Eigen-Analysis of Nonlinear PCA with Polynomial Kernels

Zhiyu Liang, *The Ohio State University*
Yoonkyung Lee, *The Ohio State University*

Technical Report No. 872

October, 2012

Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247

Eigen-Analysis of Nonlinear PCA with Polynomial Kernels

Zhiyu Liang[*] Yoonkyung Lee[†]

Abstract

There has been growing interest in kernel methods for classification, clustering and dimension reduction. For example, kernel linear discriminant analysis, spectral clustering and kernel principal component analysis are widely used in statistical learning and data mining applications. The empirical success of the kernel method is generally attributed to nonlinear feature mapping induced by the kernel, which in turn determines a low dimensional data embedding. It is important to understand the effect of a kernel and its associated kernel parameter(s) on the embedding in relation to data distributions. In this paper, we examine the geometry of the nonlinear embedding for kernel PCA when polynomial kernels are used. We carry out eigen-analysis of the polynomial kernel operator associated with data distributions and investigate the effect of the degree of polynomial. The results provide both insights into the geometry of nonlinear data embedding and practical guidelines for choosing an appropriate degree for dimension reduction with polynomial kernels.

Keywords: Gaussian kernel, kernel methods, kernel PCA, nonlinear embedding, polynomial kernel

1 Introduction

Kernel methods have drawn great attention in machine learning and data mining in recent years (Schölkopf and Smola 2002, Hofmann et al. 2008). They are given as nonlinear generalization of linear methods by mapping data into a high dimensional feature space and applying the linear methods in the so-called feature space (Aizerman et al. 1964). Kernels are the functions that define the inner product of the feature vectors and play an important role in capturing nonlinear mapping desired for data analysis. Historically, they are closely related to reproducing kernels used in statistics for nonparametric function estimation; see Wahba (1990) for spline models. The explicit form of feature mapping is not required. Instead, specification of a kernel is sufficient for kernel methods. Application of the nonlinear generalization through kernels has led to various methods for classification, clustering and dimension reduction. Examples include support vector machines (SVM) (Schölkopf et al. 1999, Vapnik 1995), kernel linear discriminant analysis (kernel LDA) (Mika et al. 1999), spectral clustering (Scott and Longuet-Higgins 1990, von Luxburg 2007), and kernel principal component analysis (kernel PCA) (Schölkopf et al. 1998).

There have been many studies examining the effect of a kernel function and its associated parameters on the performance of kernel methods. For example, Brown et al. (2000), Ahn (2010) and Baudat and Anouar (2000) investigated how to select the bandwidth of Gaussian kernel for SVM and kernel LDA.

[*] Zhiyu Liang is Graduate Student, Department of Statistics, The Ohio State University, Columbus, OH 43210.

[†] Yoonkyung Lee is Associate Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210. Lee's research was supported in part by National Security Agency grant H98230-10-1-0202 and National Science Foundation grant DMS-12-09194.

In spectral clustering and kernel PCA, the kernel determines the projections or data embeddings to be used for uncovering clusters or for effective low dimensional representation of data, which are given as the leading eigenvectors of the kernel matrix. As kernel PCA regards the spectral analysis of a finite-dimensional kernel matrix, we can consider the eigen-analysis of the kernel operator as an infinite dimensional analogue, where eigenfunctions are viewed as a continuous version of the eigenvectors of the kernel matrix. Such eigen-analysis can provide a view point of the method at the population level. In general, it is important to understand the effect of a kernel on nonlinear data embedding in relation to data distributions. In this paper, we examine the geometry of the data embedding for kernel PCA.

Zhu et al. (1998), Williams and Seeger (2000) and Shi et al. (2009) studied the relation between Gaussian kernels and the eigenfunctions of the corresponding kernel operator under normal distributions. Zhu et al. (1998) computed the eigenvalues and eigenfunctions of the Gaussian kernel operator explicitly when data follow a univariate normal distribution. Williams and Seeger (2000) investigated how eigenvalues and eigenfunctions change depending on the input density function, and stated that the eigenfunctions with relatively large eigenvalues are useful in classification, in the context of approximating the kernel matrix using low rank eigen-expansion. Shi et al. (2009) extended the discussion for spectral clustering, explaining which eigenvectors to use for clustering when the distribution is a mixture of multiple components.

Among the kernel functions, Gaussian kernel and polynomial kernels are commonly used. Kaufmann (1999) discussed the application of polynomial kernels to handwritten digits recognition and checkerboard problem in the context of classification using support vector machines, which produced decent results. Extending the current studies of the Gaussian kernel operator, we carry out eigen-analysis of the polynomial kernel operator under various data distributions. In addition, we investigate the effect of the degree on the geometry of the nonlinear embedding with polynomial kernels.

In Section 2, we describe general results of eigen-analysis of the polynomial kernel operator defined through data distributions and show that the matrix of moments determines the eigenvalues and eigenfunctions. In Section 3, numerical examples are given to illustrate the relationship between the eigenvectors of a sample kernel matrix and the eigenfunctions from the theoretical analysis. We comment on the effect of degrees (especially even or odd) on data projections given by the leading eigenvectors, in relation to some features of the data distribution in the original input space. We also discuss how the eigenfunctions can explain some geometric patterns observed in data projections. In Section 4, we present kernel principal component analysis of the handwritten digit data from Le Cun et al. (1990) for some pairs of digits and explain the geometry of the embeddings of digit pairs through analysis of the sample moment matrices. Section 5 concludes this paper with discussions.

2 Eigen-Analysis of Polynomial Kernel Operator

In this section, we study the dependence of eigenfunctions and eigenvalues of the kernel operator on the data density distribution when the polynomial kernels are used.

2.1 Preliminaries

Suppose that data ($\mathcal{D} = \{x_1, \ldots, x_n\}$) consist of iid sample from a probability distribution P and the input domain for the data is $\mathcal{X}$. Then a kernel function K is defined as a semi-positive definite mapping from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$, and there is a unique function space $\mathcal{H}_K$ (called a reproducing kernel Hilbert space) corresponding to the kernel. See Aronszajn (1950) for kernel functions and reproducing kernel Hilber spaces in general.

Given a probability distribution P with density function $p(x)$ and a kernel function K, the distribution-dependent kernel operator is defined as

$$\mathcal{K}_p f(y) = \int_{\mathcal{X}} K(x, y) f(x) p(x) dx \tag{1}$$

as a mapping from $\mathcal{H}_K$ to $\mathcal{H}_K$. Then an eigenfunction $\phi \in \mathcal{H}_K$ and the corresponding eigenvalue λ for the operator $\mathcal{K}_p$ are defined through the equation

$$\mathcal{K}_p \phi = \lambda \phi \quad \text{or} \quad \int_{\mathcal{X}} K(x, y) \phi(x) p(x) dx = \lambda \phi(y). \tag{2}$$

Note that the eigenvalue and eigenfunction depend on both the kernel and probability distribution.

To see the connection between the kernel operator and kernel matrix as its sample version, consider the $n \times n$ kernel matrix, $K_n = [K(x_i, x_j)]$. Kernel PCA (Schölkopf et al. 1998) finds nonlinear data embeddings for dimension reduction through eigen analysis of the kernel matrix. Suppose that λ_n and $\mathbf{v} = (v_1, \ldots, v_n)^t$ are a pair of eigenvalue and eigenvector of K_n such that $K_n \mathbf{v} = \lambda_n \mathbf{v}$. Then for each $i = 1, 2, \ldots, n$, we have

$$\frac{1}{n} \sum_{j=1}^{n} K(x_i, x_j) v_j = \frac{\lambda_n}{n} v_i.$$

When $x_1, \ldots, x_n$ are sampled from the distribution with density $p(x)$ and $\mathbf{v}$ is considered as a discrete version of $\phi(\cdot)$ at data points, $(\phi(x_1), \ldots, \phi(x_n))^t$, we can see that the left-hand side of the above equation is an approximation to its integral counterpart:

$$\frac{1}{n} \sum_{j=1}^{n} K(x_i, x_j) \phi(x_j) \approx \int_{\mathcal{X}} K(x, x_i) \phi(x) p(x) dx.$$

As a result, λ_n / n can be viewed as an approximation to the eigenvalue λ of the kernel operator with eigenfunction ϕ. The pair of λ_n and $\mathbf{v}$ yield a nonlinear principal component or nonlinear embedding from $\mathcal{X}$ to $\mathbb{R}$ given by

$$\hat{\phi}(x) = \frac{1}{\lambda_n} \sum_{i=1}^{n} v_i K(x_i, x).$$

Hence, eigen-analysis of the kernel operator amounts to an infinite-dimensional analogue of kernel PCA. See Williams and Seeger (2000) and Shi et al. (2009) for more discussions about the connection.

We examine eigen-expansion of the polynomial kernel operator based on the equation (2) when $\mathcal{X} = \mathbb{R}^p$, and establish the dependence of the eigen-expansion on the data distribution. There are two types of polynomial kernels of degree d: i) $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^t \mathbf{y})^d$ and ii) $K^*(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^t \mathbf{y})^d$. We begin with eigen-analysis for the first type in two dimensional setting in Section 2.2 and generalize it to p-dimensional setting in Section 2.3. Then we extend the analysis further to the second type with an additional constant in Section 2.4.

