EIGEN-ANALYSIS OF KERNEL OPERATORS FOR NONLINEAR DIMENSION REDUCTION AND DISCRIMINATION

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of the Ohio State University

By

Zhiyu Liang

Graduate Program in Statistics

The Ohio State University

2014

Dissertation Committee:

Yoonkyung Lee, Advisor

Tao Shi

Vincent Vu

© Copyright by

Zhiyu Liang

2014

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 dissertation, we examine the geometry of the nonlinear embeddings for kernel PCA and kernel LDA through spectral analysis of the corresponding kernel operators. In particular, we carry out eigen-analysis of the polynomial kernel operator associated with data distributions and investigate the effect of the degree of polynomial on the data embedding. We also investigate the effect of centering kernels on the spectral property of both polynomial and Gaussian kernel operators. In addition, we extend the framework of the eigen-analysis of kernel PCA to kernel LDA by considering between-class and within-class variation operators for polynomial kernels. The results provide both insights into the geometry of nonlinear data embeddings given by kernel methods and practical guidelines for choosing an appropriate degree for dimension reduction and discrimination with polynomial kernels.

This is dedicated to my parents, Yigui Liang and Yunmei Hou; my sister, Zhiyan Liang and my husband, Sungmin Kim.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation and gratitude first and foremost to my advisor, Dr. Yoonkyung Lee, for her continuous advice and encouragement throughout my doctoral studies and dissertation work. Without her excellent guidance, patience and constructive comments, I could have never finished my dissertation successfully.

My gratitude also goes to the members of my committee, Dr. Tao Shi, Dr. Vincent Vu, and Dr. Prem Goel for their guidance in my research and valuable comments. Special thanks also go to Dr. Randolph Moses for willing to participate in my final defense committee and giving constructive comments afterwards. I also thank Dr. Elizabeth Stasny for her help throughout my Ph.D. study; Dr. Rebecca Sela and Dr. Nader Gemayel for their valuable advice that made my internship fruitful and exciting.

I also thank my parents and sister for always supporting me and encouraging me with their best wishes.

Last but not least, I would like express my love, appreciation, and gratitude to my husband Sungmin Kim for his continuous help, support, love and many valuable academic advices.

VITA

2004	B.Sc. in Applied Mathematics, Shanghai University of Finance and Economics, Shanghai, China
2008-Present	Graduate Teaching /Research Associate, The Ohio State University, Columbus, OH

PUBLICATIONS

Liang, Z. and Lee, Y. (2013), *Eigen-analysis of Nonlinear PCA with Polynomial Kernels.* Statistical Analysis and Data Mining, Vol 6, Issue 6, pp 529-544.

FIELDS OF STUDY

Major Field: Statistics

TABLE OF CONTENTS

Abstract ii

Dedication iii

Acknowledgments iv

Vita v

List of Figures viii

CHAPTER PAGE

1 Introduction 1

2 Kernel Methods 6

2.1 Kernel 6
2.2 Kernel method 9
2.2.1 Examples of kernel methods 10
2.3 Kernel Operator 14
2.3.1 Definition 14
2.3.2 Eigen-analysis of the Gaussian kernel operator 16

3 Eigen-analysis of Kernel Operators for Nonlinear Dimension Reduction 18

3.1 Eigen-analysis of the Polynomial Kernel Operator 18
3.1.1 Two-dimensional setting 19
3.1.2 Multi-dimensional setting 22
3.1.3 Polynomial kernel with constant 24
3.2 Simulation Studies 26
3.2.1 Uniform example 26
3.2.2 Mixture normal example 29
3.2.3 Effect of degree 33
3.2.4 Restriction in embeddings 35
3.3 Analysis of Handwritten Digit Data 37

4 On the Effect of Centering Kernels in Kernel PCA 43

4.1 Centering in the feature space 44
4.1.1 Simple illustration 47
4.1.2 Centered data with centered kernel 48
4.1.3 Uncentered data with uncentered kernel 50
4.1.4 Uncentered data with centered kernel 53
4.1.5 Extension 56
4.2 General Eigen-analysis for Centered Kernel Operator 57
4.2.1 Derivation of eigenfunction-eigenvalue pair 57
4.2.2 Orthogonality of eigenfunctions 61
4.2.3 Connection with the centered polynomial kernel operator result 62
4.3 Analysis of Centered Gaussian Kernel Operator 63
4.3.1 Centered Gaussian kernel operator 63
4.3.2 One-component normal 64
4.3.3 Mixture normal distribution 72

5 Eigen-analysis of Kernel Operators for Nonlinear Discrimination . . . 77

5.1 The Population Version of Kernel LDA 78
5.2 Eigen-analysis of the Polynomial Kernel Operator 80
5.2.1 Two-dimensional setting 80
5.2.2 Multi-dimensional setting 84
5.3 Simulation Studies 87
5.4 Effect of Degree 93

6 Conclusion and Discussion 99

6.1 Conclusion 99
6.2 Discussion 100

APPENDICES

A Proof for the form of leading eigenfunctions in a simple centered data example 103

B Example for the centered data with centered kernel 109

C Remarks on $\mathcal{K}_p$ being a valid mapping 112

Bibliography 114

LIST OF FIGURES

FIGURE PAGE

3.1 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. 28

3.2 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. 31

3.3 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. 32

3.4 The mixture normal data and their projections through principal components with polynomial kernel of varying degrees. The colors distinguish the two normal components. 34

3.5 “Wheel” data and their projections through principal components with polynomial kernel of varying degrees. The colors distinguish the two clusters. 36

3.6 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 $\phi_2(\mathbf{x}) = -0.5x_1x_2$. 37

3.7 Projections of handwritten digits 3 and 8 by kernel PCA with polynomial kernel of degree 1 to 4. 39

3.8 Images corresponding to a 5×5 grid over the first two principal components for kernel PCA of the handwritten digits 3 and 8. . . . 40

3.9 Projections of handwritten digits 7 and 9 by kernel PCA with polynomial kernel of degree 1 to 4. 41

3.10 Projections of digits given by approximate eigenfunctions of kernel PCA that are based on the sample moment matrices. 42

4.1 Comparison of the contours of the leading eigenfunction for the uncentered kernel operator (left) and centered kernel operator (right) in the bivariate normal example with $k = 2$. 50

4.2 Contour plots of the leading eigenfunctions of the uncentered kernel operator for the distribution setting in (4.6) when the center of data distribution gradually moves along the x_1 axis away from the origin. 52

4.3 The trichotomy of the leading eigenfunction form for the uncentered data distribution case in (4.6) with centered kernel. 55

4.4 Contours of the leading eigenfunction as m increases when $k = 1.5$ for the uncentered data distribution in (4.6) with centered kernel operator. 56

4.5 Contours of the leading eigenfunction when the center of the data distribution moves from the origin along the 45 degree line. 58

4.6 The first row shows the first five eigenvectors of an uncentered Gaussian kernel matrix with the bandwidth $w = 1.5$ for data sampled from normal distribution $N(2, 1^2)$ and the second row shows the first five eigenvectors of the centered kernel matrix. The third row shows the linear combinations of eigenvectors with the coefficients derived from our analysis; the fourth row shows the first five theoretical eigenfunctions for the centered kernel operator; the fifth row shows the five eigenfunctions for the uncentered kernel operator. . . 66

4.7 The inner product $\langle \phi_1, \frac{\phi_*}{\|\phi_*\|} \rangle_p$ versus the value of w. 69

4.8 The top row shows five leading eigenvectors of a Gaussian kernel matrix of data sampled from a mixture normal distribution $0.6N(2, 1^2) + 0.4N(-2, 1^2)$, the second row shows the five leading eigenvectors of the uncentered kernel matrix. The third row shows the theoretical eigenfunctions of the centered kernel operator we obtained. 76

5.1 The contours of the probability density function of a mixture normal example with two classes. The red circles and blue crosses show the data points generated from the distribution. 88

5.2 The left panel shows the contours of the empirical discriminant function from kernel LDA algorithm with linear kernel; the right panel shows the contours of theoretical discriminant function. . . . 90

5.3 The left panel shows the contours of the empirical discriminant function from kernel LDA algorithm; the right panel shows the contours of theoretical discriminant function ($d = 2$). 92

5.4 The left panel shows the embeddings of kernel LDA algorithm; the right panel shows the contours of the discriminant function ($d = 3$). 94

5.5 "Wheel" data and contours of the theoretical discriminant functions of kernel LDA with polynomial kernel of varying degrees based on the sample moments. 95

5.6 Contours of the empirical discriminant functions for "wheel" data. . 96

5.7 The scatterplot of two explicit features x_1^3 vs $\sqrt{3}x_1^2x_2$ (polynomial kernel with $d = 3$) for the bivariate normal example in (5.4). Red circles and blue crosses represent two classes. 97

5.8 The scatterplot of two features x_1^2 vs x_2^2 (polynomial kernel with $d = 2$) for the "wheel" data. Black circles and red crosses represent the outer circle and inner cluster in the original data. 98

5.9 Comparison between the first principal component of kernel PCA and the discriminant function of kernel LDA for polynomial kernel with $d = 2$ over "wheel" data. 98

CHAPTER 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 (SVMs) (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.

In spectral clustering and kernel PCA, the kernel determines the projections or data embeddings to be used for uncovering clusters or for representing data effectively in a low dimensional space, 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 dissertation, 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. Although Gaussian kernel is generally more flexible as a universal approximator, the two kernels have different merits, and the polynomial kernel with appropriate degrees can be often as effective as Gaussian kernel. For example, 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 standard PCA, eigen-decomposition is performed on the covariance matrix to obtain the principal components. Analogous to this standard practice, data are centered in the feature space and the corresponding centered version of the kernel matrix is commonly used in kernel PCA. We explore the effect of centering kernels on the spectral property of both polynomial kernel and Gaussian kernel operators, using the explicit form of the centered kernel operator. In particular, we characterize the change in the spectrum from the uncentered counterpart.

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. We extend the eigen-analysis of the kernel operator for kernel PCA to the general eigen-problem associated with kernel LDA, which leads to better understanding of the kernel LDA projections in relation to the underlying data distribution on the nonlinear embedding for discrimination. We mainly investigate the eigen-analysis of the polynomial kernel operator for kernel LDA and comment on the effect of the degree.

Chapter 2 gives introduction to technical details of the kernel, kernel operator and kernel methods. It also provides a review on the eigen-analysis of the Gaussian kernel operator. Chapter 3 presents the eigen-analysis of nonlinear PCA with polynomial kernels. Section 3.1 includes the general results of the eigen-analysis of the polynomial kernel operator defined through data distributions, and we show that the matrix of moments determines the eigenvalues and eigenfunctions. In Section 3.2, 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 3.3, 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. Chapter 4 mainly focuses on the effect of centering kernels. Section 4.1 regards how centering kernel affects the spectral property of the polynomial kernel operator. We show examples using both centered and uncentered polynomial kernels to illustrate the difference. In Section 4.2, we use Mercer's theorem to express the kernel function for general analysis of the centered kernel operator, which encompasses the result for the polynomial kernel operator. Section 4.3 examines the effect of centering kernels on the spectral property of the Gaussian kernel operator. We investigate both one-component normal and multi-component normal examples and describe the change in the spectrum after centering. Chapter 5 extends the current framework for analysis of kernel PCA to kernel LDA. By solving the general eigen-problem associated with the population version of kernel LDA, we characterize the theoretical discriminant function that maximizes the between-class variation relative to within-class variation. The polynomial kernel function is used in this derivation. Numerical examples are given in Section 5.3 and Section 5.4 to compare the empirical discriminant function and theoretical discriminant function. Chapter 6 concludes the dissertation with discussions.

CHAPTER 2

KERNEL METHODS

2.1 Kernel

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 $\mathcal{X}$, e.g. $\mathcal{X} = \mathbb{R}^p$. Then a kernel function is defined as a semi-positive definite mapping from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$, i.e.:

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, (x_i, x_j) \to K(x_i, x_j).$$

The kernel function is symmetric, which means $K(x, y) = K(y, x)$. Besides, there are some properties of a kernel that are worth noting:

$$K(x, x) \geq 0$$
$$K(u, v) \leq \sqrt{K(u, u)K(v, v)}.$$

For the corresponding data point (x_i, x_j), the kernel function can be expressed in terms of inner product as follows:

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle,$$

where Φ is typically a nonlinear map from the input space to an inner product space $\mathcal{H}$,

$$\Phi : \mathcal{X} \to \mathcal{H}.$$

The reason we introduce the inner product space is that being able to compute the inner product allows us to perform related geometrical constructions with the information of angles, distances or lengths. Given such formulation, we call the similarity measure function K a kernel, Φ its feature map and $\mathcal{H}$ the corresponding feature space. We say that kernel K corresponds to inner products in the feature space $\mathcal{H}$ via a feature mapping Φ.

Schölkopf and Smola (2002) showed that kernels which correspond to the inner products in the feature space coincide with the class of non-negative definite kernels. Some examples of non-negative definite kernels can be found to be evaluated efficiently even if they correspond to inner products in infinite dimensional inner product space. The correspondence is thus critical.

Historically, kernels are closely related to reproducing kernels typically used for nonparametric function estimation. The following summary of construction of reproducing kernel Hilbert space and reproducing kernel gives an example of well-defined kernel space and the corresponding kernel.

To define reproducing kernels, consider a Hilbert space $\mathcal{H}_K$ of real valued functions on an input domain $\mathcal{X}$. Note that a Hilbert space $\mathcal{H}_K$ is a complete inner product linear space, which is different from the feature space $\mathcal{H}$. In Wahba (1990), a reproducing kernel Hilbert space is defined as a Hilbert space of real valued functions, where for each $x \in \mathcal{X}$, the evaluation functional $L_x(f) = f(x)$ is bounded in $\mathcal{H}_K$.

By the Riesz representation theorem, if $\mathcal{H}_K$ is a reproducing kernel Hilbert space, then there exists an element $K_x \in \mathcal{H}_K$, the representer of evaluation at x, such that

$$L_x(f) = \langle K_x, f \rangle = f(x), \quad \forall f \in \mathcal{H}_K$$

see Aronszajn (1950) for details. The symmetric bivariate function $K(x, y)$ (Note that $K(x, y) = K_x(y) = \langle K_x, K_y \rangle = \langle K_y, K_x \rangle$) is called the reproducing kernel and it has the reproducing property $\langle K(x, \cdot), f(\cdot) \rangle = f(x)$. It can be shown that any reproducing kernel is non-negative definite.

There exists a one-to-one correspondence between reproducing kernel Hilbert spaces and non-negative definite functions. The Moore-Aronszajn theorem states that for every reproducing kernel Hilbert space $\mathcal{H}_K$ of functions, there corresponds a unique reproducing kernel $K(x, y)$, which is non-negative definite. Conversely, given a non-negative definite function $K(s, t)$ on $\mathcal{X}$, we can construct a unique reproducing kernel Hilbert space $\mathcal{H}_K$ that has $K(s, t)$ as its reproducing kernel.

Given the kernel function, we define the kernel matrix in the following way. Let $x_1, \dots, x_n \in \mathcal{X}$ be an iid sample and K be the kernel function, the kernel matrix is given as a $n \times n$ matrix:

$$K_n = [K(x_i, x_j)].$$

We say a kernel matrix is non-negative definite if it satisfies the condition

$$\sum_{i,j} c_i c_j K(x_i, x_j) \geq 0.$$

A kernel function which generates a non-negative definite kernel matrix K_n is called a non-negative definite kernel.

2.2 Kernel method

Kernel methods are given as nonlinear generalization of linear methods by mapping data into a high dimensional feature space and applying the linear methods in the feature space. In most kernel methods, the key step is to replace the inner product with the kernel so that the explicit form of feature mapping is not required. This substitution is called the "kernel trick" in machine learning. Such trick allows us to handle problems which are difficult to solve in the high or even infinite dimensional feature space directly.

