uiowa.edu

(SOORAN KANG) COLLEGE OF GENERAL EDUCATION, CHOONGANG UNIVERSITY, SEOUL, KOREA

Email address: sooran09@cau.ac.kr

(MYUNG-SIN SONG) DEPARTMENT OF MATHEMATICS AND STATISTICS, SOUTHERN ILLINOIS UNIVERSITY EDWARDSVILLE, EDWARDSVILLE, IL 62026, USA

Email address: msong@siue.edu

(JAMES F. TIAN) MATHEMATICAL REVIEWS, 416 4TH STREET ANN ARBOR, MI 48103-4816, U.S.A.

Email address: jft@ams.org