2.2 Two-dimensional data

Suppose that data arise from two-dimensional setting, $\mathcal{X} = \mathbb{R}^2$ with the distribution with density $p(\mathbf{x})$. For polynomial kernel of degree d, $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^t\mathbf{y})^d$, we want to find λ and $\phi(\cdot)$ satisfying

$$\int_{\mathbb{R}^2} K(\mathbf{x},\mathbf{y})p(\mathbf{x})\phi(\mathbf{x})\, d\mathbf{x} = \lambda\phi(\mathbf{y})$$
$$i.e. \int_{\mathbb{R}^2} (\mathbf{x}^t\mathbf{y})^d p(\mathbf{x})\phi(\mathbf{x})\, d\mathbf{x} = \lambda\phi(\mathbf{y}).$$

Since $(\mathbf{x}^t\mathbf{y})^d = (x_1y_1 + x_2y_2)^d = \sum_{j=0}^{d}\binom{d}{j}(x_1y_1)^j(x_2y_2)^{d-j}$, the equation becomes

$$\int \left[\sum_{j=0}^{d}\binom{d}{j}(x_1y_1)^j(x_2y_2)^{d-j}\right] p(\mathbf{x})\phi(\mathbf{x})\, d\mathbf{x} = \lambda\phi(\mathbf{y}),$$

which is re-expressed as

$$\sum_{j=0}^{d}\binom{d}{j}y_1^j y_2^{d-j}\left[\int x_1^j x_2^{d-j} p(\mathbf{x})\phi(\mathbf{x})\, d\mathbf{x}\right] = \lambda\phi(\mathbf{y}).$$

Let $C_j = \int x_1^j x_2^{d-j} p(\mathbf{x})\phi(\mathbf{x})\, d\mathbf{x}$ be a distribution-dependent constant for $j = 0, \dots, d$. Then for $\lambda \neq 0$, the corresponding eigenfunction $\phi(\cdot)$ should be of the form

$$\phi(\mathbf{y}) = \frac{1}{\lambda}\sum_{k=0}^{d}\binom{d}{k}C_k y_1^k y_2^{d-k}. \tag{3}$$

By plugging $\phi(\mathbf{x})$ back into the defining equation for C_j, we get the following equations for the constants $(j = 0, \dots, d)$,

$$C_j = \frac{1}{\lambda}\int x_1^j x_2^{d-j} p(\mathbf{x})\sum_{k=1}^{d}\binom{d}{k}C_k x_1^k x_2^{d-k} d\mathbf{x},$$

which leads to

$$\lambda C_j = \sum_{k=1}^{d}\binom{d}{k}C_k \int x_1^{j+k} x_2^{2d-(j+k)} p(\mathbf{x}) d\mathbf{x}.$$

Note that $\int x_1^{j+k}x_2^{2d-(j+k)}p(\mathbf{x})d\mathbf{x}$ is $E(X_1^{j+k}X_2^{2d-(j+k)})$, a moment of the random vector $\mathbf{X} = (X_1, X_2)^t$ distributed with $p(\mathbf{x})$. Let $\mu_{j+k,2d-(j+k)}$ denote the moment. Then the set of the equations can be written as

$$\sum_{k=0}^{d}\binom{d}{k}\mu_{j+k,2d-(j+k)}C_k = \lambda C_j \quad \text{for } j = 0, \dots, d. \tag{4}$$

Defining the $(d + 1) \times (d + 1)$ matrix with entries given by moments of total degree $2d$ as follows

$$M_2^d = \begin{bmatrix} \binom{d}{0}\mu_{0,2d} & \binom{d}{1}\mu_{1,2d-1} & \cdots & \binom{d}{d}\mu_{d,d} \\ \binom{d}{0}\mu_{1,2d-1} & \binom{d}{1}\mu_{2,2d-2} & \cdots & \binom{d}{d}\mu_{d+1,d-1} \\ \vdots & \vdots & \ddots & \vdots \\ \binom{d}{0}\mu_{d,d} & \binom{d}{1}\mu_{d+1,d-1} & \cdots & \binom{d}{d}\mu_{2d,0} \end{bmatrix}, \tag{5}$$

we can succinctly express the set of equations as

$$M_2^d \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_d \end{bmatrix} = \lambda \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_d \end{bmatrix}. \tag{6}$$

For the moment matrix M_2^d, the subscript indicates the input dimension, and the superscript refers to the degree of the polynomial kernel. From the equation (6), we can see that the pairs of eigenvalue and eigenfunction for the polynomial kernel operator are determined by the spectral decomposition of the moment matrix M_2^d. However, the eigenvectors of M_2^d need to be scaled so that $\int \phi^2(\mathbf{x})p(\mathbf{x})\,d\mathbf{x} = 1$. Obviously, the determinant of $M_2^d - \lambda I$ is a polynomial of degree $(d+1)$. Therefore, there are at most $(d+1)$ nonzero eigenvalues of the polynomial kernel operator. The statements so far lead to the following theorem.

Theorem 1. *Suppose that the probability distribution $p(x_1, x_2)$ defined on $\mathbb{R}^2$ has finite $2d$th moments $\mu_{j,2d-j} = E(X_1^j X_2^{2d-j}), j = 0, \ldots, 2d$. For the polynomial kernel of degree d, $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^t\mathbf{y})^d$,*

(i) The eigenvalues of the polynomial kernel operator are given by the eigenvalues of the moment matrix

$$M_2^d = \begin{bmatrix} \binom{d}{0}\mu_{0,2d} & \binom{d}{1}\mu_{1,2d-1} & \cdots & \binom{d}{d}\mu_{d,d} \\ \binom{d}{0}\mu_{1,2d-1} & \binom{d}{1}\mu_{2,2d-2} & \cdots & \binom{d}{d}\mu_{d+1,d-1} \\ \vdots & \vdots & \ddots & \vdots \\ \binom{d}{0}\mu_{d,d} & \binom{d}{1}\mu_{d+1,d-1} & \cdots & \binom{d}{d}\mu_{2d,0} \end{bmatrix}.$$

(ii) There are at most $d+1$ nonzero eigenvalues.

(iii) The eigenfunctions are polynomials of total degree d of the form in (3) with coefficients determined by the eigenvectors of M_2^d.

2.3 p-dimensional data

In general, consider the p-dimensional input space ($\mathcal{X} = \mathbb{R}^p$) for data. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, the kernel function can be expanded as

$$(\mathbf{x}^t\mathbf{y})^d = (\sum_{i=1}^p x_i y_i)^d = \sum_{j_1+\cdots+j_p=d} \binom{d}{j_1, \cdots, j_p} \prod_{k=1}^p (x_k y_k)^{j_k},$$

and the equation (2) becomes

$$\int \sum_{j_1+\cdots+j_p=d} \binom{d}{j_1, \cdots, j_p} \prod_{k=1}^p (x_k y_k)^{j_k} p(\mathbf{x})\phi(\mathbf{x})\,d\mathbf{x} = \lambda\phi(\mathbf{y})$$

$$i.e. \sum_{j_1+\cdots+j_p=d} \binom{d}{j_1, \cdots, j_p} \prod_{k=1}^p y_k^{j_k} \left[\int \prod_{k=1}^p x_k^{j_k} p(\mathbf{x})\phi(\mathbf{x})\,d\mathbf{x}\right] = \lambda\phi(\mathbf{y}).$$

Letting $C_{j_1,\cdots,j_p} = \int \prod_{k=1}^p x_k^{j_k} p(\mathbf{x})\phi(\mathbf{x})\,d\mathbf{x}$, we can write the eigenfunction $\phi(\cdot)$ as

$$\phi(\mathbf{x}) = \frac{1}{\lambda} \sum_{j_1+\cdots+j_p=d} \binom{d}{j_1, \cdots, j_p} C_{j_1,\cdots,j_p} \prod_{k=1}^p x_k^{j_k}. \tag{7}$$

Again, by plugging this expansion $\phi(\mathbf{x})$ in the equation that defines $C_{j_1,\cdots,j_p}$, we get a set of equations for the constants:

$$C_{j_1,\cdots,j_p} = \frac{1}{\lambda}\int \prod_{k=1}^{p} x_k^{j_k} p(\mathbf{x}) \left[\sum_{i_1+\cdots+i_p=d} \binom{d}{i_1,\cdots,i_p} C_{i_1,\cdots,i_p} \prod_{k=1}^{p} x_k^{i_k} \right] d\mathbf{x},$$

which is rewritten as

$$\lambda C_{j_1,\cdots,j_p} = \sum_{i_1+\cdots+i_p=d} \binom{d}{i_1,\cdots,i_p} C_{i_1,\cdots,i_p} \int \prod_{k=1}^{p} x_k^{i_k+j_k} p(\mathbf{x})\, d\mathbf{x}.$$

Let $\mu_{j_1+i_1,\cdots,j_p+i_p}$ denote the moment $E(\prod_{k=1}^{p} X_k^{j_k+i_k}) = \int \prod_{k=1}^{p} x_k^{i_k+j_k} p(\mathbf{x})\, d\mathbf{x}$ for $(i_1,\ldots,i_p)$ with $i_1+\cdots+i_p = d$ and $(j_1,\ldots,j_p)$ with $j_1+\cdots+j_p = d$. Then we have

$$\sum_{i_1+\cdots+i_p=d} \binom{d}{i_1,\cdots,i_p} \mu_{j_1+i_1,\cdots,j_p+i_p} C_{i_1,\cdots,i_p} = \lambda C_{j_1,\cdots,j_p}. \tag{8}$$

To express the above equation in matrix form, we generalize the moment matrix M_2^d to M_p^d with entries given by $\binom{d}{i_1,\cdots,i_p}\mu_{j_1+i_1,\cdots,j_p+i_p}$. The dimension of M_p^d is the number of combinations of non-negative integers j_k's satisfying $j_1+\cdots+j_p = d$, which is $d_p = \binom{d+p-1}{d}$. Then the equation (8) is written as

$$M_p^d C = \lambda C,$$

where $C = \left[C_{j_1,\cdots,j_p}, j_1+\cdots+j_p = d\right]^t$. Applying the similar argument used for two-dimensional data, we conclude that there are at most $d_p = \binom{d+p-1}{d}$ nonzero eigenvalues of the polynomial kernel operator, and d_p depends on both the input dimension and the degree of the polynomial kernel. Thus we arrive at the following theorem.