We introduce three popular kernel methods in the following section where the kernel trick is applied. Some examples of positive definite kernels in those kernel methods include Gaussian kernel

$$K(x, x') = e^{-\|x-x'\|^2/2\sigma^2},$$

polynomial kernel of degree d

$$K(x, x') = (1 + \langle x, x' \rangle)^d,$$

sigmoid kernels

$$K(x, x') = \tanh(\kappa(x \cdot x') + \Theta)$$

and so on.

2.2.1 Examples of kernel methods

(a) Support Vector Machines (SVM)

Several authors came up with the class of hyperplanes based on the data $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, with the input domain of $x_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$,

$$\mathbf{x}^t\boldsymbol{\beta} + \beta_0 = 0 \quad \text{where} \quad \boldsymbol{\beta} \in \mathbb{R},$$

corresponding to the following classification rule

$$f(\mathbf{x}) = \text{sign}(\mathbf{x}^t\boldsymbol{\beta} + \beta_0),$$

and proposed a learning algorithm for problems which are linearly separable by finding hyperplane that create the largest margin between the points for different classes through optimization problem (Vapnik and Lerner 1963; Vapnik and Chervonenkis 1964). We call the above classifier which finds linear boundaries in the input space the support vector classifier (Hastie et al. 2009). In the non-separable case, where the classes overlap, the optimization problem can be generalized by allowing some points on the wrong side of the margin, which leads to the objective function

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i'=1}^{n} \alpha_i \alpha_{i'} y_i y_{i'} x_i^t x_{i'}.$$

While the support vector classifier finds the linear boundary, the procedures can be more flexible by mapping the data into the feature space. We call this extension the Support Vector Machines. It produces the nonlinear boundaries in the input space by constructing linear boundaries in the feature space to achieve better separation. Through feature mapping, the objective function has the form

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} y_i y_{i'} \langle \Phi(x_i), \Phi(x_{i'}) \rangle.$$

By applying the kernel trick, we replace $K(x_i, x_i^{'}) = \langle \Phi(x_i), \Phi(x_{i'}) \rangle$, then the the support vector machines for two-class classification problems have the following form

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + \beta_0.$$

The support vector machine is generally used to solve two-class problems. It can also be extended to multiclass problems by solving many two-class problems. SVMs have applications in many supervised and unsupervised learning problems.

(b) Kernel PCA

The standard PCA is a powerful tool for extracting a linear structure in the data. The principal components are obtained by eigen-decomposition of the covariance matrix. Schölkopf et al. (1998) proposed kernel PCA by computing the inner product in the feature space using kernel functions in the input space. In this kernel method, one can compute the principal components in a high-dimensional feature space, which is related to input space by some nonlinear feature mapping. Similar to SVM, this kernel method enables the construction of nonlinear version of principle component analysis.

Suppose the covariance matrix for the centered observations $x_i, i = 1, \cdots, n$, is $C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^t$. PCA computes the principal components by solving the equation

$$C\mathbf{v} = \lambda\mathbf{v}$$

By mapping the data into the feature space $\mathcal{H}$, the covariance matrix in $\mathcal{H}$ can be written in the form of $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^t$. The problem is thus turned into finding the eigen-decomposition of $\bar{C}$,

$$\bar{C}\mathbf{u} = \lambda\mathbf{u}.$$

Notice that the computation involved is prohibitive when the feature space is very high dimensional. Replacing $\mathbf{u} = \sum_{i=1}^{n} \alpha_i \Phi(x_i)$ in the above equation, we have the eigenvalue problem of the kernel matrix

$$K_n\boldsymbol{\alpha} = n\lambda\boldsymbol{\alpha},$$

where the kernel matrix is given by $K_n = [K(x_i, x_j)] = [\langle\Phi(x_i), \Phi(x_j)\rangle]$. We thus have the matrix $\boldsymbol{\alpha}$ with its columns as the eigenvectors of the kernel matrix, let $\boldsymbol{\alpha}^k$ indicate the eigenvectors with respect to the eigenvalue λ_k. Correspondingly, the projections of any point x onto the normalized eigenvectors $\mathbf{u}^k$ of the covariance matrix in the feature space can be derived as $\sum_{i=1}^{n} \alpha_i^k K(x_i, x)$. We are thus able to obtain embeddings in the feature space for any data point in kernel PCA setting.

(c) Kernel LDA

In kernel PCA, we aim to find the principal components explaining as much variance of data as possible, which are for describing the data. When it comes to classfication, we look for the features which discriminate between the two classes given the label information. The classical classification algorithms include linear and quadratic discriminant analysis, which assume the Gaussian distribution for each class. Fisher's linear discriminant direction is obtained through maximizing the between-class variance relative to the within-class variance.

Mika et al. (1999) proposed a nonlinear classification technique based on Fisher's linear discriminant analysis, which could be useful when the classification boundary is not clear. By mapping the data into the feature space, the kernel trick allows us to find Fisher's linear discriminant in the feature space $\mathcal{H}$, leading to a nonlinear discriminant direction in the input space.

Assume we have two classes for this classification problem and let $\mathcal{D}_1 = \{x_1^1, \cdots, x_{n_1}^1\}$ and $\mathcal{D}_0 = \{x_1^0, \cdots, x_{n_0}^0\}$ be samples from those two different classes. Then the sample size is $n = n_1 + n_0$. Let Φ be the feature mapping into the feature space $\mathcal{H}$. To find the linear discriminant in $\mathcal{H}$, we need to find the direction w which maximizes the between-class variation relative to within-class variation in the feature space, i.e.

$$J(w) = \frac{w^t S_B w}{w^t S_W w}. \tag{2.1}$$

Here $w \in \mathcal{H}$ and S_B and S_W are the matrices in the feature space,

$$\begin{aligned} S_B &= (m_1^\Phi - m_0^\Phi)(m_1^\Phi - m_0^\Phi)^t \quad \text{and} \\ S_W &= \sum_{l=1,0} \sum_{x \in \mathcal{D}_l} (\Phi(x) - m_l^\Phi)(\Phi(x) - m_l^\Phi)^t, \end{aligned}$$

where $m_l^\Phi = \frac{1}{n_l} \sum_{j=1}^{n_l} \Phi(x_j^l), l = 1, 0$ is the mean of feature vectors in class l.

When $w \in \mathcal{H}$ is in the span of all training samples in the feature space, w can be written as $w = \sum_{i=1}^{n} \alpha_i \Phi(x_i)$. Plugging w into the equation (2.1) and expanding both the numerator and denominator in terms of α using the kernel trick, we have

$$w^t S_B w = \alpha^t B \alpha \quad \text{and} \quad w^t S_W w = \alpha^t W \alpha,$$

where B and W are defined based on the kernel matrix, see Section 5.1 for details of B and W. Therefore, Fisher's linear discriminant can be found through the eigen-problem:

$$B\alpha = \lambda W \alpha,$$

Similar to the kernel PCA, any projection of a new pattern on the direction w in the feature space is given by the linear combination of the coefficients α_i and kernel functions evaluated at the new point and original data $K(x_i, x)$, which gives the empirical discriminant function $\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ in kernel LDA.

2.3 Kernel Operator

2.3.1 Definition

As we mentioned before, our kernel function K_n 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. Given a probability distribution P with density function $p(x)$ and a kernel function K, the distribution-dependent kernel operator is defined as

$$\mathcal{K}_p f(y) = \int_{\mathcal{X}} K(x, y) f(x) p(x) dx \tag{2.2}$$

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

$$\mathcal{K}_p\phi = \lambda\phi \quad \text{or} \quad \int_{\mathcal{X}} K(x,y)\phi(x)p(x)dx = \lambda\phi(y). \tag{2.3}$$

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, $K_n = [K(x_i, x_j)]$. From the discussion about kernel PCA in Section 2.2.1, we know that kernel PCA 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. Baker (1977) gives the theory of the numerical solution of eigenvalue problems, showing that the eigenvalues of K_n converge to eigenvalues of the kernel operator as $n \to \infty$. The eigen-analysis of kernel PCA is useful for understanding the kernel method on the population level.

2.3.2 Eigen-analysis of the Gaussian kernel operator

As we discussed in the introduction, 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 the normal distributions.

Shi et al. (2009) obtained the refined version of analytic results in Zhu et al. (1998) for the spectrum of Gaussian kernel operator with the univariate Gaussian case. When the probability density function is normal with $P \sim N(\mu, \sigma^2)$ and the kernel function $K(x,y) = \exp(-\frac{(x-y)^2}{2w^2})$, the eigenvalues and eigenfunctions are given explicitly by

$$\lambda_i = \sqrt{\frac{2}{(1+\beta+\sqrt{1+2\beta})}} \left(\frac{\beta}{1+\beta+\sqrt{1+2\beta}} \right)^{i-1}$$

$$\phi_i(x) = \frac{(1+2\beta)^{1/8}}{\sqrt{2^{i-1}(i-1)!}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \frac{\sqrt{1+2\beta}-1}{2} \right) H_{i-1}\left(\left(\frac{1}{4} + \frac{\beta}{2} \right)^{1/4} \frac{x-\mu}{\sigma} \right),$$

for $i = 1, 2, \cdots$, where $\beta = 2\sigma^2/w^2$, H_i is the ith order Hermite polynomial; see Koekoek and Swarttouw (1998) for more details about Hermite polynomials. Williams and Seeger (2000) investigated the dependence of the eigenfunction on the Gaussian input density and discussed how this dependence determines the basis functions for classification problems. Shi et al. (2009) explored this connection in an attempt to understand spectral clustering method from a population point of view.

Many clustering algorithms use the top eigenvectors of the kernel matrix or its normalized version (Scott and Longuet-Higgins 1990; Perona and Freeman 1998; Shi and Malik 2000). Despite their empirical success, some limitations of the above standard spectral clustering are noted in Nadler and Galun (2007). For example, they pointed out that those clustering algorithms based on the kernel matrix with a single parameter (e.g, Gaussian kernel) would fail when dealing with clusters of different scales.

Shi et al. (2009) investigated the spectral clustering from a population level when the distribution P includes several separate high-density components. They found that when there is enough separation among the components, each of the top eigenfunctions of the kernel operator corresponds to one of the separate components, with the order of eigenfunctions determined by the mixture proportion and the eigenvalues. They also showed that the top eigenfunction of the kernel operator for separate components is the only eigenfunction with no sign change. Hence, when each mixture component has enough separation from the other components, the number of eigenfunctions with no sign change of the kernel operator $\mathcal{K}_P$ suggests the number of components of the distribution. Using the relationship between kernel matrix and kernel operator, we can estimate the number of clusters by the number of eigenvectors that have no sign change up to some precision.

CHAPTER 3

EIGEN-ANALYSIS OF KERNEL OPERATORS FOR NONLINEAR DIMENSION REDUCTION

3.1 Eigen-analysis of the 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. We examine eigen-expansion of the polynomial kernel operator based on the equation (2.3) 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 3.1.1 and generalize it to p-dimensional setting in Section 3.1.2. Then we extend the analysis further to the second type with an additional constant in Section 3.1.3.

3.1.1 Two-dimensional setting

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

$$\int_{\mathbb{R}^2} (\mathbf{x}^t\mathbf{y})^d \phi(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x} = \lambda\phi(\mathbf{y}) \tag{3.1}$$

in this setting. More explicitly,

$$K(\mathbf{x}, \mathbf{y}) = (x_1y_1 + x_2y_2)^d = \sum_{j=0}^{d} \binom{d}{j} (x_1y_1)^{d-j}(x_2y_2)^j = \sum_{j=0}^{d} \binom{d}{j} (x_1^{d-j}x_2^j)(y_1^{d-j}y_2^j).$$

Note that the polynomial kernel can be also expressed as the inner product of the so-called feature vectors, $\Phi(\mathbf{x})^t\Phi(\mathbf{y})$, through the feature map,

$$\Phi(\mathbf{x}) = \left(\binom{d}{0}^{\frac{1}{2}} x_1^d, \binom{d}{1}^{\frac{1}{2}} x_1^{d-1}x_2, \cdots, \binom{d}{d}^{\frac{1}{2}} x_2^d \right)^t.$$

Appendix C comments on that the mapping $\mathcal{K}_p$ is valid with the polynomial kernel. With the explicit expression of K, the equation (3.1) becomes

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

which is re-expressed as

$$\sum_{j=0}^{d} \binom{d}{j}^{\frac{1}{2}} y_1^{d-j}y_2^j \left[\int \binom{d}{j}^{\frac{1}{2}} x_1^{d-j}x_2^j\ \phi(\mathbf{x})p(\mathbf{x})\, d\mathbf{x} \right] = \lambda\phi(\mathbf{y}).$$

Let $C_j = \binom{d}{j}^{\frac{1}{2}} \int x_1^{d-j}x_2^j\ \phi(\mathbf{x})p(\mathbf{x})\, d\mathbf{x}$ be a distribution-dependent constant for $j = 0, \ldots, d$. Then for $\lambda \neq 0$, the corresponding eigenfunction $\phi(\cdot)$ should be of the form

$$\phi(\mathbf{y}) = \frac{1}{\lambda} \sum_{k=0}^{d} \binom{d}{k}^{\frac{1}{2}} C_k y_1^{d-k} y_2^k. \tag{3.2}$$

By substituting (3.2) for $\phi(\mathbf{x})$ in the defining equation for C_j, we get the following equations for the constants ($j = 0, \ldots, d$):

$$C_j = \frac{1}{\lambda}\binom{d}{j}^{\frac{1}{2}} \int x_1^{d-j} x_2^{j} \left[\sum_{k=0}^{d} \binom{d}{k}^{\frac{1}{2}} C_k x_1^{d-k} x_2^{k}\right] p(\mathbf{x}) d\mathbf{x},$$

which leads to

$$\lambda C_j = \sum_{k=0}^{d} \binom{d}{k}^{\frac{1}{2}} \binom{d}{j}^{\frac{1}{2}} C_k \int x_1^{2d-(j+k)} x_2^{(j+k)} p(\mathbf{x}) d\mathbf{x}.$$

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

$$\sum_{k=0}^{d} \binom{d}{k}^{\frac{1}{2}} \binom{d}{j}^{\frac{1}{2}} \mu_{2d-(j+k),(j+k})C_k = \lambda C_j \quad \text{for } j = 0, \ldots, d. \tag{3.3}$$

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

$$M_2^d = \begin{bmatrix} \binom{d}{0}\mu_{2d,0} & \binom{d}{0}^{\frac{1}{2}}\binom{d}{1}^{\frac{1}{2}}\mu_{2d-1,1} & \cdots & \binom{d}{0}^{\frac{1}{2}}\binom{d}{d}^{\frac{1}{2}}\mu_{d,d} \\ \binom{d}{1}^{\frac{1}{2}}\binom{d}{0}^{\frac{1}{2}}\mu_{2d-1,1} & \binom{d}{1}^{\frac{1}{2}}\binom{d}{1}^{\frac{1}{2}}\mu_{2d-2,2} & \cdots & \binom{d}{1}^{\frac{1}{2}}\binom{d}{d}^{\frac{1}{2}}\mu_{d-1,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ \binom{d}{d}^{\frac{1}{2}}\binom{d}{0}^{\frac{1}{2}}\mu_{d,d} & \binom{d}{d}^{\frac{1}{2}}\binom{d}{1}^{\frac{1}{2}}\mu_{d-1,d+1} & \cdots & \binom{d}{d}\mu_{0,2d} \end{bmatrix}, \tag{3.4}$$

we can succinctly express the set of equations as

$$M_2^d \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_d \end{bmatrix} = \lambda \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_d \end{bmatrix}. \tag{3.5}$$

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 (3.5), 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. Note that 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 2dth moments $\mu_{2d-j,j} = E(X_1^{2d-j}X_2^j), j = 0, \ldots, 2d$. 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

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

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

(iii) The eigenfunctions are polynomials of total degree d of the form in (3.2) with coefficients determined by the eigenvectors of M_2^d.