Theorem 2. *Suppose that the probability distribution $p(x_1, x_2, \cdots, x_p)$ defined on $\mathbb{R}^p$ has finite $2d$th moments, $\mu_{i_1+j_1,\cdots,i_p+j_p} = E(\prod_{k=1}^{p} X_k^{i_k+j_k})$ for $j_1+\cdots+j_p = d, i_1+\cdots+i_p = d$. For the polynomial kernel of degree d, $K(\mathbf{x},\mathbf{y}) = (\mathbf{x}^t\mathbf{y})^d$,*

(i) The eigenvalues of the polynomial kernel operator are given by the eigenvalues of the moment matrix M_p^d.

(ii) There are at most $d_p = \binom{d+p-1}{d}$ nonzero eigenvalues.

(iii) The eigenfunctions are polynomials of total degree d of the form in (7) *with coefficients given by the eigenvectors of M_p^d.*

2.4 Polynomial kernel with constant

The kernel operator for the second type of polynomial kernel with constant can be treated as a special case of what we have discussed in the previous section.

For example, $K^*(\mathbf{x},\mathbf{y}) = (1 + x_1y_1 + x_2y_2)^2$ in the two-dimensional setting can be viewed as $K(\mathbf{x},\mathbf{y}) = (x_1y_1 + x_2y_2 + x_3y_3)^2$ in the three-dimensional setting with $x_3 = y_3 = 1$. Using the connection between K^* and K, we know that the number of nonzero eigenvalues for the kernel operator with K^* is at most $\binom{2+3-1}{2} = \binom{4}{2} = 6$ from the theorem in Section 2.3. The eigenfunctions in this case are of the following form,

$$\phi(\mathbf{x}) = \frac{1}{\lambda} \sum_{j_1+j_2+j_3=2} \binom{2}{j_1, j_2, j_3} C_{j_1,j_2,j_3} x_1^{j_1} x_2^{j_2}. \tag{9}$$

There are six combinations of non-negative integers j_k's which satisfy $j_1 + j_2 + j_3 = 2$. M_3^2 in general is given as follows:

$$M_3^2 = \begin{bmatrix} \mu_{0,0,4} & 2\mu_{0,1,3} & \mu_{0,2,2} & 2\mu_{1,0,3} & 2\mu_{1,1,2} & \mu_{2,0,2} \\ \mu_{0,1,3} & 2\mu_{0,2,2} & \mu_{0,3,1} & 2\mu_{1,1,2} & 2\mu_{1,2,1} & \mu_{2,1,1} \\ \mu_{0,2,2} & 2\mu_{0,3,1} & \mu_{0,4,0} & 2\mu_{1,2,1} & 2\mu_{1,3,0} & \mu_{2,2,0} \\ \mu_{1,0,3} & 2\mu_{1,1,2} & \mu_{1,2,1} & 2\mu_{2,0,2} & 2\mu_{2,1,1} & \mu_{3,0,1} \\ \mu_{1,1,2} & 2\mu_{1,2,1} & \mu_{1,3,0} & 2\mu_{2,1,1} & 2\mu_{2,2,0} & \mu_{3,1,0} \\ \mu_{2,0,2} & 2\mu_{2,1,1} & \mu_{2,2,0} & 2\mu_{3,0,1} & 2\mu_{3,1,0} & \mu_{4,0,0} \end{bmatrix},$$

and the vector C with constants C_{j_1,j_2,j_3} satisfies the following equation

$$M_3^2 \begin{bmatrix} C_{0,0,2} \\ C_{0,1,1} \\ C_{0,2,0} \\ C_{1,0,1} \\ C_{1,1,0} \\ C_{2,0,0} \end{bmatrix} = \lambda \begin{bmatrix} C_{0,0,2} \\ C_{0,1,1} \\ C_{0,2,0} \\ C_{1,0,1} \\ C_{1,1,0} \\ C_{2,0,0} \end{bmatrix}.$$

Since $X_3 = 1$, the moments $\mu_{i_1+j_1,i_2+j_2,i_3+j_3} = E(\prod_{k=1}^{3} X_k^{i_k+j_k})$ are simplified to $\mu^*_{i_1+j_1,i_2+j_2} = E(\prod_{k=1}^{2} X_k^{i_k+j_k})$.

In summary, we conclude that for the data distribution in $\mathbb{R}^p$ and polynomial kernel of degree d with constant term, the resulting eigenvalues and eigenfunctions of the kernel operator can be obtained on the basis of Theorem 2. The extensions are accomplished by application of the result with polynomial kernel of degree d for data distribution in $\mathbb{R}^{p+1}$ with X_{p+1} fixed at 1, where the moments $\mu_{i_1+j_1,\cdots,i_{p+1}+j_{p+1}} = E(\prod_{k=1}^{p+1} X_k^{i_k+j_k})$ reduce to $\mu^*_{i_1+j_1,\cdots,i_p+j_p} = E(\prod_{k=1}^{p} X_k^{i_k+j_k})$.

3 Simulation Studies

We present simulation studies to illustrate the relationship between the theoretical eigenfunctions and sample eigenvectors for kernel PCA. First we consider two simulation settings in $\mathbb{R}^2$ and examine the explicit forms of the eigenfunctions using Theorem 1. With an additional example, we investigate the effect of degree (the parameter for polynomial kernels) on the nonlinear data embeddings induced by the kernel, which can be used for uncovering data clusters or discriminating different classes. Furthermore, we explore how eigenfunctions can be used to understand certain geometric patterns observed in data projections.

3.1 Uniform example

For $\mathbf{X} = (X_1, X_2)^t$, let X_1 and X_2 be iid with uniform distribution on $(0, 1)$. Suppose that we use the second-order polynomial kernel, $K(\mathbf{x}, \mathbf{y}) = (x_1y_1 + x_2y_2)^2$. Since all the fourth moments

$\mu_{j,4-j} = E(X_1^j X_2^{4-j})$, $j = 0, \ldots, 4$ are finite in this case, we can compute the theoretical moment matrix M_2^2 explicitly, and it is given by

$$M_2^2 = \begin{bmatrix} \mu_{0,4} & 2\mu_{1,3} & \mu_{2,2} \\ \mu_{1,3} & 2\mu_{2,2} & \mu_{3,1} \\ \mu_{2,2} & 2\mu_{3,1} & \mu_{4,0} \end{bmatrix} = \begin{bmatrix} \frac{1}{5} & \frac{1}{4} & \frac{1}{9} \\ \frac{1}{8} & \frac{2}{9} & \frac{1}{8} \\ \frac{1}{9} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}.$$

Notice that there is symmetry in the moments due to the exchangeability of X_1 and X_2 (e.g. $\mu_{1,3} = \mu_{3,1}$).

The eigenvalues of the kernel operator are the same as the eigenvalues of M_2^2. We can get the eigenvalues of the matrix numerically, and they are given by $\lambda_1 = 0.5206$, $\lambda_2 = 0.0889$, and $\lambda_3 = 0.0127$. According to Theorem 1, given each eigenvalue λ, the corresponding eigenfunction can be written explicitly in the form,

$$\phi(\mathbf{x}) = \frac{1}{\lambda}\left(C_2 x_1^2 + 2C_1 x_1 x_2 + C_0 x_2^2\right),$$

where $(C_0, C_1, C_2)^t$ is a scaled version of the eigenvector of M_2^2 corresponding to λ. For simplicity of exposition, we choose not to scale the eigenfunctions to the unit norm but to go with the scale given by the eigenvectors throughout our numerical studies. With the unit-normed eigenvectors, we have the following eigenfunctions for the uniform distribution:

$$\begin{aligned} \phi_1(\mathbf{x}) &= \frac{1}{0.5206}(-0.608x_1^2 - 1.019x_1x_2 - 0.608x_2^2), \\ \phi_2(\mathbf{x}) &= \frac{1}{0.0889}(0.707x_1^2 - 0.707x_2^2), \\ \phi_3(\mathbf{x}) &= \frac{1}{0.0127}(0.540x_1^2 - 1.290x_1x_2 + 0.540x_2^2). \end{aligned}$$

To make numerical comparisons, we took a sample of size 400 from the distribution and computed the sample kernel matrix for the second-order polynomial kernel. Then we obtained its eigenvalues and corresponding eigenvectors. There are three non-zero eigenvalues, and they are $\hat{\lambda}_1 = 0.5118$, $\hat{\lambda}_2 = 0.0913$, and $\hat{\lambda}_3 = 0.0138$ after being scaled by the sample size n as discussed in Section 2. The sample eigenvalues are quite close to the theoretical ones.