(iv) The eigenfunctions, ϕ_i, are orthogonal in the sense of

$$\langle \phi_i, \phi_j \rangle_p = \int_{\mathbb{R}^2} \phi_i(\mathbf{x})\phi_j(\mathbf{x})p(\mathbf{x})d\mathbf{x} = 0 \quad \textit{for} \quad i \neq j.$$

Proof. We prove the statement (*iv*). Let $\mathbf{C}_i$ and $\mathbf{C}_j$ be the eigenvectors of M_2^d corresponding to a pair of eigenfunctions ϕ_i and $\phi_j \in \mathcal{H}_K$ with eigenvalues λ_i and λ_j. Then for $i \neq j$,

$$\int_{\mathbb{R}^2} \phi_i(\mathbf{x})\phi_j(\mathbf{x})p(\mathbf{x})d\mathbf{x} \propto \mathbf{C}_i^t M_2^d \mathbf{C}_j = \mathbf{C}_i^t(\lambda_j \mathbf{C}_j) = \lambda_j \mathbf{C}_i^t \mathbf{C}_j = 0.$$

□

3.1.2 Multi-dimensional setting

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

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

and the equation (2.3) becomes

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

$$i.e. \sum_{j_1+\cdots+j_p=d} \binom{d}{j_1,\cdots,j_p}^{\frac{1}{2}} \prod_{k=1}^p y_k^{j_k} \left[\int \binom{d}{j_1,\cdots,j_p}^{\frac{1}{2}} \prod_{k=1}^p x_k^{j_k}\phi(\mathbf{x})p(\mathbf{x})\,d\mathbf{x}\right] = \lambda\phi(\mathbf{y}).$$

Letting $C_{j_1,\cdots,j_p} = \binom{d}{j_1,\cdots,j_p}^{\frac{1}{2}} \int \prod_{k=1}^p x_k^{j_k}\ \phi(\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$, we can write the eigenfunction $\phi(\cdot)$ as

$$\phi(\mathbf{x}) = \frac{1}{\lambda} \sum_{j_1+\cdots+j_p=d} \binom{d}{j_1,\cdots,j_p}^{\frac{1}{2}} C_{j_1,\cdots,j_p} \prod_{k=1}^p x_k^{j_k}. \tag{3.6}$$

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:

$$C_{j_1,\cdots,j_p}$$

$$=\frac{1}{\lambda}\binom{d}{j_1,\cdots,j_p}^{\frac{1}{2}}\int\prod_{k=1}^{p}x_k^{j_k}\left[\sum_{i_1+\cdots+i_p=d}\binom{d}{i_1,\cdots,i_p}^{\frac{1}{2}}C_{i_1,\cdots,i_p}\prod_{k=1}^{p}x_k^{i_k}\right]p(\mathbf{x})d\mathbf{x},$$

which is rewritten as

$$\lambda C_{j_1,\cdots,j_p}=\sum_{i_1+\cdots+i_p=d}\binom{d}{i_1,\cdots,i_p}^{\frac{1}{2}}\binom{d}{j_1,\cdots,j_p}^{\frac{1}{2}}C_{i_1,\cdots,i_p}\int\prod_{k=1}^{p}x_k^{i_k+j_k}p(\mathbf{x})\,d\mathbf{x}.$$

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

$$\sum_{i_1+\cdots+i_p=d}\binom{d}{i_1,\cdots,i_p}^{\frac{1}{2}}\binom{d}{j_1,\cdots,j_p}^{\frac{1}{2}}\mu_{j_1+i_1,\cdots,j_p+i_p}C_{i_1,\cdots,i_p}=\lambda C_{j_1,\cdots,j_p}. \qquad (3.7)$$

To express the above equation in matrix form, we generalize the moment matrix M_2^d to M_p^d with entries given by $\binom{d}{i_1,\cdots,i_p}^{\frac{1}{2}}\binom{d}{j_1,\cdots,j_p}^{\frac{1}{2}}\mu_{j_1+i_1,\cdots,j_p+i_p}$. 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 (3.7) is written as

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

where C is a d_p-vector with entries $C_{j_1,\cdots,j_p}$ for $j_1+\cdots+j_p=d$. Applying the similar argument used for the two-dimensional setting, 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 2dth 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 (3.6) *with coefficients given by the eigenvectors of* M_p^d.

(iv) The eigenfunctions are orthogonal with respect to the inner product, $\langle \phi_i, \phi_j \rangle_p = \int_{\mathbb{R}^p} \phi_i(\mathbf{x})\phi_j(\mathbf{x})p(\mathbf{x})d\mathbf{x}$.

3.1.3 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 Theorem 2. The eigenfunctions in this case are of the following form:

$$\phi(\mathbf{x}) = \frac{1}{\lambda} \sum_{j_1+j_2+j_3=2} \binom{2}{j_1, j_2, j_3}^{\frac{1}{2}} C_{j_1,j_2,j_3} x_1^{j_1} x_2^{j_2}. \tag{3.8}$$

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

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

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

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

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 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 $\mu^*_{i_1+j_1,\cdots,i_p+j_p} = E(\prod_{k=1}^{p} X_k^{i_k+j_k})$.

3.2 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.

3.2.1 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

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

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 those of M_2^2. We can get the eigenvalues of the matrix numerically, which 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,

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

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:

$$\begin{aligned}
\phi_1(\mathbf{x}) &= \frac{1}{0.5206}(-0.542x_1^2 - 0.908x_1x_2 - 0.542x_2^2),\\
\phi_2(\mathbf{x}) &= \frac{1}{0.0889}(0.707x_1^2 - 0.707x_2^2),\\
\phi_3(\mathbf{x}) &= \frac{1}{0.0127}(0.454x_1^2 - 1.084x_1x_2 + 0.454x_2^2).
\end{aligned}$$

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 $\hat{\lambda}_1 = 0.5409$, $\hat{\lambda}_2 = 0.0944$, and $\hat{\lambda}_3 = 0.0139$ after being scaled by the sample size n as discussed in Section 3.1. The sample eigenvalues are quite close to the theoretical ones.

Figure 3.1 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 blue 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 3.1 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.

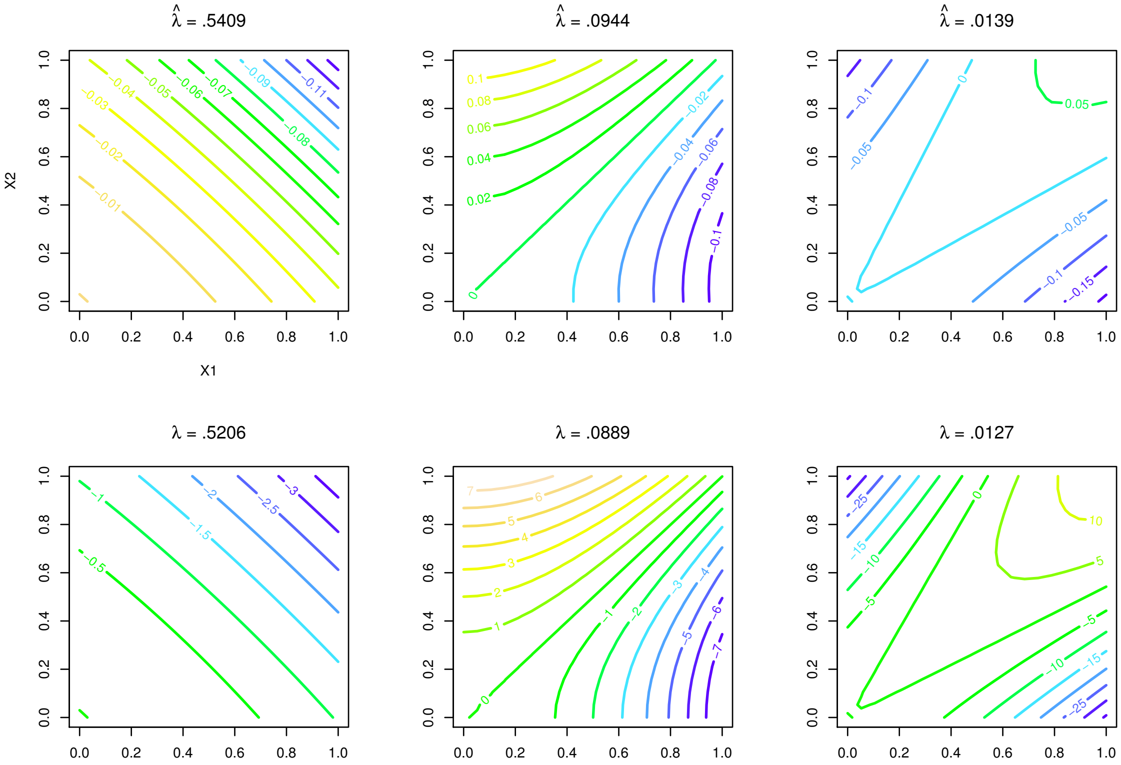


Figure 3.1: Comparison of the contours of the nonlinear embeddings given by three leading eigenvectors and the theoretical eigenfunctions for the uniform data. The upper three panels are for the embeddings induced by the eigenvectors for three nonzero eigenvalues, and the lower three panels are for the corresponding eigenfunctions.

3.2.2 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.

When degree is 2

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

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

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:

$$\phi_1(\mathbf{x}) = \frac{1}{110.593}(0.886x_1^2 - 0.597x_1x_2 + 0.191x_2^2),$$
$$\phi_2(\mathbf{x}) = \frac{1}{17.415}(-0.459x_1^2 - 1.050x_1x_2 + 0.487x_2^2),$$
$$\phi_3(\mathbf{x}) = \frac{1}{2.492}(0.064x_1^2 + 0.735x_1x_2 + 0.852x_2^2).$$

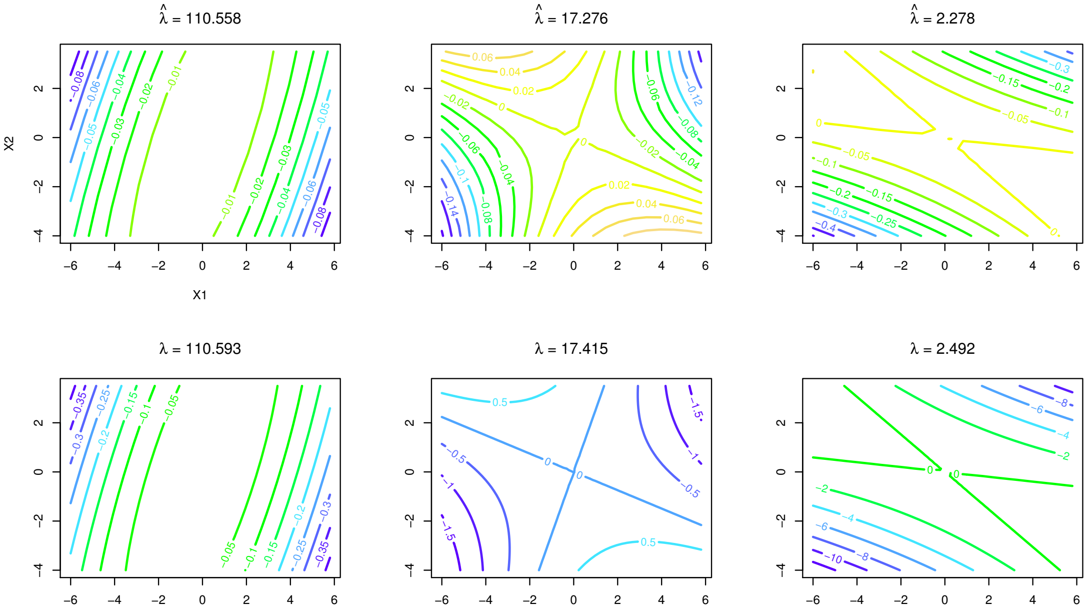
Contours of these eigenfunctions are displayed in the bottom panels of Figure 3.2.

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 3.4. Three nonzero eigenvalues from the kernel matrix are found to be $\hat{\lambda}_1 = 110.558$, $\hat{\lambda}_2 = 17.276$, and $\hat{\lambda}_3 = 2.278$. The top panels of Figure 3.2 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 approximate symmetry around the origin 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.

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

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



$\hat{\lambda} = 110.558$ $\hat{\lambda} = 17.276$ $\hat{\lambda} = 2.278$

$\lambda = 110.593$ $\lambda = 17.415$ $\lambda = 2.492$

Figure 3.2: 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.

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

$$\begin{aligned}
\phi_1(\mathbf{x}) &= \frac{1}{1862.615}(0.848x_1^3 - 0.791x_1^2x_2 + 0.437x_1x_2^2 - 0.097x_2^3),\\
\phi_2(\mathbf{x}) &= \frac{1}{343.748}(-0.518x_1^3 - 1.073x_1^2x_2 + 0.862x_1x_2^2 - 0.317x_2^3),\\
\phi_3(\mathbf{x}) &= \frac{1}{59.266}(-0.112x_1^3 - 1.079x_1^2x_2 - 0.929x_1x_2^2 + 0.559x_2^3),\\
\phi_4(\mathbf{x}) &= \frac{1}{9.870}(-0.026x_1^3 + 0.245x_1^2x_2 + 1.097x_1x_2^2 + 0.761x_2^3).
\end{aligned}$$

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 $\hat{\lambda}_1 = 1892.775$, $\hat{\lambda}_2 = 346.190$, $\hat{\lambda}_3 = 58.954$, and $\hat{\lambda}_4 = 8.365$.

Figure 3.3 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.

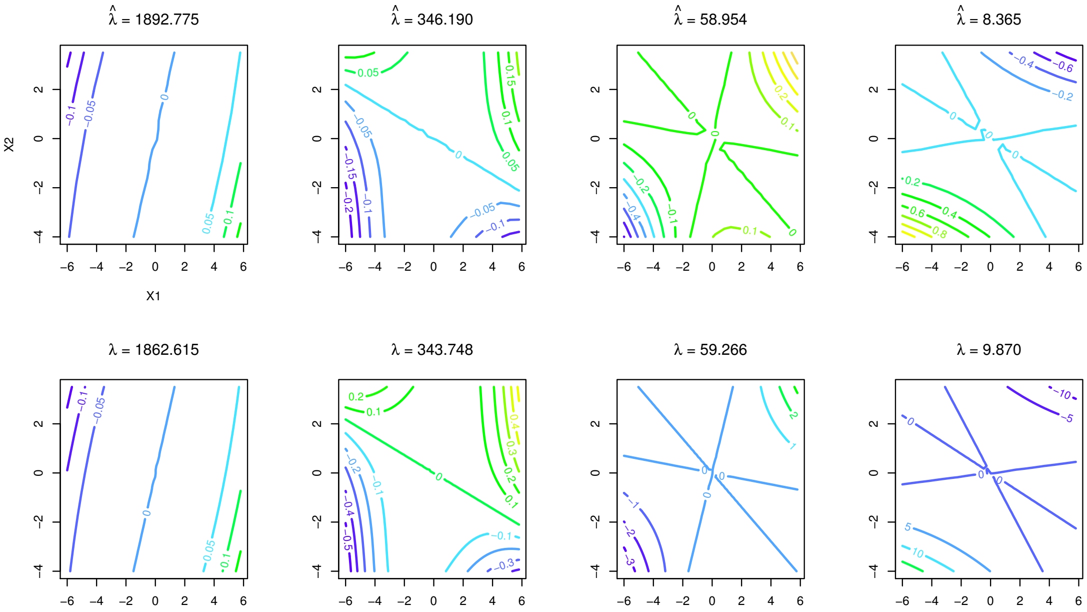


Figure 3.3: Comparison of the contours of the nonlinear embeddings given by four leading eigenvectors (top panels) and the theoretical eigenfunctions (bottom panels) for the mixture normal data when degree is 3.

For the polynomial kernel with constant

For the kernel function form $K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^t\mathbf{y})^2$ in the 2-dimensional case, the eigenfunctions can be derived based on the moment matrix M_3^2, as was mentioned in section 3.1.3,

$$M_3^2 = \begin{bmatrix} 90.5 & -25\sqrt{2} & -11\sqrt{2} & 15 & 2.5\sqrt{2} & 7.5 \\ -25\sqrt{2} & 30 & 5 & -10\sqrt{2} & -2 & -2.5\sqrt{2} \\ -11\sqrt{2} & 5 & 15 & -\sqrt{2} & -5 & -.5\sqrt{2} \\ 15 & -10\sqrt{2} & -\sqrt{2} & 10 & 0 & 2 \\ 2.5\sqrt{2} & -2 & -5 & 0 & 4 & 0 \\ 7.5 & -2.5\sqrt{2} & -.5\sqrt{2} & 2 & 0 & 1 \end{bmatrix}.$$

The corresponding eigenfunctions are of the following form

$$\phi(\mathbf{x}) = \frac{1}{\lambda}\{C_{2,0,0}x_1^2 + \sqrt{2}C_{1,1,0}x_1x_2 + \sqrt{2}C_{1,0,1}x_1 + C_{0,2,0}x_2^2 + \sqrt{2}C_{0,1,0}x_2 + C_{0,0,2}\}.$$

The first three leading eigenfunctions are

$$\begin{aligned} \phi_1(\mathbf{x}) &= \frac{1}{114.101}\{.873x_1^2 - .583x_1x_2 - .231x_1 + .185x_2^2 + .061x_2 + .075\}, \\ \phi_2(\mathbf{x}) &= \frac{1}{18.857}\{.328x_1^2 + .956x_1x_2 - .640x_1 - .459x_2^2 + .197x_2 - .030\}, \\ \phi_3(\mathbf{x}) &= \frac{1}{12.821}\{-.347x_1^2 - .463x_1x_2 - 1.100x_1 + .153x_2^2 + .532x_2 - .050\}. \end{aligned}$$

Compared to the kernel function form $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^t\mathbf{y})^2$, we see that the basis functions are different.

3.2.3 Effect of degree

Figure 3.4 shows a scatter plot of the mixture normal data used in the example discussed in the previous section 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 3.2 and 3.3, 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.

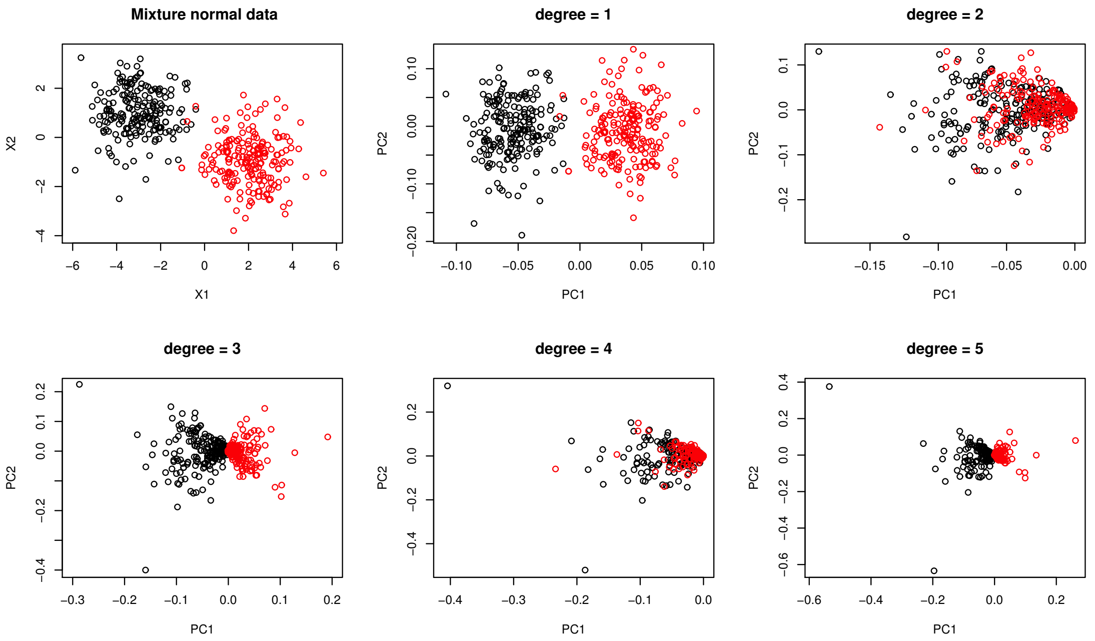


Figure 3.4: The mixture normal data and their projections through principal components with polynomial kernel of varying degrees. The colors distinguish the two normal components.

Figure 3.5 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 3.4 and 3.5 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 3.5 clearly suggests that lower even degree ($d = 2$) works better than higher even degree ($d = 6$) in providing separation of clusters for the example.

3.2.4 Restriction in embeddings

Figures 3.4 and 3.5 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

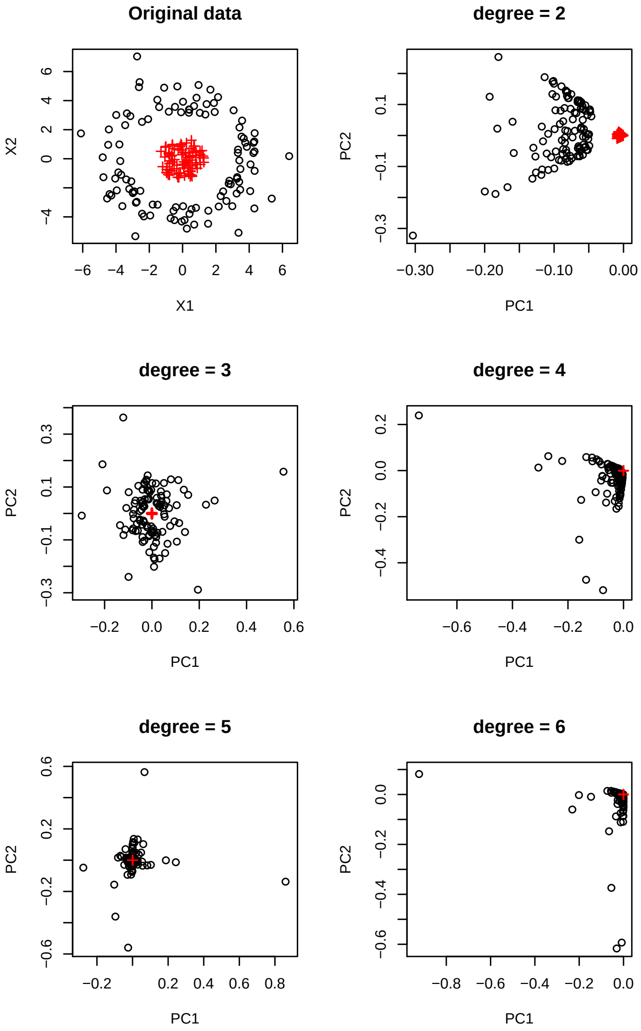


Figure 3.5: "Wheel" data and their projections through principal components with polynomial kernel of varying degrees. The colors distinguish the two clusters.

by considering a level set of the first eigenfunction and the values of the second eigenfunction corresponding to the level set. When $\phi_1(\mathbf{x}) = c_1x_1^2 + c_2x_2^2 = k$,

$$\phi_2^2(\mathbf{x}) = c_{12}^2x_1^2x_2^2 = c_{12}^2x_1^2\left[\frac{k - c_1x_1^2}{c_2}\right] = -\left(\frac{c_{12}^2c_1}{c_2}\right)x_1^4 + \left(\frac{c_{12}^2k}{c_2}\right)x_1^2.$$

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 3.6 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 3.4 and 3.5.

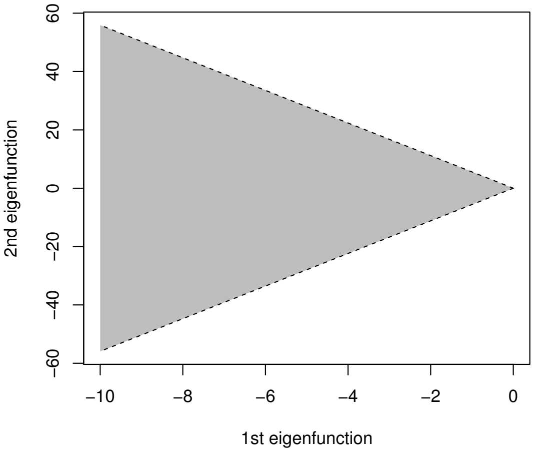


Figure 3.6: Restricted projection space for kernel PCA with quadratic kernel when the leading eigenfunctions are $\phi_1(\mathbf{x}) = -0.1x_1^2 - 0.02x_2^2$ and $\phi_2(\mathbf{x}) = -0.5x_1x_2$.

3.3 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 intensity values in each image are then scaled and translated to fall within the range from -1 to 1. 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 3.7 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 two principal components for degrees 1 and 2, we sampled the images closest to a regular grid for the principal components to visualize the variation. For each principal component, we computed the 5% and 95% quantiles, and considered five equi-distant values in the range, which lead to a total of 25 grid points. Figure 3.8 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

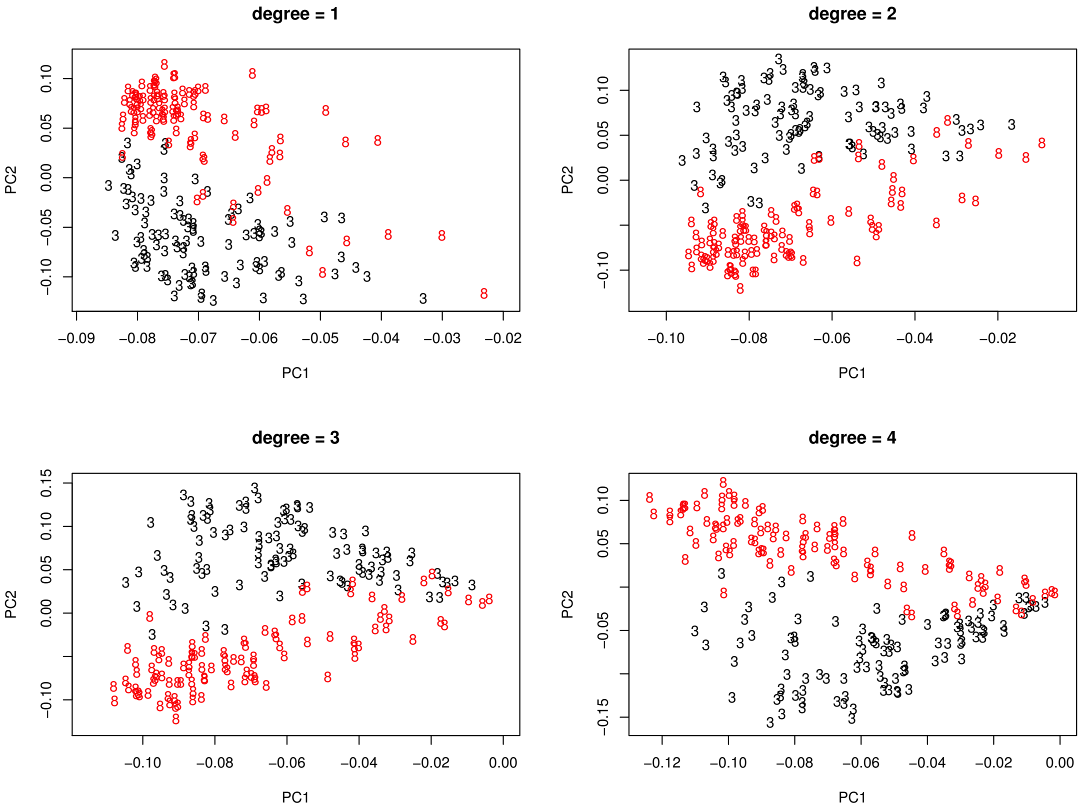
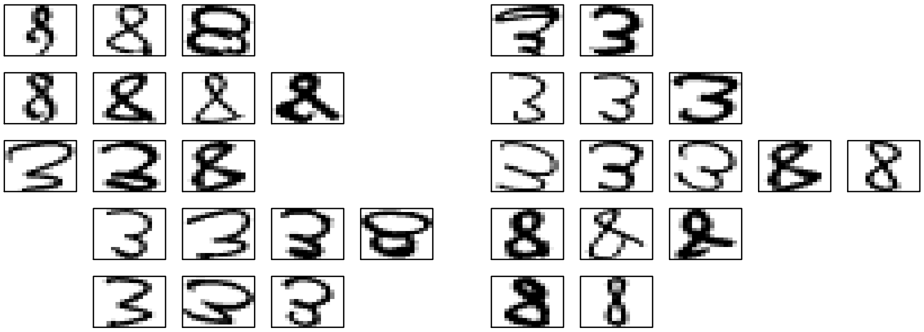


Figure 3.7: Projections of handwritten digits 3 and 8 by kernel PCA with polynomial kernel of degree 1 to 4.

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 3.9 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 3.7. A possible explanation for this lies in high dimensionality of the data, where the dimension $p = 256$



(a) degree=1

(b) degree=2

Figure 3.8: Images corresponding to a 5 × 5 grid over the first two principal components for kernel PCA of the handwritten digits 3 and 8.

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 3.7 and 3.9, 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 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 3.10 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

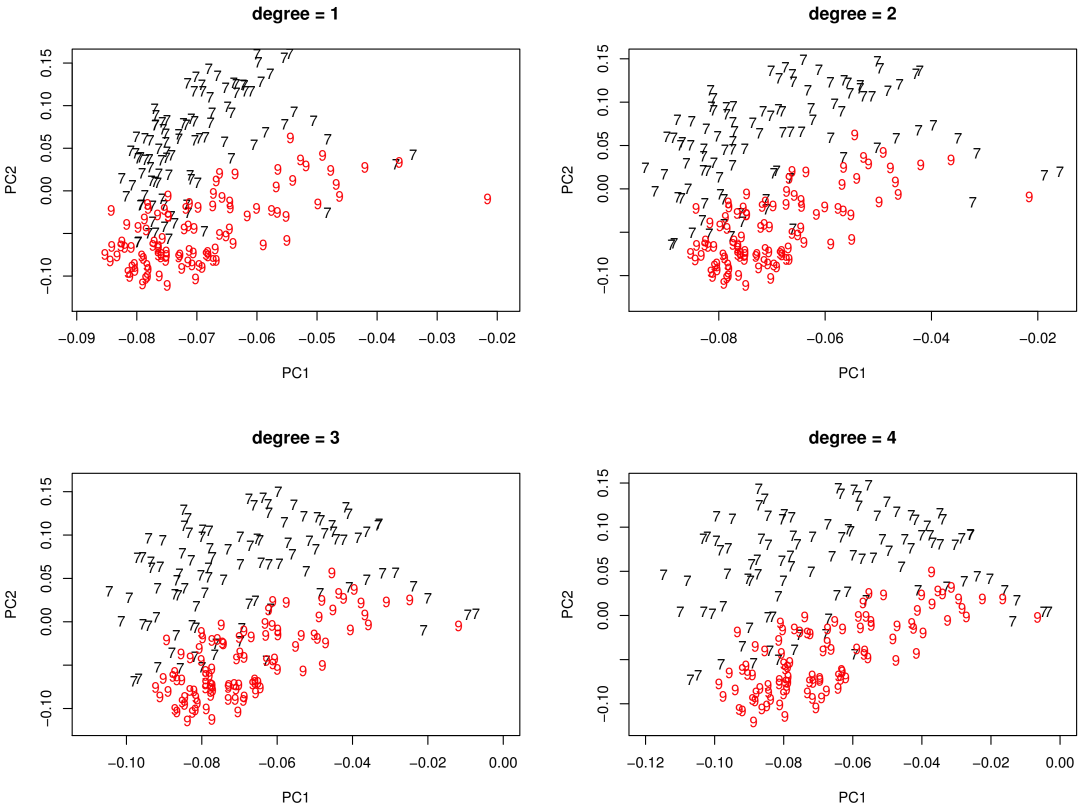


Figure 3.9: Projections of handwritten digits 7 and 9 by kernel PCA with polynomial kernel of degree 1 to 4.