Figure 1 compares the contour plots of the nonlinear embeddings given by the sample eigenvectors and the theoretical eigenfunctions. The top panels are for the embeddings induced by the leading eigenvectors of the kernel matrix, while the bottom panels are for the theoretical eigenfunctions obtained from the moment matrix. The change in color from red to yellow in each panel indicates increase in values. There is great similarity between the contours of the true eigenfunction and its sample version through eigenvector in terms of the shape and the gradient indicated by the color change. We also observe in Figure 1 that the nonlinear embeddings given by the first two leading eigenvectors and eigenfunctions of the second-order polynomial kernel are roughly along the two diagonal lines of the unit square $(0,1)^2$, which correspond to the directions of the largest variation in the uniform distribution.

3.2 Mixture normal example

We turn to a mixture of normal distributions for $(X_1, X_2)^t$. Suppose that X_1 and X_2 are two independent variables distributed with the following mixture Gaussian distribution:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim 0.5\ N\left(\begin{pmatrix} -3 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) + 0.5\ N\left(\begin{pmatrix} 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

For this example, we consider the polynomial kernels of degrees 2 and 3.

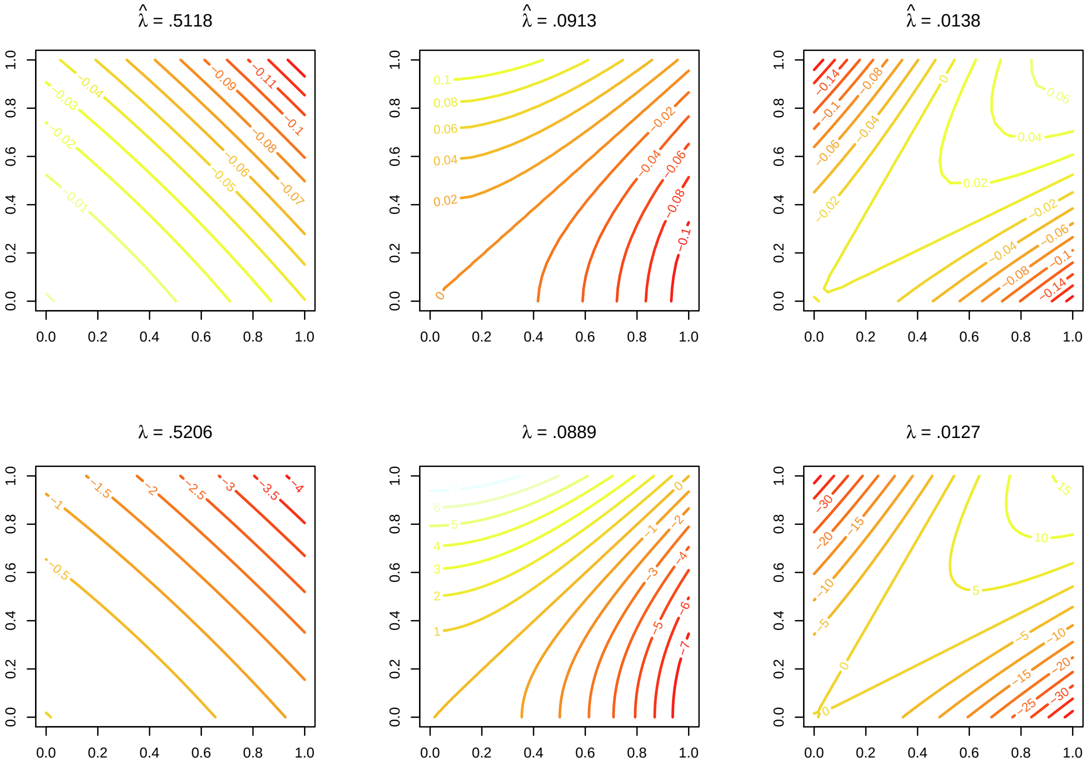


Figure 1: Comparison of the contours of the nonlinear embeddings given by three leading eigenvectors and the theoretical eigenfunctions for the uniform data. The upper three panels are for the embeddings induced by the eigenvectors for three nonzero eigenvalues, and the lower three panels are for the corresponding eigenfunctions.

When degree is 2

The moment matrix for the mixture distribution can be obtained as follows:

$$M_2^2 = \begin{bmatrix} \mu_{0,4} & 2\mu_{1,3} & \mu_{2,2} \\ \mu_{1,3} & 2\mu_{2,2} & \mu_{3,1} \\ \mu_{2,2} & 2\mu_{3,1} & \mu_{4,0} \end{bmatrix} = \begin{bmatrix} 10 & -20 & 15 \\ -10 & 30 & -25 \\ 15 & -50 & 90.5 \end{bmatrix}.$$

Three nonzero eigenvalues of the matrix are $\lambda_1 = 110.593$, $\lambda_2 = 17.415$, and $\lambda_3 = 2.492$. With the corresponding eigenvectors of the moment matrix, we get the following three eigenfunctions:

$$\phi_1(\mathbf{x}) = \frac{1}{110.593}(-0.928x_1^2 + 0.626x_1x_2 - 0.201x_2^2),$$
$$\phi_2(\mathbf{x}) = \frac{1}{17.415}(0.54x_1^2 + 1.234x_1x_2 - 0.573x_2^2),$$
$$\phi_3(\mathbf{x}) = \frac{1}{2.492}(0.068x_1^2 + 0.79x_1x_2 + 0.916x_2^2).$$

Contours of these eigenfunctions are displayed in the bottom panels of Figure 2.

For their sample counterparts, we generated a random sample of size 400 from the mixture of two normals. A scatter plot of the sample is displayed in the top left panel of Figure 4. Three nonzero eigenvalues from the kernel matrix are found to be $\hat{\lambda}_1 = 104.985$, $\hat{\lambda}_2 = 20.437$, and $\hat{\lambda}_3 = 3.180$. The top panels of Figure 2 show the contours of the data embeddings given by the corresponding eigenvectors. The data embeddings and eigenfunctions for this mixture normal example also exhibit strong similarity. The contours of the leading embedding and eigenfunction are ellipses centered at the origin. It appears that the minor axis of the ellipses for the leading eigenfuncion corresponds to the line connecting the two mean vectors of the mixture distribution, capturing the largest data variation, and the major axis is perpendicular to the mean difference. The contours of the second leading eigenfunction are hyperbolas centered at the origin. The asymptotes of the hyperbolas for the eigenfunction are the same as the major and minor axes for the leading eigenfunction. Although the symmetry that the data embeddings and eigenfunctions exhibit reflects that of the underlying distribution, information about the two normal components is lost after projection. If dimension reduction is to be used primarily for identifying the clusters later, then the quadratic kernel would not be useful in this case.

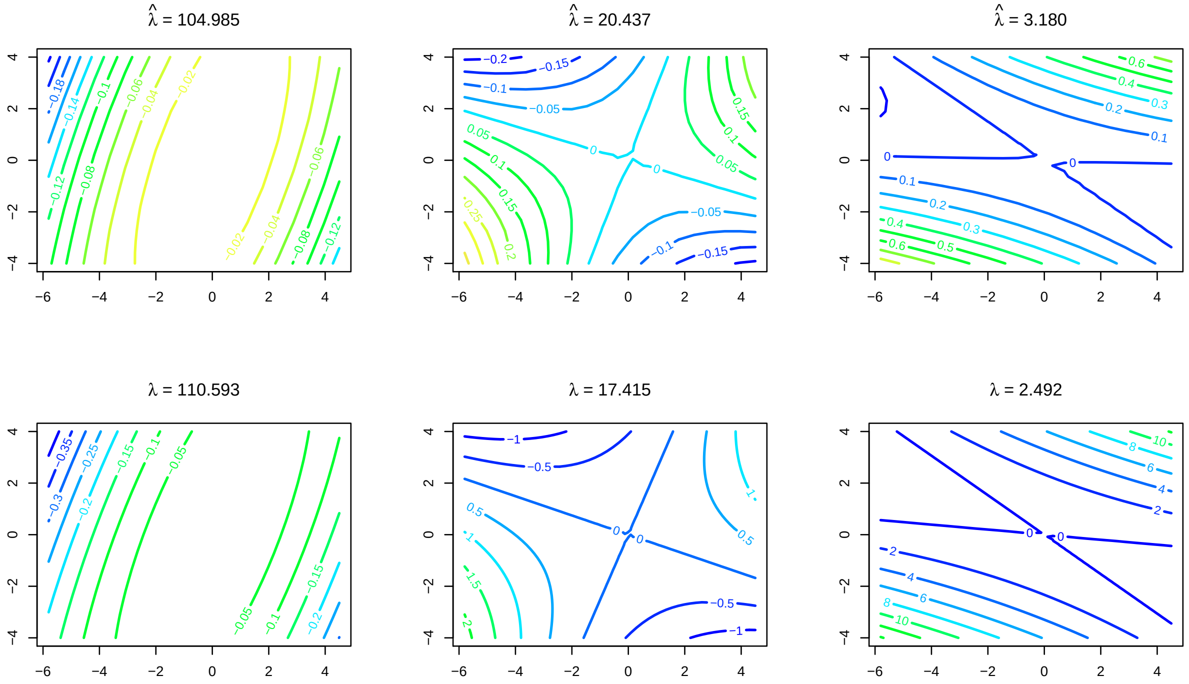


Figure 2: Comparison of the contours of the nonlinear embeddings given by three leading eigenvectors (top panels) and the theoretical eigenfunctions (bottom panels) for the mixture normal data when degree is 2.