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 3.10 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 3.10 shows similar geometric patterns as in Figures 3.7 and 3.9 for both digit pairs, indicating good agreement between the theoretical eigenfunctions and estimated principal components.

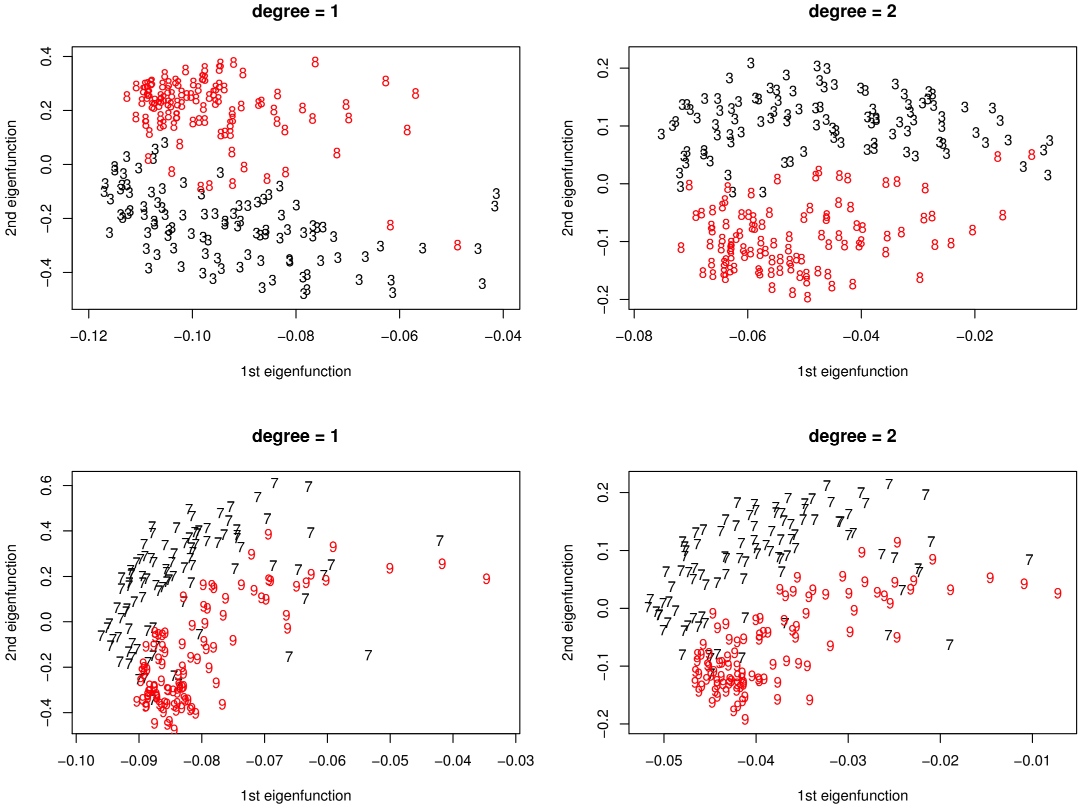


Figure 3.10: Projections of digits given by approximate eigenfunctions of kernel PCA that are based on the sample moment matrices.

CHAPTER 4
ON THE EFFECT OF CENTERING KERNELS IN KERNEL PCA

The eigen-analysis so far regards the kernel operator that is not centered in the feature space. Centering a kernel in the feature space differs from simply centering data. We note that most numerical examples in Sections 3.2 and 3.3 involve approximately centered data. In standard PCA, the principal components are obtained by eigen-decomposition of the covariance matrix which is based on variables centered around the origin. Analogous to the covariance matrix in standard PCA, centering data in the feature space leads to a centered version of the kernel matrix, which is commonly used in kernel PCA for dimension reduction. Just as the covariance matrix differs from the second moment matrix that combines mean and variance, the centered kernel is expected to be different from the uncentered counterpart in the data projection it produces. It would be interesting to examine the way centering a kernel affects the nonlinear PCA embedding, depending on the kernel type. In this chapter, we study the effect of centering kernels on the spectral property for both polynomial and Gaussian kernel operators and examine the difference between the centered and uncentered kernel operators in terms of the eigenfunctions.

4.1 Centering in the feature space

The centered kernel matrix $\tilde{K}_n$ is given by the inner product of the centered feature vectors, namely, $\tilde{\Phi}(x_i) = \Phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\Phi(x_j)$; see Schölkopf et al. (1998). With the representation of $K_{ij} \equiv K(x_i, x_j) = \Phi(x_i)^t\Phi(x_j)$, the ijth entry of $\tilde{K}_n$ is given as

$$\tilde{K}_{ij} = K_{ij} - \frac{1}{n}\sum_{l=1}^{n} K_{lj} - \frac{1}{n}\sum_{l=1}^{n} K_{il} + \frac{1}{n^2}\sum_{l=1}^{n}\sum_{m=1}^{n} K_{lm}.$$

The above sample version motivates the following function:

$$\tilde{K}(x,y) = K(x,y) - E_X[K(X,y)] - E_Y[K(x,Y)] + E_X E_Y[K(X,Y)] \tag{4.1}$$

as the population version of centered kernel function, which defines the centered kernel operator,

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

To contrast the centered and uncentered kernel operators through simple illustration, we focus on the case with $p = 2$. From $K(\mathbf{x},\mathbf{y}) = (\mathbf{x}^t\mathbf{y})^d = (x_1y_1 + x_2y_2)^d = \sum_{j=0}^{d}\binom{d}{j}(x_1y_1)^{d-j}(x_2y_2)^j$, we have the following expressions for the terms in $\tilde{K}$,

$$E_X[K(X,y)] = \int \sum_{j=0}^{d}\binom{d}{j}(x_1y_1)^{d-j}(x_2y_2)^j p(\mathbf{x})d\mathbf{x} = \sum_{j=0}^{d}\binom{d}{j}E[X_1^{d-j}X_2^j]y_1^{d-j}y_2^j,$$

$$E_Y[K(x,Y)] = \int \sum_{j=0}^{d}\binom{d}{j}(x_1y_1)^{d-j}(x_2y_2)^j p(\mathbf{y})d\mathbf{y} = \sum_{j=0}^{d}\binom{d}{j}E[Y_1^{d-j}Y_2^j]x_1^{d-j}x_2^j,$$

$$E_X E_Y[K(X,Y)] = E_X\left\{\sum_{j=0}^{d}\binom{d}{j}x_1^{d-j}x_2^j E[Y_1^{d-j}Y_2^j]\right\} = \sum_{j=0}^{d}\binom{d}{j}E[X_1^{d-j}X_2^j]E[Y_1^{d-j}Y_2^j].$$

Let $\mu_{d-j,j} = E(X_1^{d-j}X_2^j)$. Analogous to the steps in Section 3.1.1, we replace the centered kernel $\tilde{K}$ with the expansion to get following equation for the eigenfunction $\tilde{\phi}$ and eigenvalue λ:

$$\int [\sum_{j=0}^{d} \binom{d}{j} (x_1y_1)^{d-j}(x_2y_2)^j - \sum_{j=0}^{d} \binom{d}{j} \mu_{d-j,j} y_1^{d-j} y_2^j - \sum_{j=0}^{d} \binom{d}{j} \mu_{d-j,j} x_1^{d-j} x_2^j$$
$$+ \sum_{j=0}^{d} \binom{d}{j} \mu_{d-j,j}^2] \tilde{\phi}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \lambda \tilde{\phi}(\mathbf{y}).$$

The eigen equation is equivalent to:

$$\sum_{j=0}^{d} \binom{d}{j} y_1^{d-j} y_2^j \int x_1^{d-j} x_2^j \tilde{\phi}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - \sum_{j=0}^{d} \binom{d}{j} \mu_{d-j,j} y_1^{d-j} y_2^j \int \tilde{\phi}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$
$$- \sum_{j=0}^{d} \binom{d}{j} \mu_{d-j,j} \int x_1^{d-j} x_2^j \tilde{\phi}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} + \sum_{j=0}^{d} \binom{d}{j} \mu_{d-j,j}^2 \int \tilde{\phi}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$
$$= \lambda \tilde{\phi}(\mathbf{y}).$$

Using the same notation as in the previous chapter, let $C_j = \int \binom{d}{j}^{\frac{1}{2}} x_1^{d-j} x_2^j \tilde{\phi}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$, and let $m_{\tilde{\phi}} = E\tilde{\phi}(\mathbf{x}) = \int \tilde{\phi}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$. The equation can be written as

$$\sum_{j=0}^{d} \binom{d}{j}^{\frac{1}{2}} C_j y_1^{d-j} y_2^j - m_{\tilde{\phi}} \sum_{j=0}^{d} \binom{d}{j} \mu_{d-j,j} y_1^{d-j} y_2^j \tag{4.2}$$
$$- \sum_{j=0}^{d} \binom{d}{j}^{\frac{1}{2}} C_j \mu_{d-j,j} + m_{\tilde{\phi}} \sum_{j=0}^{d} \binom{d}{j} \mu_{d-j,j}^2 = \lambda \tilde{\phi}(\mathbf{y}).$$

By noting that $\tilde{\mathcal{K}}_p(1) = E_X \tilde{K}(X, y) = 0$, we can immediately see that $\lambda_0 = 0$ is the eigenvalue corresponding to the eigenfunction $\tilde{\phi}_0(x) = 1$. For any eigenfunction $\tilde{\phi}$ with nonzero eigenvalue, the equation (4.2) can be simplified by using the orthogonality of eigenfunctions, $\langle \tilde{\phi}, \tilde{\phi}_0 \rangle_p = 0$, which yields $m_{\tilde{\phi}} = 0$.

Therefore, given that $m_{\tilde{\phi}} = 0$, the equation (4.2) leads to

$$\tilde{\phi}(\mathbf{y}) = \frac{1}{\lambda} \sum_{k=0}^{d} \binom{d}{k}^{\frac{1}{2}} C_k \left(y_1^{d-k} y_2^k - \mu_{d-k,k} \right). \tag{4.3}$$

By plugging $\tilde{\phi}(\mathbf{x})$ back into C_j, we have

$$\int \binom{d}{j}^{\frac{1}{2}} x_1^{d-j} x_2^{j} \left[\sum_{k=0}^{d} \binom{d}{k}^{\frac{1}{2}} C_k \left(x_1^{d-k} x_2^{k} - \mu_{d-k,k} \right) \right] p(\mathbf{x}) d\mathbf{x} = \lambda C_j.$$

which can be expressed as

$$\sum_{k=0}^{d} \binom{d}{j}^{\frac{1}{2}} \binom{d}{k}^{\frac{1}{2}} \left(\mu_{2d-(j+k),j+k} - \mu_{d-k,k}\mu_{d-j,j} \right) C_k = \lambda C_j.$$

Defining the $(d+1) \times (d+1)$ matrix $\tilde{M}_2^d$ as,

$$\tilde{M}_2^d =$$
$$\begin{bmatrix} \binom{d}{0}(\mu_{2d,0} - \mu_{d,0}\mu_{d,0}) & \cdots & \binom{d}{0}^{\frac{1}{2}} \binom{d}{d}^{\frac{1}{2}} (\mu_{d,d} - \mu_{0,d}\mu_{d,0}) \\ \binom{d}{1}^{\frac{1}{2}} \binom{d}{0}^{\frac{1}{2}} (\mu_{2d-1,1} - \mu_{d,0}\mu_{d-1,1}) & \cdots & \binom{d}{1}^{\frac{1}{2}} \binom{d}{d}^{\frac{1}{2}} (\mu_{d-1,d+1} - \mu_{0,d}\mu_{d-1,1}) \\ \vdots & \ddots & \vdots \\ \binom{d}{d}^{\frac{1}{2}} \binom{d}{0}^{\frac{1}{2}} (\mu_{d,d} - \mu_{d,0}\mu_{0,d}) & \cdots & \binom{d}{d}(\mu_{0,2d} - \mu_{0,d}\mu_{0,d}) \end{bmatrix},$$

we can express the set of equations for C_j as

$$\tilde{M}_2^d \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_d \end{bmatrix} = \lambda \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_d \end{bmatrix}.$$

Therefore, the eigenfunctions and eigenvalues of the centered polynomial kernel operator are determined by the spectral decomposition of the "central" moment matrix $\tilde{M}_2^d$, which have different entries from its uncentered counterpart M_2^d in Theorem 1. We have the following theorem regarding the centered polynomial kernel operator.

Theorem 3. *Suppose that the probability distribution $p(x_1, x_2)$ defined on $\mathbb{R}^2$ has finite 2dth moments $\mu_{2d-j,j} = E(X_1^{2d-j}X_2^j) for j = 0, \ldots, 2d$. For the polynomial kernel of degree d, $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^t\mathbf{y})^d$,*

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

$$\tilde{M}_2^d =$$
$$\begin{bmatrix} \binom{d}{0}(\mu_{2d,0} - \mu_{d,0}\mu_{d,0}) & \cdots & \binom{d}{0}^{\frac{1}{2}}\binom{d}{d}^{\frac{1}{2}}(\mu_{d,d} - \mu_{0,d}\mu_{d,0}) \\ \binom{d}{1}^{\frac{1}{2}}\binom{d}{0}^{\frac{1}{2}}(\mu_{2d-1,1} - \mu_{d,0}\mu_{d-1,1}) & \cdots & \binom{d}{1}^{\frac{1}{2}}\binom{d}{d}^{\frac{1}{2}}(\mu_{d-1,d+1} - \mu_{0,d}\mu_{d-1,1}) \\ \vdots & \ddots & \vdots \\ \binom{d}{d}^{\frac{1}{2}}\binom{d}{0}^{\frac{1}{2}}(\mu_{d,d} - \mu_{d,0}\mu_{0,d}) & \cdots & \binom{d}{d}(\mu_{0,2d} - \mu_{0,d}\mu_{0,d}) \end{bmatrix}.$$

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

(iii) The eigenfunctions are of the form in (4.3) *with coefficients determined by the eigenvectors of $\tilde{M}_2^d$.*

4.1.1 Simple illustration

To simply illustrate the result above, we take $d = 2$ as an example. The "central" moment matrix is given by

$$\tilde{M}_2^2 = \begin{bmatrix} \mu_{4,0} - \mu_{2,0}\mu_{2,0} & \sqrt{2}(\mu_{3,1} - \mu_{1,1}\mu_{2,0}) & \mu_{2,2} - \mu_{0,2}\mu_{2,0} \\ \sqrt{2}(\mu_{3,1} - \mu_{2,0}\mu_{1,1}) & 2(\mu_{2,2} - \mu_{1,1}\mu_{1,1}) & \sqrt{2}(\mu_{1,3} - \mu_{0,2}\mu_{1,1}) \\ \mu_{2,2} - \mu_{2,0}\mu_{0,2} & \sqrt{2}(\mu_{1,3} - \mu_{1,1}\mu_{0,2}) & \mu_{0,4} - \mu_{0,2}\mu_{0,2} \end{bmatrix}. \quad (4.4)$$

Notice that $\mu_{4,0} - \mu_{2,0}\mu_{2,0} = E(X_1^4) - E(X_1^2)E(X_1^2) = \mathrm{var}(X_1^2)$ for instance. By expressing other terms in a similar way, we can also write the "central" moment matrix as

$$\tilde{M}_2^2 = \begin{bmatrix} \mathrm{var}(X_1^2) & \sqrt{2}\mathrm{cov}(X_1^2, X_1X_2) & \mathrm{cov}(X_1^2, X_2^2) \\ \sqrt{2}\mathrm{cov}(X_1^2, X_1X_2) & 2\mathrm{var}(X_1X_2) & \sqrt{2}\mathrm{cov}(X_2^2, X_1X_2) \\ \mathrm{cov}(X_2^2, X_1^2) & \sqrt{2}\mathrm{cov}(X_2^2, X_1X_2) & \mathrm{var}(X_2^2) \end{bmatrix}.$$

As a centered version, this matrix has the variance and covariance of the terms of total degree 2 as entries instead of the original moments around zero. Notice that $\tilde{M}_2^2$ is the same as the variance-covariance matrix of the feature vectors. The eigenvectors of $\tilde{M}_2^2$ determine the coefficients of eigenfunctions of $\tilde{\mathcal{K}}_p$.

We will investigate the effect of centering data and centering a kernel separately in the following subsections based on simple examples.

4.1.2 Centered data with centered kernel

For comparison between two kernel operators, we consider the following data distribution,

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & k\sigma^2 \end{pmatrix} \right), \tag{4.5}$$

where k is a positive constant. Note that the distribution is centered at the origin. We obtain the following "central" moment matrix based on (4.4) for the centered kernel operator:

$$\tilde{M}_2^2 = \begin{bmatrix} \mu_{4,0} - \mu_{2,0}\mu_{2,0} & \sqrt{2}(\mu_{3,1} - \mu_{1,1}\mu_{2,0}) & \mu_{2,2} - \mu_{0,2}\mu_{2,0} \\ \sqrt{2}(\mu_{3,1} - \mu_{1,1}\mu_{2,0}) & 2(\mu_{2,2} - \mu_{1,1}\mu_{1,1}) & \sqrt{2}(\mu_{1,3} - \mu_{0,2}\mu_{1,1}) \\ \mu_{2,2} - \mu_{2,0}\mu_{0,2} & \sqrt{2}(\mu_{1,3} - \mu_{1,1}\mu_{0,2}) & \mu_{0,4} - \mu_{0,2}\mu_{0,2} \end{bmatrix}$$
$$= \begin{bmatrix} 2\sigma^4 & 0 & 0 \\ 0 & 2k\sigma^4 & 0 \\ 0 & 0 & 2k^2\sigma^4 \end{bmatrix},$$

while the moment matrix for the uncentered counterpart is

$$M_2^2 = \begin{bmatrix} \mu_{4,0} & \sqrt{2}\mu_{3,1} & \mu_{2,2} \\ \sqrt{2}\mu_{3,1} & 2\mu_{2,2} & \sqrt{2}\mu_{1,3} \\ \mu_{2,2} & \sqrt{2}\mu_{1,3} & \mu_{0,4} \end{bmatrix} = \begin{bmatrix} 3\sigma^4 & 0 & k\sigma^4 \\ 0 & 2k\sigma^4 & 0 \\ k\sigma^4 & 0 & 3k^2\sigma^4 \end{bmatrix}.$$

The moment matrix M_2^2 gives the leading eigenfunction for $\mathcal{K}_p$ whose contours are ellipses with the minor axes along the direction of the largest variation in the distribution (for example, x_2-axis when $k > 1$), see Appendix A for details of the eigen-analysis of $\mathcal{K}_P$. The left panel of Figure 4.1 shows contours of the leading eigenfunction of the uncentered polynomial kernel operator for the bivariate normal example with $k = 2$. Similarly, it can be shown that the contours of the second leading eigenfunction form hyperbolas with their asymptotes given by the major and minor axes of the leading eigenfunction.

By contrast, the diagonal "central" moment matrix $\tilde{M}_2^2$ makes the leading eigenfunctions of the centered kernel operator $\tilde{\mathcal{K}}_p$ proportional to x_1^2, x_2^2 or x_1x_2. In particular, when $k > 1$, we have $\tilde{\phi}_1(\mathbf{x}) \propto x_2^2$, $\tilde{\phi}_2(\mathbf{x}) \propto x_1x_2$, and $\tilde{\phi}_3(\mathbf{x}) \propto x_1^2$; when $0 < k < 1$, $\tilde{\phi}_1(\mathbf{x}) \propto x_1^2$, $\tilde{\phi}_2(\mathbf{x}) \propto x_1x_2$, and $\tilde{\phi}_3(\mathbf{x}) \propto x_2^2$; when $k = 1$, the leading eigenvalue has multiplicity of 3, and thus x_1^2, x_2^2 and x_1x_2 span the corresponding eigenfunction space. The right panel of Figure 4.1 shows contours of the leading eigenfunction of the centered polynomial kernel operator for the same example. The result indicates that the centered kernel operator allows us to extract similar information about the direction of maximum variation as the uncentered version, however, with a different functional form.

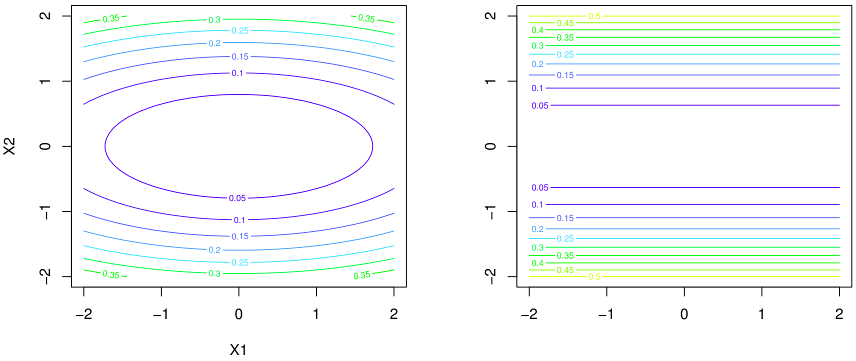


Figure 4.1: Comparison of the contours of the leading eigenfunction for the uncentered kernel operator (left) and centered kernel operator (right) in the bivariate normal example with $k = 2$.

4.1.3 Uncentered data with uncentered kernel

Now, we shift the center of the distribution from the origin along the x_1-axis by $m\sigma$. Correspondingly, we have

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} m\sigma \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & k\sigma^2 \end{pmatrix}\right).$$

This distribution has $E(X_1^2) = (m^2+1)\sigma^2, E(X_1^3) = (m^3+3m)\sigma^3, E(X_1^4) = (m^4+6m^2+3)\sigma^4$, $E(X_2^2) = k\sigma^2, E(X_2^3) = 0$, and $E(X_2^4) = 3k^2\sigma^4$. Given that X_1 and X_2 are independent, we obtain the following moment matrix:

$$M_2^2 = \begin{bmatrix} (m^4+6m^2+3)\sigma^4 & 0 & k(m^2+1)\sigma^4 \\ 0 & 2k(m^2+1)\sigma^4 & 0 \\ k(m^2+1)\sigma^4 & 0 & 3k^2\sigma^4 \end{bmatrix}.$$

The corresponding eigenvalues are $2k(m^2+1)\sigma^4$ and $\frac{1}{2}\left(3k^2\sigma^4 + (m^4+6m^2+3)\sigma^4 \pm \sqrt{\Delta}\right)$, where $\Delta = [3k^2\sigma^4 + (m^4+6m^2+3)\sigma^4]^2 - 8(k^2m^4 + 8k^2m^2 + 4k^2)\sigma^8$. The order of the three eigenvalues are determined by the relative size of k and m.

By changing m, we can see how the leading eigenfunction changes as we shift the center of the data distribution. To see the effect of moving m, we focus on the setting with $k = 2$ and $\sigma = 1$, where

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} m \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\right). \tag{4.6}$$

The corresponding moment matrix reduces to

$$M_2^2 = \begin{bmatrix} m^4+6m^2+3 & 0 & 2(m^2+1) \\ 0 & 4(m^2+1) & 0 \\ 2(m^2+1) & 0 & 12 \end{bmatrix}.$$

We now choose 6 different values of m ranging from 0 to 5 to observe qualitative changes in the contours of the leading eigenfunction.

Figure 4.2 shows the contours of the leading eigenfunctions of the uncentered kernel operator for the data distributions with six values of m. From the figure, we find that as the center of the data distribution moves away from the origin, contours of the leading eigenfunction change in shape. The contours for the data distribution with the center close to the origin form ellipses with the minor axis pointing to the maximum variation in the data. However, as the center moves away from the origin, the leading eigenfunction is more affected by the overall distance of data from the origin than the variation within the data distribution, with the minor axis along the x_1-axis.

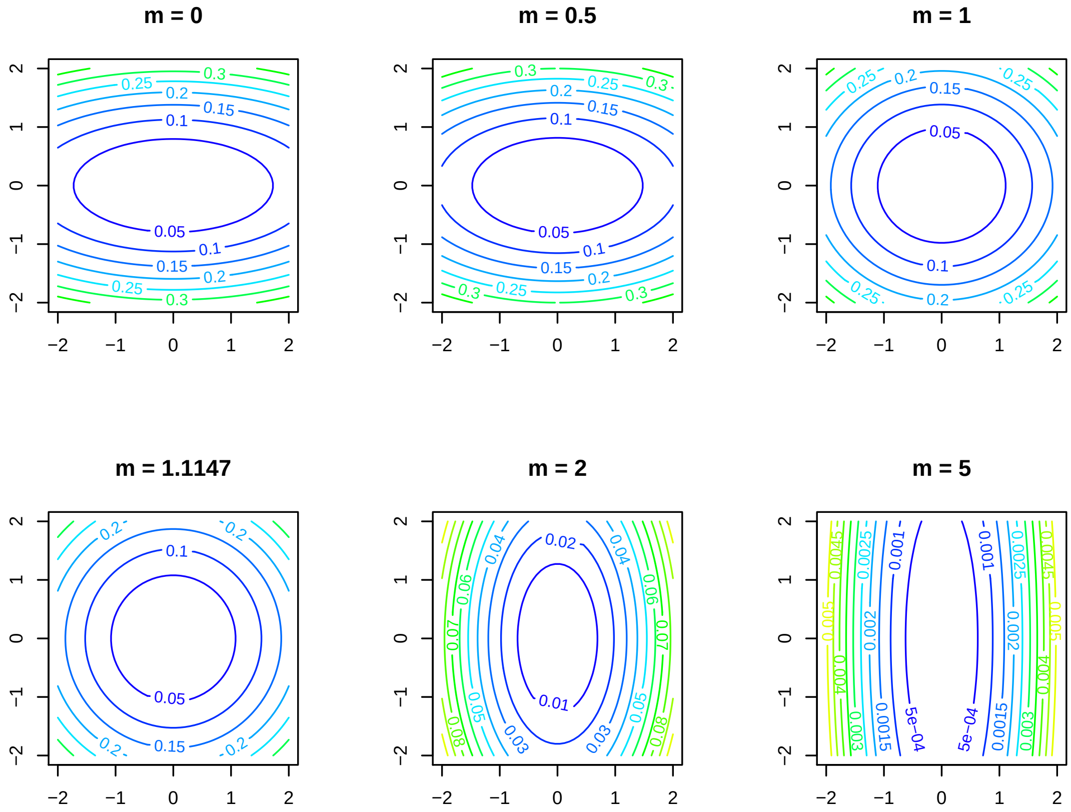


Figure 4.2: Contour plots of the leading eigenfunctions of the uncentered kernel operator for the distribution setting in (4.6) when the center of data distribution gradually moves along the x_1 axis away from the origin.

To see the effect of the relative size of m and k on the form of the leading eigenfunction, suppose $k = o(m^2)$. As m increases, the leading eigenvalue of the moment matrix is given by $\frac{1}{2}\left(3k^2\sigma^4 + (m^4 + 6m^2 + 3)\sigma^4 + \sqrt{\Delta}\right)$, which can be approximated by the dominating term $\frac{1}{2}m^4\sigma^4$. Besides, the moment matrix with $\sigma = 1$ can be approximated by,

$$M_2^2 \approx \begin{bmatrix} m^4 & 0 & km^2 \\ 0 & 2km^2 & 0 \\ km^2 & 0 & 3k^2 \end{bmatrix}.$$

Then the equations for the coefficients of eigenfunctions are approximately

$$\begin{bmatrix} m^4 & 0 & km^2 \\ 0 & 2km^2 & 0 \\ km^2 & 0 & 3k^2 \end{bmatrix} \begin{bmatrix} C_0 \\ C_1 \\ C_2 \end{bmatrix} \approx m^4 \begin{bmatrix} C_0 \\ C_1 \\ C_2 \end{bmatrix}$$

or

$$\begin{cases} m^4C_0 + km^2C_2 & \approx m^4C_0 \\ 2km^2C_1 & \approx m^4C_1 \\ km^2C_0 + 3k^2C_2 & \approx m^4C_2. \end{cases}$$

For large m, the approximate equations yield $C_1 \approx 0$ and $C_2 \approx 0$. Thus, the leading eigenfunction is approximately of the form $C_0x_1^2$, which captures the direction of the displacement of the data distribution from the origin.

4.1.4 Uncentered data with centered kernel

Consider the same bivariate normal data example as in Section 4.1.3, where the center of the distribution is shifted along x_1-axis by $m\sigma$.

For the centered kernel operator, the moment matrix can be obtained from (4.4) as follows:

$$\tilde{M}_2^2 = \begin{bmatrix} 2(2m^2+1)\sigma^4 & 0 & 0 \\ 0 & 2k(m^2+1)\sigma^4 & 0 \\ 0 & 0 & 2k^2\sigma^4 \end{bmatrix}.$$

Given that the moment matrix is diagonal, there is a trichotomy in the form of the leading eigenfunction:

- i) If $k^2 > 2m^2 + 1$ and $k > m^2 + 1$, the top eigenfunction $\tilde{\phi}_1(\mathbf{x}) \propto x_2^2$. Note that $k > m^2 + 1$ implies $k^2 > 2m^2 + 1$. Thus, as long as $k > m^2 + 1$, $\tilde{\phi}_1(\mathbf{x}) \propto x_2^2$.
- ii) If $k^2 < 2m^2+1$ and $k < \dfrac{2m^2+1}{m^2+1}$, the top eigenfunction $\tilde{\phi}_1(\mathbf{x}) \propto x_1^2$. Since $k < \dfrac{2m^2+1}{m^2+1}$ implies $k^2 < 2m^2 + 1$, as long as $k < \dfrac{2m^2+1}{m^2+1}$, $\tilde{\phi}_1(\mathbf{x}) \propto x_1^2$.
- iii) If $\dfrac{2m^2+1}{m^2+1} < k < m^2 + 1$, $\tilde{\phi}_1(\mathbf{x}) \propto x_1 x_2$.