When degree is 3

The moment matrix for $d = 3$ involves the moments up to order 6, and for the mixture distribution, it is explicitly given by

$$M_3^2 = \begin{bmatrix} \mu_{0,6} & 3\mu_{1,5} & 3\mu_{2,4} & \mu_{3,3} \\ \mu_{1,5} & 3\mu_{2,4} & 3\mu_{3,3} & \mu_{4,2} \\ \mu_{2,4} & 3\mu_{3,3} & 3\mu_{4,2} & \mu_{5,1} \\ \mu_{3,3} & 3\mu_{4,2} & 3\mu_{5,1} & \mu_{6,0} \end{bmatrix} = \begin{bmatrix} 76 & -195 & 225 & -100 \\ -65 & 225 & -300 & 181 \\ 75 & -300 & 543 & -350 \\ -100 & 543 & -1050 & 1431.5 \end{bmatrix}.$$

The matrix has four nonzero eigenvalues, $\lambda_1 = 1862.615$, $\lambda_2 = 343.748$, $\lambda_3 = 59.266$, and $\lambda_4 = 9.870$, and the corresponding eigenfunctions are

$$\begin{aligned}
\phi_1(\mathbf{x}) &= \frac{1}{1862.615}(-0.937x_1^3 + 0.873x_1^2x_2 - 0.483x_1x_2^2 + 0.107x_2^3),\\
\phi_2(\mathbf{x}) &= \frac{1}{343.748}(-0.681x_1^3 + 1.41x_1^2x_2 - 1.134x_1x_2^2 + 0.416x_2^3),\\
\phi_3(\mathbf{x}) &= \frac{1}{59.266}(0.151x_1^3 + 1.455x_1^2x_2 + 1.254x_1x_2^2 - 0.753x_2^3),\\
\phi_4(\mathbf{x}) &= \frac{1}{53.388}(0.03x_1^3 - 0.288x_1^2x_2 - 1.293x_1x_2^2 - 0.897x_2^3).
\end{aligned}$$

We obtained the kernel matrix with the polynomial kernel of degree 3 for the same data as in $d = 2$ case. Four leading eigenvalues for this kernel matrix are $\hat{\lambda}_1 = 1665.553$, $\hat{\lambda}_2 = 417.996$, $\hat{\lambda}_3 = 81.182$, and $\hat{\lambda}_4 = 11.536$.

Figure 3 shows the contours of the projections given by the sample eigenvectors of the kernel matrix and their theoretical counterparts when degree is 3. As in the case of degree 2, the leading eigenfunction and data embedding capture the largest variance along the direction of the difference between the two normal means. However, in contrast with degree 2, their contours show a monotone change along the direction, which allows identification of the two normal components if classification or clustering is concerned.

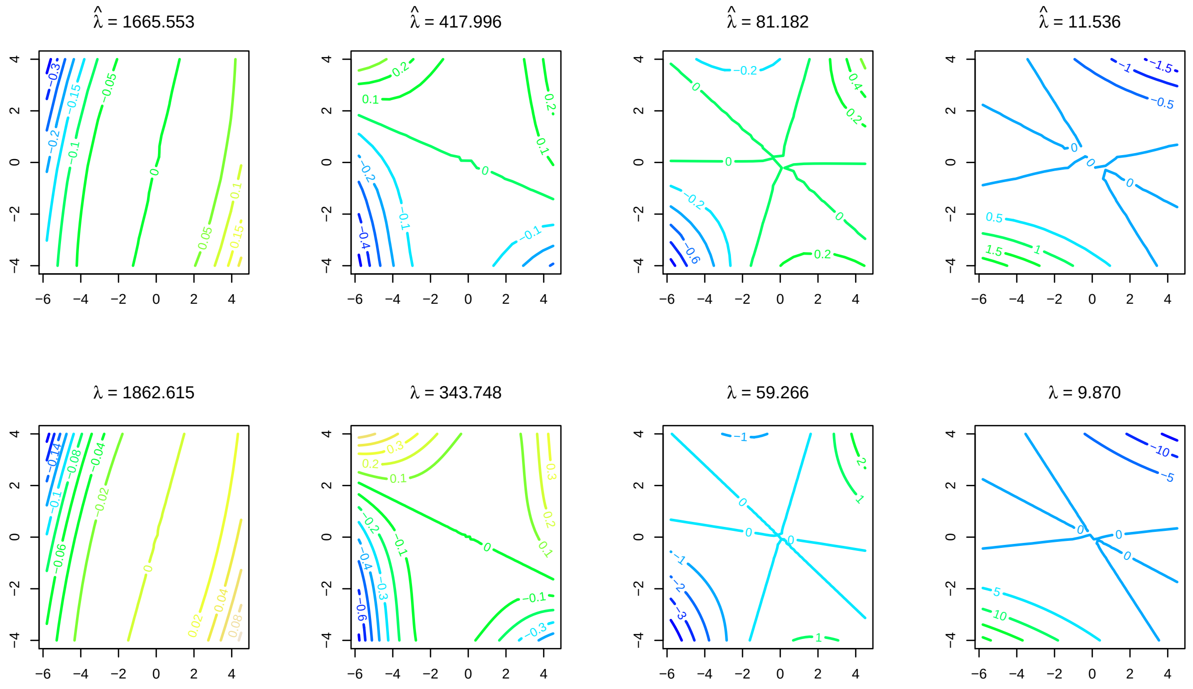


Figure 3: Comparison of the contours of the nonlinear embeddings given by four leading eigenvectors (top panels) and the theoretical eigenfunctions (bottom panels) for the mixture normal data when degree is 3.

3.3 Effect of degree

Figure 4 shows a scatter plot of the mixture normal data used in the example discussed above along with their projections through the first two principal components of kernel PCA with polynomial kernel from degree 1 to 5. As observed in the analysis of degrees 2 and 3 and illustrated in Figures 2 and 3, the degree of polynomial kernel has different effect on the projections. Odd-degree polynomial kernels provide projections that can separate the two components (or classes), while even-degree polynomial kernels project the two clusters into the same region, overlaying them on top of each other. The fact that odd degrees work for the mixture normal example while even degrees mask the difference between the clusters is closely tied to the underlying data structure.

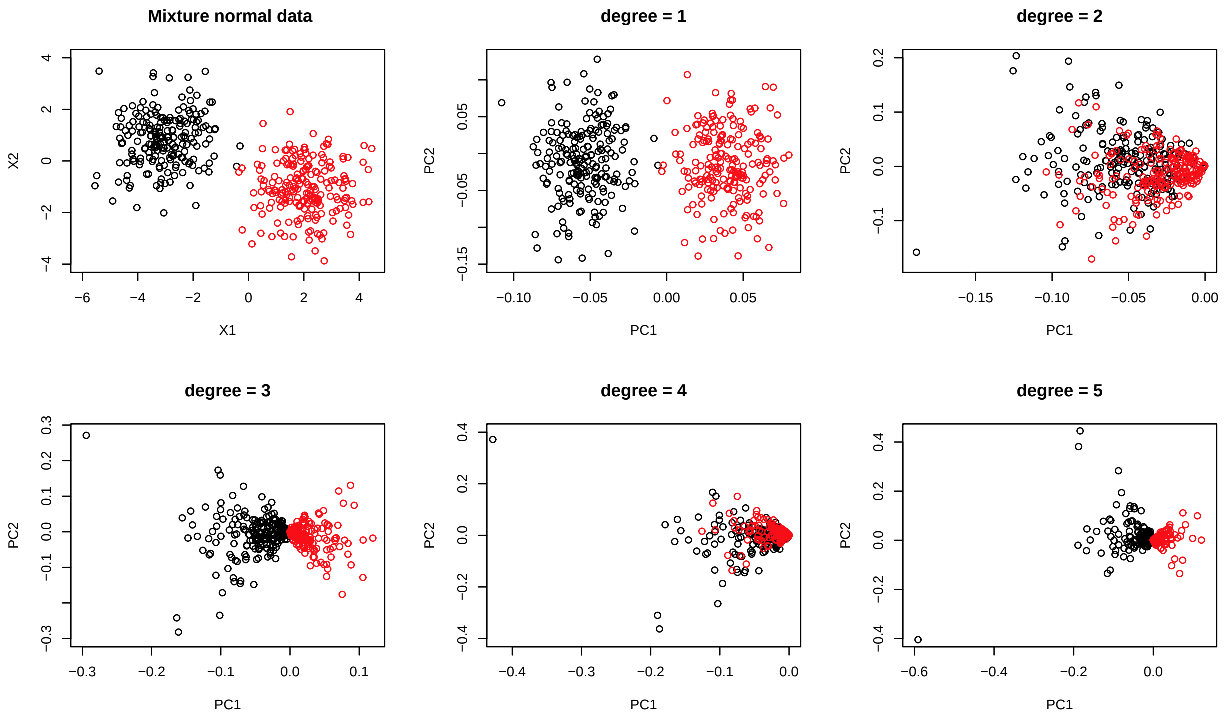


Figure 4: The mixture normal data and their projections through principal components with polynomial kernel of varying degrees. The colors distinguish the two normal components.