Therefore, the relationship between k and m determines the shape of the leading eigenfunction. As we can see from the summary above, the effect of m is not the same as what is observed for the uncentered kernel in Section 4.1.3. Figure 4.3 depicts the trichotomy of the form of the top eigenfunction depending on the relationship between k and m. As we can see from the plot, k has to be less than 2 for the eigenfunction to be of the form x_1^2. When k is greater or equal to 2, the form of the top eigenfunction is either x_1x_2 or x_2^2 depending on m.

For instance, when $k = 2$, the leading eigenfunction is of the form x_2^2 if $m \leq 1$ while it is of the form x_1x_2 otherwise.

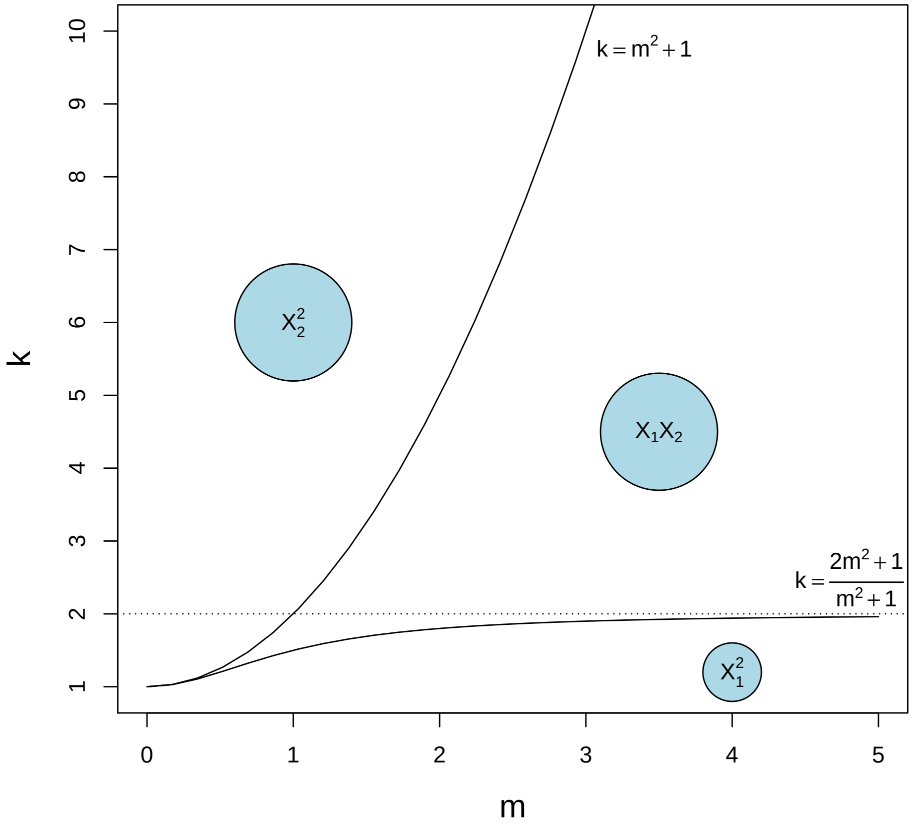


Figure 4.3: The trichotomy of the leading eigenfunction form for the uncentered data distribution case in (4.6) with centered kernel.

When $k = 1.5$, as m changes, the form of the leading eigenfunction changes from x_2^2 to x_1x_2, and then to x_1^2. Figure 4.4 illustrates the expected pattern.

Recall that for the uncentered counterpart in section 4.1.3, the transition in the leading eigenfunction with m is not in the functional form but in the orientation of the corresponding ellipses. Clearly, centering a kernel changes the spectrum of the kernel operator and the mode of change depends on the form of the kernel.

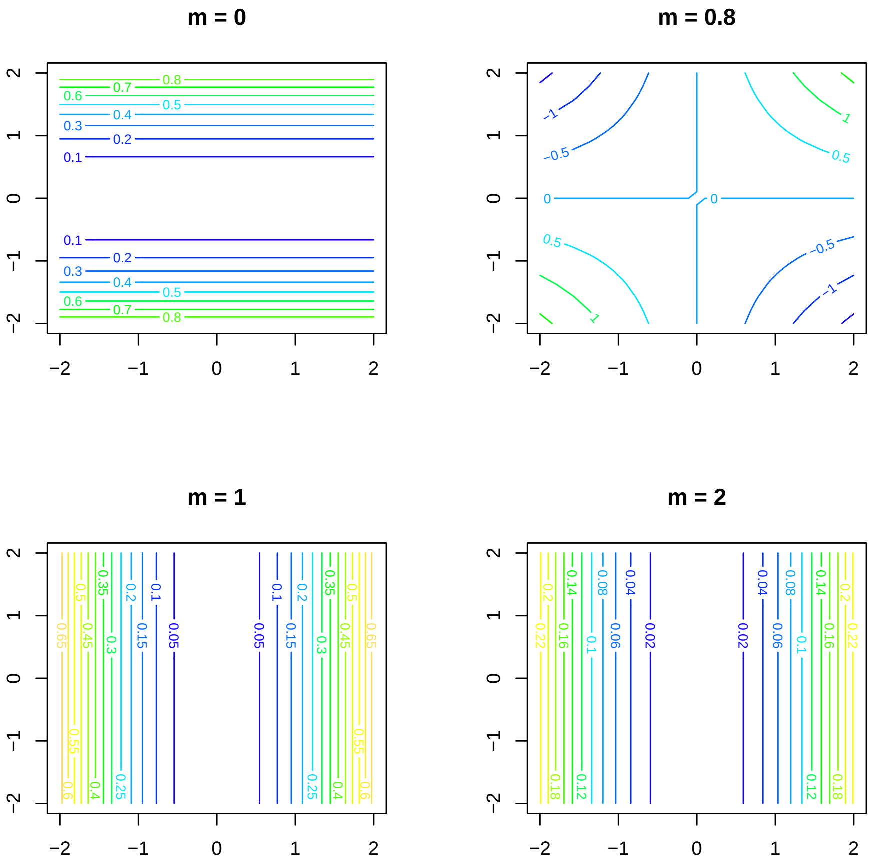


Figure 4.4: Contours of the leading eigenfunction as m increases when $k = 1.5$ for the uncentered data distribution in (4.6) with centered kernel operator.

4.1.5 Extension

In the examples above, we have only considered the displacement of the center of distribution along the x_1-axis for simplicity. In order to see more general transition in the contours of the leading eigenfunction, we extend the numerical analysis to the case where the data distribution has its center at $(m\sigma, m\sigma)$ as follows:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} m\sigma \\ m\sigma \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & k\sigma^2 \end{pmatrix} \right).$$

The moment matrix for the uncentered polynomial kernel of degree 2 is:

$$M_2^2 = \begin{bmatrix} (m^4+6m^2+3)\sigma^4 & \sqrt{2}(m^4+3m^2)\sigma^4 & (m^2+1)(m^2+k)\sigma^4 \\ \sqrt{2}(m^4+3m^2)\sigma^4 & 2(m^2+1)(m^2+k)\sigma^4 & \sqrt{2}(m^4+3km^2)\sigma^4 \\ (m^2+1)(m^2+k)\sigma^4 & \sqrt{2}(m^4+3km^2)\sigma^4 & (m^4+6km^2+3k^2)\sigma^4 \end{bmatrix}.$$

It is analytically more involved to find the exact form of eigenfunctions given the moment matrix. However, based on the discussion in the previous sections, we conjecture that as the center of the data distribution moves away from the origin, contours of the leading eigenfunction would capture the direction of displacement from the origin instead of the maximum variation in the data distribution. Figure 4.5 shows the change in the contour plots of the leading eigenfunction as we increase m when $k = 2, \sigma^2 = 1$. We see that the contours are affected by both the center of the distribution and the variation within the distribution. As the center moves away from the origin along the 45 degree line, the contours of the leading eigenfunction change their orientation from the maximum variation within the distribution (x_2-axis) to the displacement of the center ($x_2 = x_1$ direction).

4.2 General Eigen-analysis for Centered Kernel Operator

4.2.1 Derivation of eigenfunction-eigenvalue pair

We study the effect of centering a kernel on the spectrums of the kernel operator in general using Mercer-Hilbert-Schmidt theorem (Wahba (1990)). Analogous to the explicit expansion of the polynomial kernel, the theorem states that for a

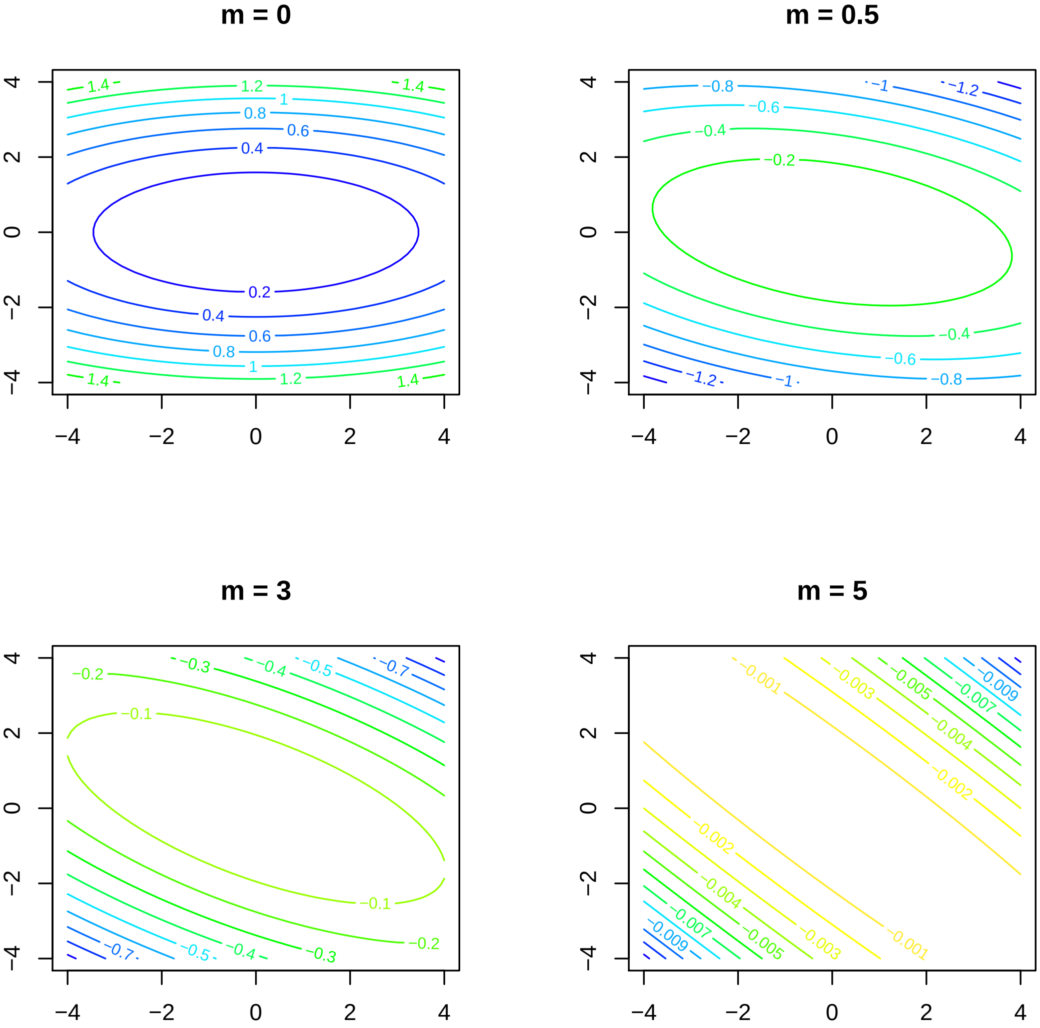


Figure 4.5: Contours of the leading eigenfunction when the center of the data distribution moves from the origin along the 45 degree line.

continuous symmetric non-negative definite kernel K, there exists an orthonormal sequence of continuous eigenfunctions, ϕ_i and eigenvalues $\lambda_i \geq 0$ of the kernel operator such that

$$K(x,y) = \sum_i \lambda_i \phi_i(x) \phi_i(y).$$

Using the eigen-expansion of K, we examine the spectrum of the centered kernel operator $\mathcal{K}_p$. For the centered kernel operator, we have already shown that $\tilde{\phi}_0(x) = 1$ and $\tilde{\lambda}_0 = 0$ form an eigenvalue-eigenfunction pair. This simplifies the eigen-equation for the centered kernel operator $\mathcal{K}$ with non-zero eigenvalues as follows:

$$\int (K(x,y) - E_Y[K(x,Y)])\tilde{\phi}(x)p(x)dx = \lambda \tilde{\phi}(y). \tag{4.7}$$

With the expression of $K(x,y) = \sum_i \sqrt{\lambda_i}\phi_i(x)\sqrt{\lambda_i}\phi_i(y)$ and $E_Y[K(x,Y)] = \sum_i \sqrt{\lambda_i}\phi_i(x)E_Y[\sqrt{\lambda_i}\phi_i(Y)]$, the equation (4.7) is written as,

$$\int (\sum_i \sqrt{\lambda_i}\phi_i(x)\sqrt{\lambda_i}\phi_i(y) - \sum_i \sqrt{\lambda_i}\phi_i(x)E_Y[\sqrt{\lambda_i}\phi_i(Y)])\tilde{\phi}(x)p(x)dx = \lambda\tilde{\phi}(y),$$

which reduces to

$$\sum_i \sqrt{\lambda_i}(\phi_i(y) - E_Y[\phi_i(Y)]) \int \sqrt{\lambda_i}\phi_i(x)\tilde{\phi}(x)p(x)dx = \lambda\tilde{\phi}(y).$$

Letting C_i denote $\int \sqrt{\lambda_i}\phi_i(x)\tilde{\phi}(x)p(x)dx$, we have

$$\sum_i C_i\sqrt{\lambda_i}\left(\phi_i(y) - E_Y[\phi_i(Y)]\right) = \lambda\tilde{\phi}(y).$$

This leads to the eigenfunction of the form,

$$\tilde{\phi}(y) = \frac{1}{\lambda}\sum_i C_i\sqrt{\lambda_i}\left(\phi_i(y) - E_Y[\phi_i(Y)]\right).$$

By plugging this form into the expression for the constant C_i, we have

$$\int \sqrt{\lambda_j}\phi_j(x)\left[\sum_i C_i\sqrt{\lambda_i}\left(\phi_i(x) - E_X[\phi_i(X)]\right)\right]p(x)dx = \lambda C_j,$$

which equates to

$$\sum_i \sqrt{\lambda_i}\sqrt{\lambda_j}\left[\int \phi_j(x)\phi_i(x)p(x)dx - E_X[\phi_i(X)]E_X[\phi_j(X)]\right] C_i = \lambda C_j. \quad (4.8)$$

Let $\mu_j = E_X[\phi_j(X)]$, and note that $\displaystyle\int \phi_i(x)\phi_j(x)p(x)dx = \langle\phi_i(x), \phi_j(x)\rangle_p = \delta_{ij}$ from the orthogonality of $\{\phi_j\}$.