Figure 5 shows a different scenario where polynomial kernels of even degree would work better than those of odd degree. The scatter plot of “wheel” data in the top left panel shows two clusters symmetric around the origin. Separation of the clusters requires a nonlinear mapping. Due to the “even” nature of clustering, the data projections for even-degree polynomial kernels (especially degree 2 in the figure) tend to make the clusters linearly separable.

Besides the difference in effect between even and odd degrees, we also observe in Figures 4 and 5 that as the degree of polynomial kernel increases, projections tend to be heavily influenced by outliers. Increase in degree creates the effect of squeezing most points increasingly more in the projection space while spreading the outliers further apart. The observed sensitivity of the data embeddings to outliers for kernel PCA cautions against the choice of high degrees, in spite of the conventional belief that as degree goes higher, kernel methods become more flexible. Figure 5 clearly suggests that lower even degree ($d = 2$) works better than higher even degree ($d = 6$) in providing separation of clusters for the example.

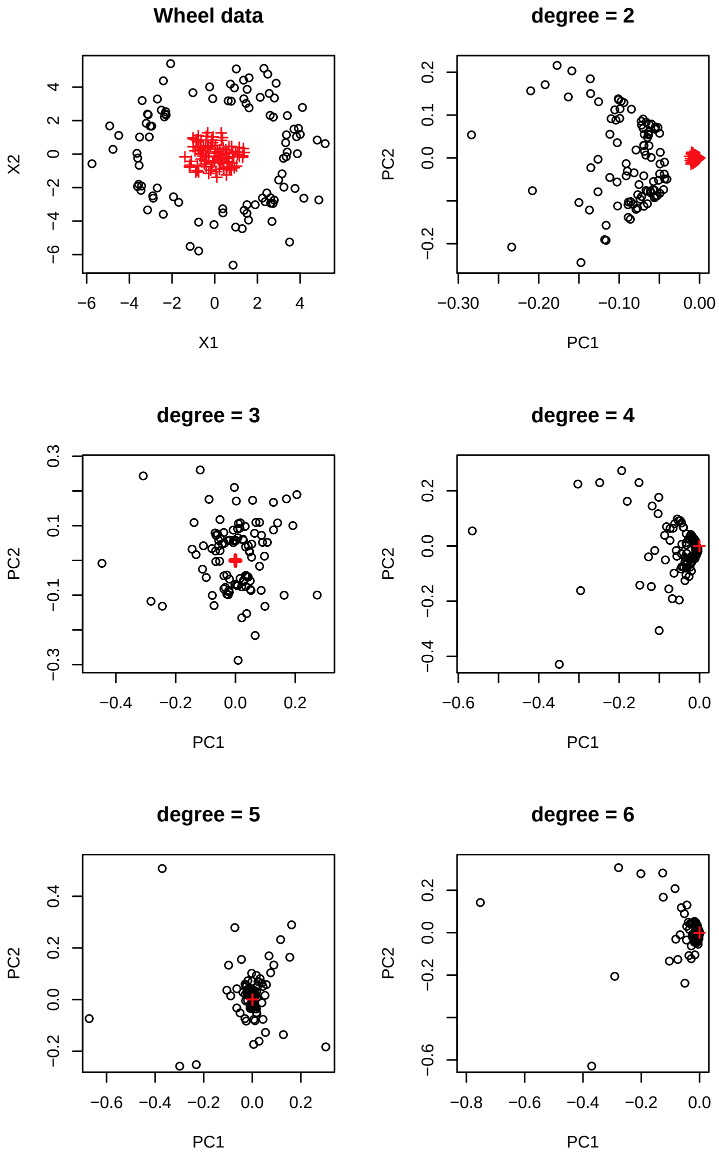


Figure 5: "Wheel" data and their projections through principal components with polynomial kernel of varying degrees. The colors distinguish the two clusters.

3.4 Restriction in embeddings

Figures 4 and 5 also suggest that there are some geometric restrictions in the pairs of nonlinear data embeddings given by the leading eigenvectors of kernel matrix. For example, when degree is 2, the data projections derived from kernel PCA exhibit a conic pattern in both examples. We explain the intrinsic restriction in the nonlinear data embeddings using their corresponding eigenfunctions.

For simplicity, suppose that the leading eigenfunctions are of the form, $\phi_1(\mathbf{x}) = c_1x_1^2 + c_2x_2^2$ and $\phi_2(\mathbf{x}) = c_{12}x_1x_2$, yielding ellipses and hyperbolas as their contours as in the mixture normal data. We can examine the relationship between the two by considering a level set of the first eigenfunction and the values of the second eigenfunction corresponding to the level set. When

$\phi_1(\mathbf{x}) = c_1 x_1^2 + c_2 x_2^2 = k$,

$$\phi_2^2(\mathbf{x}) = c_{12}^2 x_1^2 x_2^2 = c_{12}^2 x_1^2 \left[\frac{k - c_1 x_1^2}{c_2}\right] = -\left(\frac{c_{12}^2 c_1}{c_2}\right) x_1^4 + \left(\frac{c_{12}^2 k}{c_2}\right) x_1^2.$$

It is easy to see that $\phi_2^2(\mathbf{x})$ is bounded above by $c_{12}^2 k^2/(4c_1c_2)$, provided that $c_1c_2 > 0$. Hence, there is restriction in the range of $\phi_2(\mathbf{x})$ given $\phi_1(\mathbf{x}) = k$. To give a concrete example, let $\phi_1(\mathbf{x}) = -0.1x_1^2 - 0.02x_2^2$ and $\phi_2(\mathbf{x}) = -0.5x_1x_2$. Figure 6 shows the possible range for $\phi_2(\mathbf{x})$, given the value of $\phi_1(\mathbf{x})$ for this hypothetical example. The intrinsic restriction in the projection space as indicated by the shaded region explains the conic pattern observed in Figures 4 and 5.

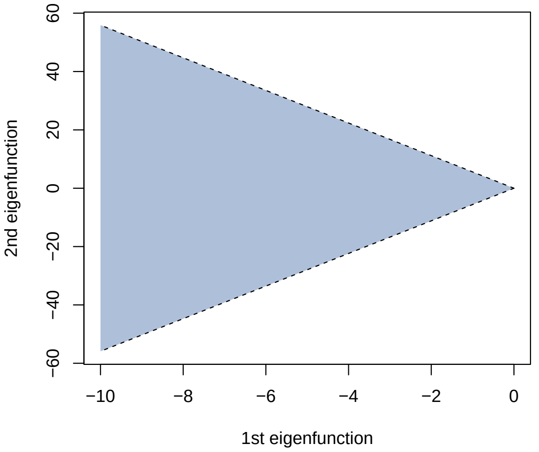


Figure 6: Restricted projection space for kernel PCA with quadratic kernel when the leading eigenfunctions are $\phi_1(\mathbf{x}) = -0.1x_1^2 - 0.02x_2^2$ and $\phi_2(\mathbf{x}) = -0.5x_1x_2$.

4 Analysis of Handwritten Digit Data

We carry out nonlinear principal component analysis of handwritten digit data from Le Cun et al. (1990) (also used in Hastie et al. (2009) for illustration) with polynomial kernels and investigate the geometry of the induced data embeddings in relation to the sample moment matrices. Kernel PCA of the data using Gaussian kernel and sigmoid kernel is discussed in Schölkopf et al. (1998). The handwritten digits are scanned from the ZIP codes written on U.S. postal envelopes. All the images of digits have been rescaled and normalized, resulting in 16 × 16 grayscale maps with the intensity of each pixel ranging from 0 to 255. The original data set includes the digits from 0 to 9, but we will focus on the digit pairs of (3, 8) and (7, 9) as they are often confused with each other.

Figure 7 shows pairwise plots of the two leading eigenvectors of the kernel matrix with polynomial kernel of degree 1 to 4 for the pair of digits, 3 and 8. Just for illustration of geometric patterns of the data projections by kernel PCA, we used only 100 randomly chosen images for each digit. In each panel, the second principal component appears to capture the difference between the two digits. The directions of the two principal components seem to alternate when the degree changes.

In order to gain insight about the major variation in the digit images captured by the first principal components for degrees 1 and 2, we sample the images closest to a regular grid for the principal components and visualize them. For each principal component, we computed the 5% and

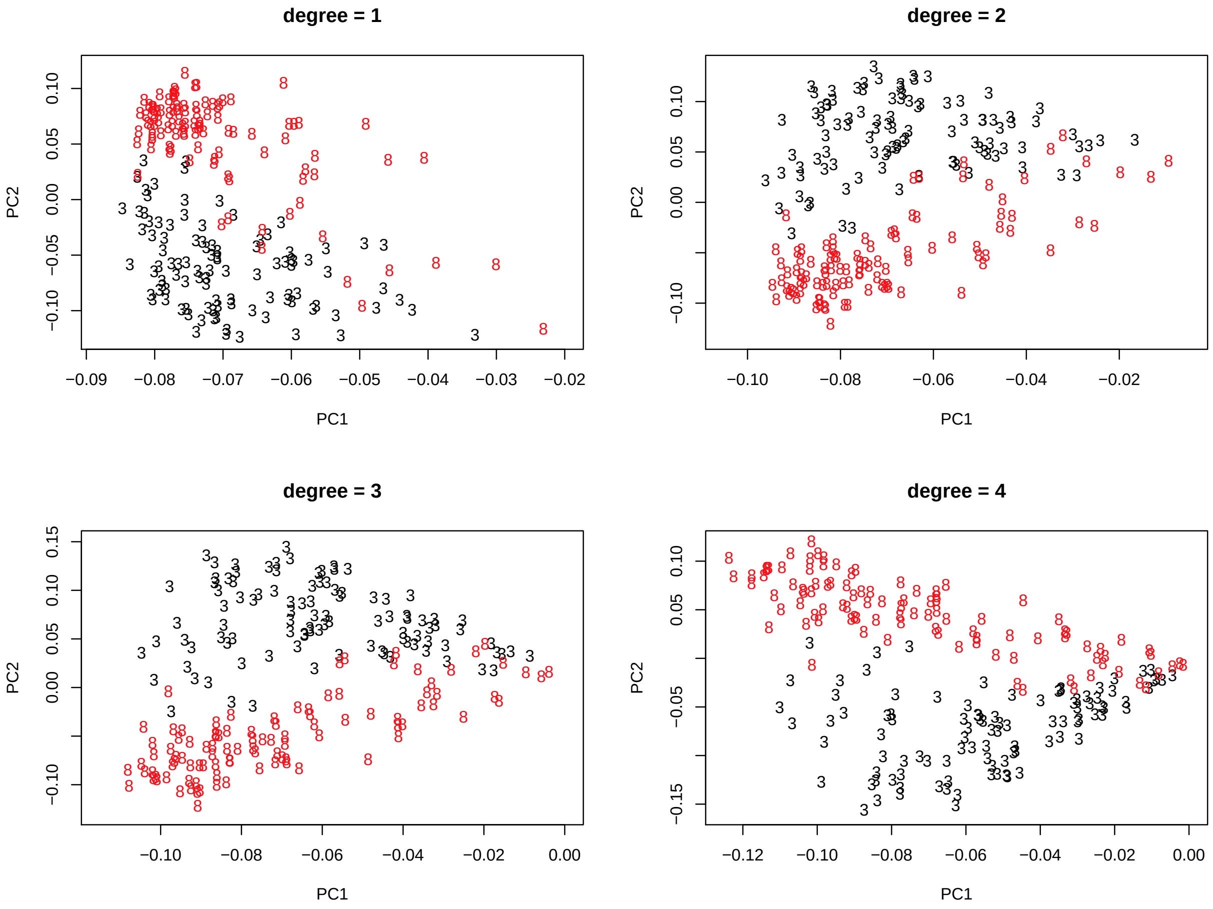
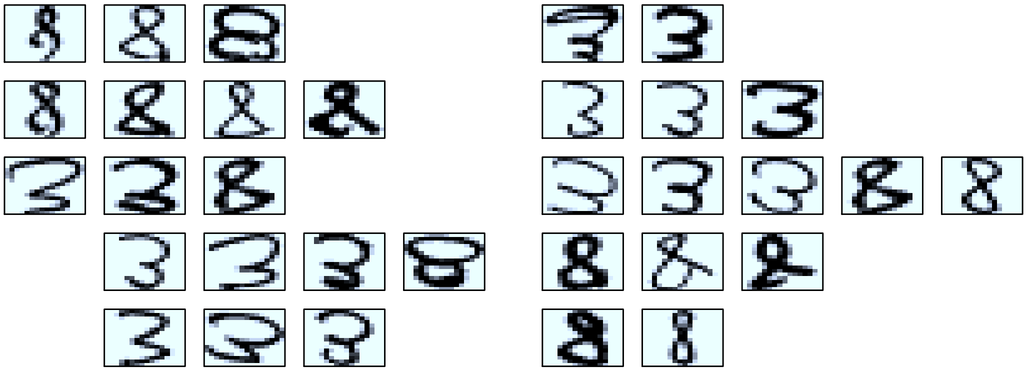


Figure 7: Projections of handwritten digits 3 and 8 by kernel PCA with polynomial kernel of degree 1 to 4.

95% quantiles, and considered five equi-distant values in the range, which lead to a total of 25 grid points. Figure 8 displays the sample images with their principal component values closest to the grid for the first and second degrees. The spatial location of the images in the figure approximately corresponds to the grid in the principal components space, and the boxes corresponding to those grid points without observed images nearby are left blank. The second principal component in each panel obviously indicates the change from one digit to the other. The first principal component seems to account for the change in digit size from small to large for degree 1 in the left panel and from large to small for degree 2 in the right panel.

Figure 9 displays similar projection plots for the digit pair of 7 and 9. When the degree of polynomial kernel varies, there seems relatively little change in the projections. The same phenomenon is observed in Figure 7. A possible explanation for this lies in high dimensionality of the data, where the dimension $p = 256$ is about the same order as the sample size $n = 200$. The theoretical analysis of the spectrum of kernel matrices in El Karoui (2010) suggests that nonlinear kernel PCA in high dimensions becomes effectively linear PCA.

To illustrate the connection between theoretical eigenfunctions and principal components in Figures 7 and 9, we consider approximation of the eigenfunctions of kernel PCA based on the sample moment matrices. For $d = 1$, there are 256 possible combinations of $i_1, \cdots, i_{256}$ which



(a) degree=1 (b) degree=2

Figure 8: Images corresponding to a 5 × 5 grid over the first two principal components for kernel PCA of the handwritten digits 3 and 8.

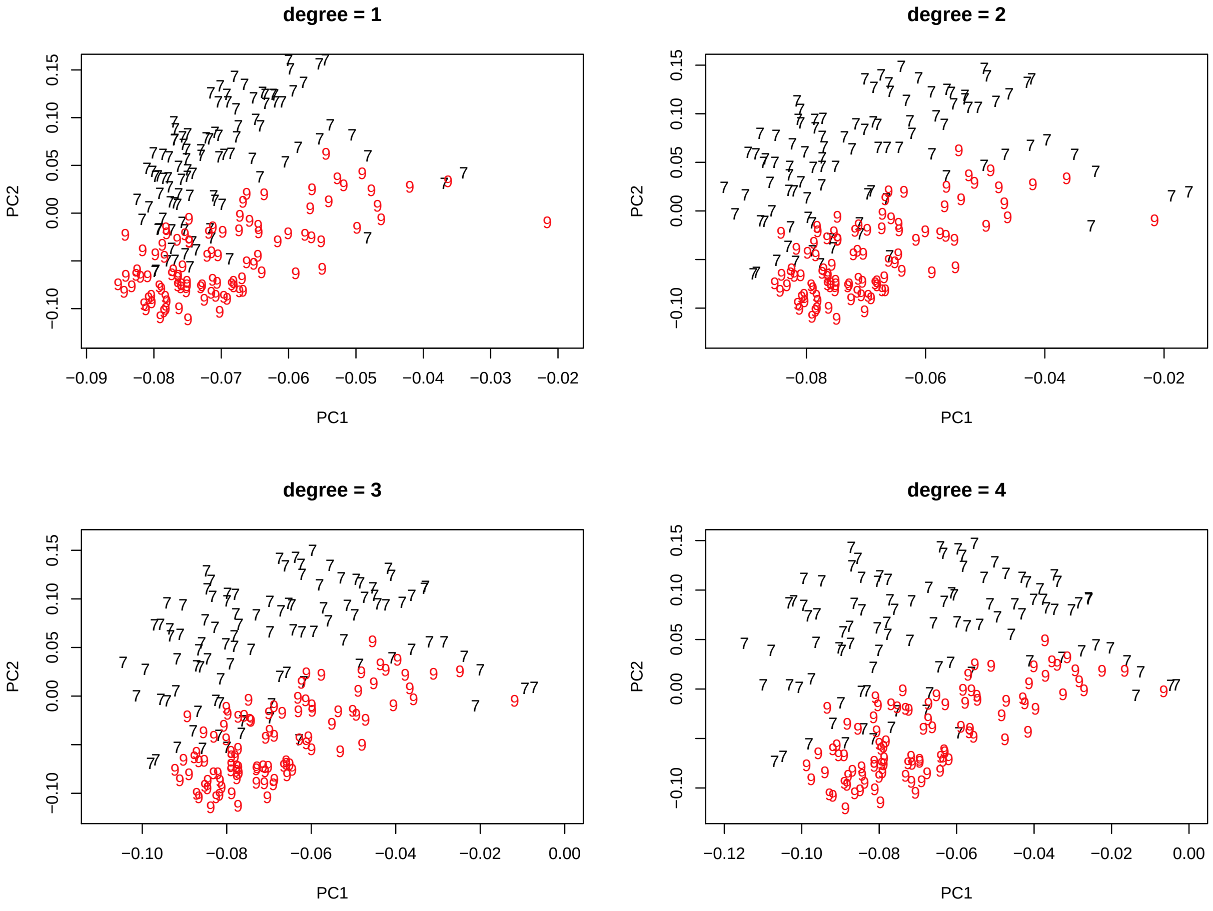


Figure 9: Projections of handwritten digits 7 and 9 by kernel PCA with polynomial kernel of degree 1 to 4.

sum up to 1, leading to a 256 × 256 moment matrix M^1_{256}. Given eigenvalue λ and eigenvector

$(C_1, \ldots, C_{256})^t$ of the moment matrix, we have the eigenfunction of the form:

$$\phi(\mathbf{x}) = \frac{1}{\lambda} \sum_{k=1}^{256} C_k x_k.$$

We approximate the moment matrix by its sample version and use the eigenvalue and eigenvector of the sample moment matrix to approximate the eigenfunction. For digit pairs (3, 8) and (7, 9), the left panels of Figure 10 show the projections of the handwritten digits given by the approximate eigenfunctions of linear PCA.

When $d = 2$, the number of possible combinations of $i_1, \cdots, i_{256}$ which add up to 2, becomes $256 + \binom{256}{2} = 32,896$. Hence, the dimension of the moment matrix M_{256}^2 is 32,896. Given that M_{256}^2 is such a huge matrix, we decided to bypass the direct eigen-analysis of the sample moment matrix of the same size and approximate the theoretical eigenfunctions by reducing the dimension of the images by a factor of 4 first and applying the same procedure to the reduced images. We reduce the dimension of images from 16×16 to 8×8 by averaging the four pixel values in the unit of 2×2 block, which yields 64 new variables instead of the original 256 variables. The right panels of Figure 10 show the projections of digits given by the approximate eigenfunctions based on the sample moment matrices of the 8×8 reduced pixel images when $d = 2$. Figure 10 shows similar geometric patterns as in Figures 7 and 9 for both digit pairs, indicating good agreement between the theoretical eigenfunctions and estimated principal components.

5 Discussion

We have shown that the spectrum of the polynomial kernel operator given a probability distribution can be characterized by the matrix of the corresponding moments. Applying the theoretical result to various examples, we have examined the effect of the degree of polynomial kernel on induced data embeddings in low dimensional setting. Further we have discussed that the form of eigenfunctions of the polynomial kernel operator brings some functional relation between them, which, in turn, restricts data projections to a certain region. In general, a proper choice between even and odd degrees depends on the features of the underlying distribution we wish to capture with low dimensional data embeddings. Contrary to the common suggestion for increasing degree to gain flexibility in kernel methods, our numerical analysis raises questions about the virtue of projections given by high-order polynomials in kernel PCA as they tend to be very sensitive to outliers.

Inspired by the mixture normal example, we want to further investigate the relationship between geometric properties of leading eigenfunctions and a family of parametric distributions. Also, it would be worthwhile to explore ways to mitigate the sensitivity of PCA with polynomial kernels to outlying observations. Some of the approaches taken to make PCA robust in the literature as in Candes et al. (2009) and Hubert et al. (2005) could be extended to kernel PCA. As another popular kernel method, kernel LDA has been used successfully in many applications. For example, Mika et al. (1999) conducted an experimental study showing that kernel LDA is competitive in comparison to other classification methods. It would be interesting to extend the eigen-analysis of the kernel operator in this paper analogously to the operator associated with kernel LDA for better understanding of the kernel LDA projections in relation to the probability distribution for data.

References

J. Ahn. A stable hyperparameter selection for the Gaussian RBF kernel for discrimination. *Statistical Analysis and Data Mining*, 3(3):142–148, 2010.

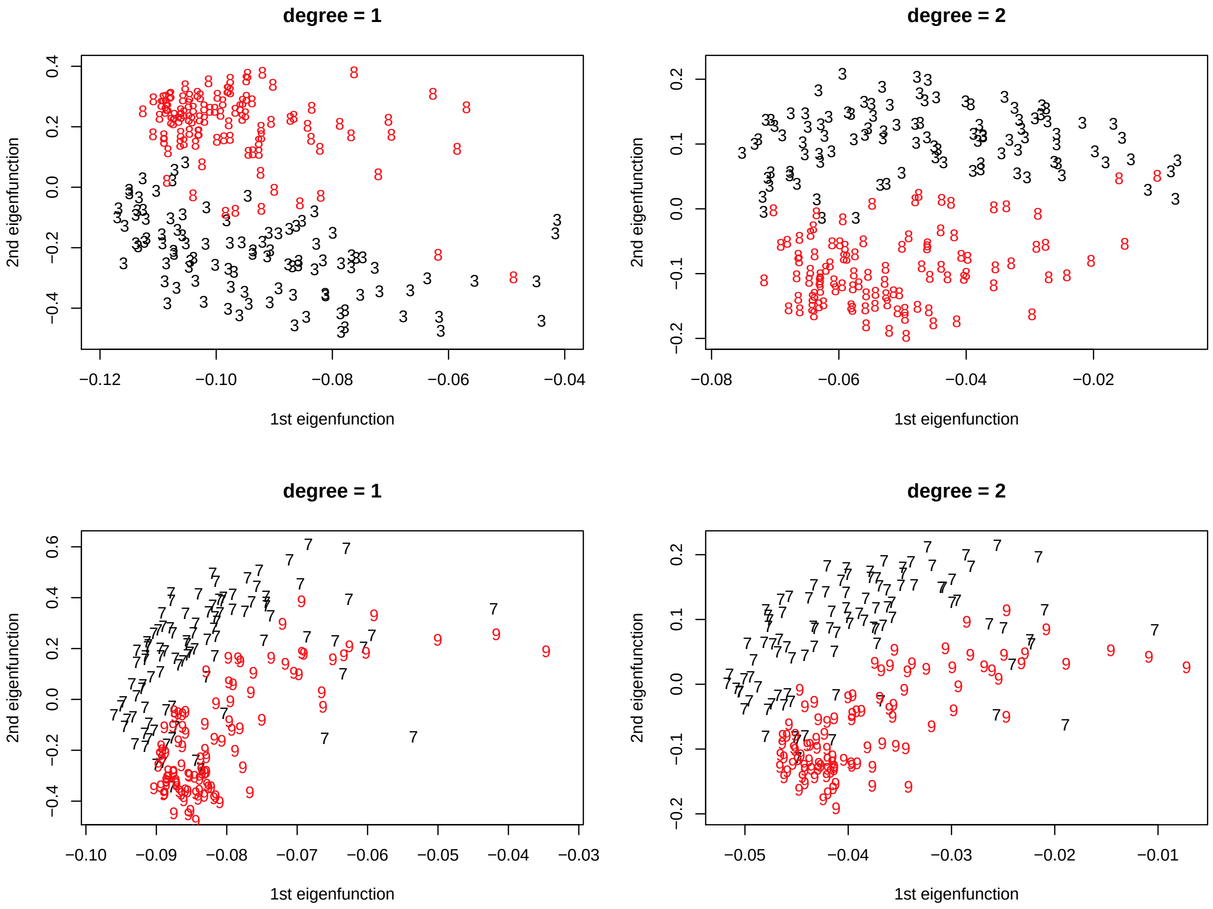


Figure 10: Projections of digits given by approximate eigenfunctions of kernel PCA that are based on the sample moment matrices.

M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

N. Aronszajn. Theory of reproducing kernel. *Transactions of the American Mathematical Society*, 68:3337–404, 1950.

G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.

E.J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *arXiv preprint arXiv:0912.3599*, 2009.

N. El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning - Data mining, Inference and Prediction*. New York: Springer, 2009.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):pp. 1171–1220, 2008.

M. Hubert, P.J. Rousseeuw, and K.V. Branden. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

L Kaufmann. Solving the quadratic programming problem arising in support vector classification. In *Advances in Kernel Methods - Support Vector Learning*, pages 147–167, Cambridge, MA, 1999. MIT Press.

Y. Le Cun, B. Boster, J. Denkern, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a backpropogation network. *In advances in Neural Information Processing Systems*, pages 396–404, 1990.

S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and K. R. Müller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–48, 1999.

B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

B. Schölkopf, A. J. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, pages 1299–1319, 1998.

B. Schölkopf, C. Burges, and A. J. Smola. *Advances in Kernel Methods - Support Vector Learning.* MIT Press, 1999.

G. L. Scott and H. C. Longuet-Higgins. Feature grouping by relocalisation of eigenvectors of proximity matrix. In *Proceedings of British Machine Vision Conference*, pages 103–108, 1990.

T. Shi, M. Belkin, and B. Yu. Data spectroscopy: eigenspaces of convolution operators and clustering. *The Annals of Statistics*, 37(6B):3960–3984, 2009.

V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, MR1367965, 1995.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

G. Wahba. *Spline models for observational data.* Philadelphia, PA: Society for Industrial and Applied Mathematics, 1990.

C. K. Williams and M. Seeger. The effect of the input density distribution on kernel-based classifiers. In *Proceedings of the 17th International Conference on Machine Learning*, pages 1159–1166. Morgan Kaufmann, 2000.

H. Zhu, C. K. Williams, R. Rhower, and M. Morciniec. Gaussian regression and optimal finite-dimensional linear models. *Neural Networks and Machine Learning*, (C. Bishop, ed.):167–184, 1998.