Hence, λ and $\tilde{\phi}$ are determined by the eigenvalue-eigenvector pair of the matrix $\tilde{M}$ in general whose ij-th entry is

$$\sqrt{\lambda_i}\sqrt{\lambda_j}(\langle\phi_i(x), \phi_j(x)\rangle_p - \mu_i\mu_j).$$

Here $\tilde{M}$ is taken as a linear operator defined on the l^2 space. If we let $\bar{\phi}_i(x) = \phi_i(x) - \mu_i$, then

$$\begin{aligned}&\langle\bar{\phi}_i(x), \bar{\phi}_j(x)\rangle_p = \langle\phi_i(x) - \mu_i, \phi_j(x) - \mu_j\rangle_p \\ =&\langle\phi_i(x), \phi_j(x)\rangle_p - \mu_i\mu_j.\end{aligned}$$

Therefore, the ijth entry of matrix $\tilde{M}$ can also be written as $\sqrt{\lambda_i}\sqrt{\lambda_j}\langle\bar{\phi}_i(x), \bar{\phi}_j(x)\rangle_p$. Then the diagonals of $\tilde{M}$ are $\lambda_j(1-\mu_j^2)$ and the off-diagonals $-\sqrt{\lambda_i}\sqrt{\lambda_j}\mu_i\mu_j$. That is,

$$\tilde{M} = \begin{bmatrix} \lambda_1(1-\mu_1^2) & -\sqrt{\lambda_1}\sqrt{\lambda_2}\mu_1\mu_2 & \dots \\ -\sqrt{\lambda_2}\sqrt{\lambda_1}\mu_2\mu_1 & \lambda_2(1-\mu_2^2) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix},$$

and equation (4.8) can be written as

$$\tilde{M}\begin{bmatrix} C_1 \\ C_2 \\ \vdots \end{bmatrix} = \lambda \begin{bmatrix} C_1 \\ C_2 \\ \vdots \end{bmatrix}.$$

The eigenvectors of $\tilde{M}$ determine the coefficients of the eigenfunctions of centered kernel operator.

In summary, we conclude that the eigen-analysis of the centered kernel operator can be done systematically based on the explicit eigenvalues and eigenfunctions of the uncentered kernel operator.

4.2.2 Orthogonality of eigenfunctions

We elaborate on the way the eigenfunctions of $\tilde{K}$ are characterized through the matrix $\tilde{M}$, and verify their properties.

For $\mathbb{C} = (C_i)_{i=1}^{\infty} \in l^2$, we can define a function $f(x) = \sum_{i=1}^{\infty} C_i \sqrt{\lambda_i} \bar{\phi}_i(x)$ properly. Note that

$$\|f\|_p^2 = \sum_{i,j} C_i C_j \sqrt{\lambda_i}\sqrt{\lambda_j} \langle \bar{\phi}_i, \bar{\phi}_j \rangle_p \leq \sum_{i,j} |C_i||C_j| \sqrt{\lambda_i}\sqrt{\lambda_j} \|\bar{\phi}_i\| \|\bar{\phi}_j\|$$
$$=(\sum_i |C_i| \sqrt{\lambda_i} \|\bar{\phi}_i\|_p)^2 \leq (\sum_i |C_i| \sqrt{\lambda_i})^2 < \infty.$$

Suppose that the ith eigenvector of $\tilde{M}$ is $\mathbf{C}_i = (C_{i1}, C_{i2}, \cdots)$ and the jth eigenvector is $\mathbf{C}_j = (C_{j1}, C_{j2}, \cdots)$, and they are orthogonal in the l^2 sense, $\langle \mathbf{C}_i, \mathbf{C}_j \rangle = \sum_{k=1}^{\infty} C_{ik} C_{jk} = 0$. Here we show that the eigenfunctions $\tilde{\phi}_i$ and $\tilde{\phi}_j$ defined by $\mathbf{C}_i$ and $\mathbf{C}_j$ are orthogonal $\langle \tilde{\phi}_i, \tilde{\phi}_j \rangle_p = 0$ for $i \neq j$.

The inner product of the eigenfunctions is

$$\langle \tilde{\phi}_i, \tilde{\phi}_j \rangle_p = \langle \frac{1}{\lambda_i} \sum_k C_{ik} \sqrt{\lambda_k} \bar{\phi}_k, \frac{1}{\lambda_j} \sum_k C_{jk} \sqrt{\lambda_k} \bar{\phi}_k \rangle_p$$
$$=\frac{1}{\lambda_i \lambda_j} \sum_{k,k'} C_{ik} C_{jk'} \sqrt{\lambda_k} \sqrt{\lambda_{k'}} \langle \bar{\phi}_k, \bar{\phi}_{k'} \rangle_p.$$

Let the $kk^{'}$th entry of matrix $\tilde{M}$, $\tilde{M}_{kk'} = \sqrt{\lambda_k}\sqrt{\lambda_{k'}}\langle\bar{\phi}_k, \bar{\phi}_{k'}\rangle_p$. Then the inner product is expressed as

$$\langle\tilde{\phi}_i, \tilde{\phi}_j\rangle_p = \frac{1}{\lambda_i\lambda_j}\sum_k C_{ik}\sum_{k'} C_{jk'}\tilde{M}_{kk'} = \frac{1}{\lambda_i\lambda_j}\langle\mathbf{C}_i, \tilde{M}\mathbf{C}_j\rangle.$$

Since $\tilde{M}\mathbf{C}_j = \lambda_j\mathbf{C}_j$,

$$\langle\tilde{\phi}_i, \tilde{\phi}_j\rangle_p = \frac{1}{\lambda_i}\langle\mathbf{C}_i, \mathbf{C}_j\rangle = 0.$$

The orthogonality condition for the eigenfunctions is therefore verified.

4.2.3 Connection with the centered polynomial kernel operator result

The discussion so far regards the general case of any kernel operator. In this section, we make the connection between the general results and the specific result for the polynomial kernel operator in Section 4.1.

The eigen-analysis for the centered polynomial kernel in a two-dimensional scenario has been investigated in Section 4.1, where $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^t\mathbf{y})^d = (x_1y_1 + x_2y_2)^d = \sum_{j=0}^{d}\binom{d}{j}(x_1y_1)^{d-j}(x_2y_2)^j$. Recall that the polynomial kernel can be expressed as the inner product of feature vectors $\Phi(\mathbf{x})^t\Phi(\mathbf{y})$ explicitly with the feature vector $\Phi(\mathbf{x}) = \left(\binom{d}{0}^{\frac{1}{2}}x_1^d, \cdots, \binom{d}{d}^{\frac{1}{2}}x_2^d\right)$. Here the features are fixed regardless of the data distribution.

On the other hand, Mercer's theorem allows us to express the kernel function generally as $K(x, y) = \sum_k \lambda_k\phi_k(x)\phi_k(y) = \sum_k \sqrt{\lambda_k}\phi_k(x)\sqrt{\lambda_k}\phi_k(y)$, which can be alternatively written as $\Phi(\mathbf{x})^t\Phi(\mathbf{y})$, with $\Phi(\mathbf{x}) = (\sqrt{\lambda_1}\phi_1(x), \sqrt{\lambda_2}\phi_2(x), \cdots)$. Despite the difference, the general form of each entry of $\tilde{M}$ applies to yield the same “central” moment matrix we have obtained for the centered polynomial kernel operator, alternatively written as

$$\begin{aligned}
&\sqrt{\lambda_i}\sqrt{\lambda_j}\langle\bar{\phi}_i,\bar{\phi}_j\rangle_p\\
=&\binom{d}{i}^{\frac{1}{2}}\binom{d}{j}^{\frac{1}{2}}\langle x_1^{d-i}x_2^i - E(X_1^{d-i}X_2^i), x_1^{d-j}x_2^j - E(X_1^{d-j}X_2^j)\rangle_p\\
=&\binom{d}{i}^{\frac{1}{2}}\binom{d}{j}^{\frac{1}{2}}\left\{E(X_1^{2d-i-j}X_2^{i+j}) - E(X_1^{d-i}X_2^i)E(X_1^{d-j}X_2^j)\right\}
\end{aligned}$$

where $i = 0, \cdots, d$ and $j = 0, \cdots, d$.

4.3 Analysis of Centered Gaussian Kernel Operator

Using the general framework laid out for the spectral analysis of a centered kernel operator, we examine the effect of centering for the Gaussian kernel. Zhu et al. (1998), Williams and Seeger (2000) and Shi et al. (2009) study the uncentered Gaussian kernel operator and we base our analysis on the theoretical result for the uncentered Gaussian kernel operator.

4.3.1 Centered Gaussian kernel operator

For simplicity, suppose that X is distributed with $N(\mu, \sigma^2)$. For the Gaussian kernel with a bandwidth parameter w, $K(x,y) = e^{-\frac{(x-y)^2}{2w^2}}$, the eigenvalues and eigenfunctions of the kernel operator are available, see Shi et al. (2009) for example. The population version of centered Gaussian kernel function in this setting can be obtained explicitly using the equation (4.1) as follows. From the fact that

$$E_X[K(X,y)] = \int e^{-\frac{(x-y)^2}{2w^2}}p(x)dx = \sqrt{\frac{w^2}{w^2+\sigma^2}}e^{-\frac{(y-\mu)^2}{2(w^2+\sigma^2)}},$$

where $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, we have

$$E_Y[K(x,Y)] = \sqrt{\frac{w^2}{w^2+\sigma^2}} e^{-\frac{(x-\mu)^2}{2(w^2+\sigma^2)}}$$

and

$$\begin{aligned} E_X E_Y[K(X,Y)] &= \frac{1}{\sqrt{2\pi\sigma^2}} \sqrt{\frac{w^2}{w^2+\sigma^2}} \int e^{-\frac{(x-\mu)^2}{2(w^2+\sigma^2)}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \sqrt{\frac{w^2}{w^2+2\sigma^2}}. \end{aligned}$$

Therefore, the centered Gaussian kernel function is expressed as

$$\tilde{K}(x,y) = e^{-\frac{(x-y)^2}{2w^2}} - \sqrt{\frac{w^2}{w^2+\sigma^2}} e^{-\frac{(y-\mu)^2}{2(w^2+\sigma^2)}} - \sqrt{\frac{w^2}{w^2+\sigma^2}} e^{-\frac{(x-\mu)^2}{2(w^2+\sigma^2)}} + \sqrt{\frac{w^2}{w^2+2\sigma^2}}.$$

We carry out a simulation study to illustrate the effect of centering for the Gaussian kernel operator in the following.

4.3.2 One-component normal

Simulation

We generate data of size 1000 from normal distribution $N(2, 1^2)$. For the Gaussian kernel function with the bandwidth $w = 1.5$, we then obtained the eigenvectors of the uncentered Gaussian kernel matrix using the data. Similarly, we obtained the eigenvectors of the centered Gaussian kernel matrix for comparison.

The first row of Figure 4.6 shows the eigenvectors for the uncentered kernel matrix while the second row shows those for the centered kernel matrix. The comparison between the two rows suggests that there is a shift in the spectrum after centering, which in turn results in some change in the information about the distribution carried by the eigenvectors. For example, while the first eigenvector of the uncentered kernel matrix has its peak at 2, the mode of the distribution, such information is lost in the first eigenvector of the centered kernel matrix. But the second eigenvector of the centered kernel matrix seems to pick up this information.

Given that it is customary to focus on the leading eigenvector, the uncentered kernel seems to capture this key information better in its first eigenvector than the centered kernel. We would like to provide explanations for the change in the spectrum.

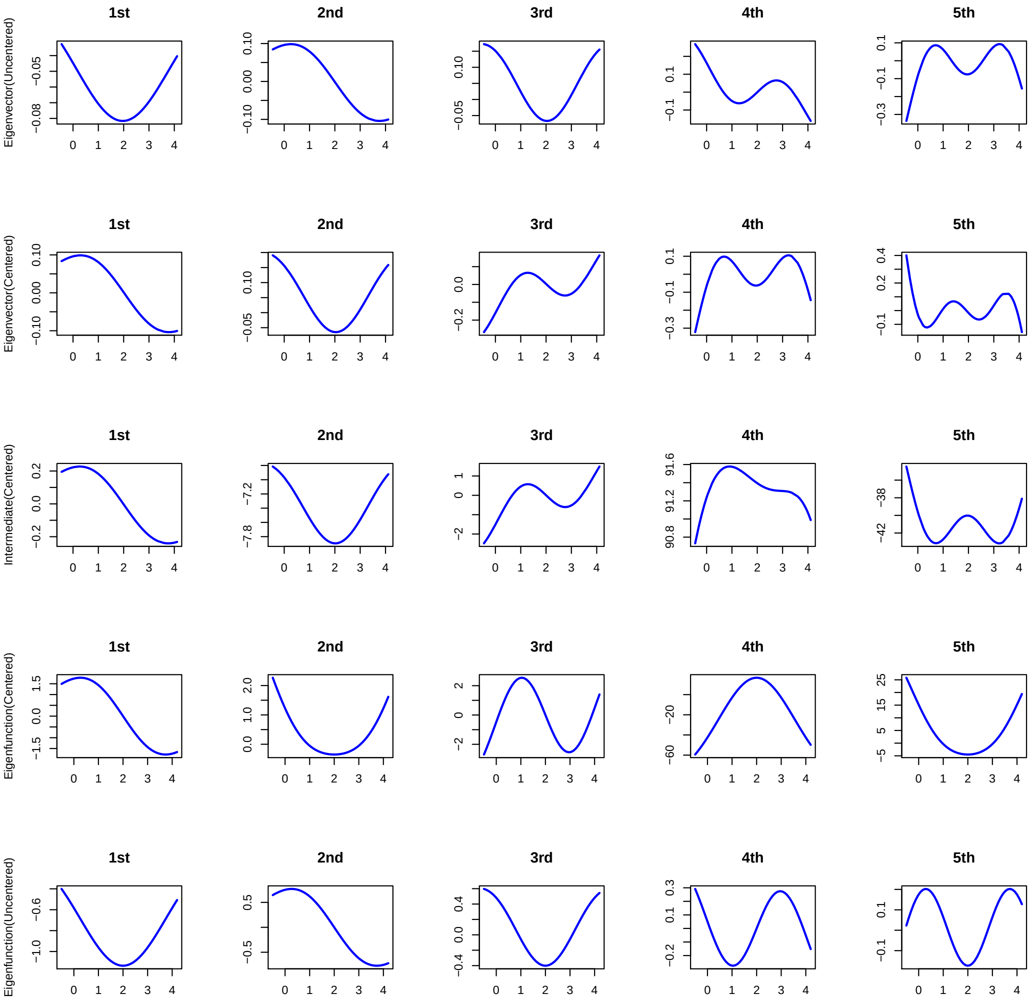


Figure 4.6: The first row shows the first five eigenvectors of an uncentered Gaussian kernel matrix with the bandwidth $w = 1.5$ for data sampled from normal distribution $N(2, 1^2)$ and the second row shows the first five eigenvectors of the centered kernel matrix. The third row shows the linear combinations of eigenvectors with the coefficients derived from our analysis; the fourth row shows the first five theoretical eigenfunctions for the centered kernel operator; the fifth row shows the five eigenfunctions for the uncentered kernel operator.

Preliminary explanation for the shifted spectrum

Since $\lambda_0 = 0$ is the eigenvalue corresponding to the eigenfunction $\tilde{\phi}_0(x) = 1$ for a centered kernel operator in general, any eigenfunction $\tilde{\phi}$ with a nonzero eigenvalue has a zero expected value, $\int \tilde{\phi}(x)\tilde{\phi}_0(x)p(x)dx = \int \tilde{\phi}(x)p(x)dx = 0$. Therefore, for the Gaussian kernel $K(x,y) = e^{-\frac{(x-y)^2}{2w^2}}$ and $\tilde{\phi}$,

$$\tilde{\mathcal{K}}_p\tilde{\phi} = \int \{e^{-\frac{(x-y)^2}{2w^2}} - \sqrt{\frac{w^2}{w^2+\sigma^2}}e^{-\frac{(x-\mu)^2}{2(w^2+\sigma^2)}}\}\tilde{\phi}(x)p(x)dx = \lambda\tilde{\phi}.$$

To see the connection between $\tilde{\mathcal{K}}_p$ and $\mathcal{K}_p$, re-express

$$\tilde{\mathcal{K}}_p\tilde{\phi} = \mathcal{K}_p\tilde{\phi} - \sqrt{\frac{w^2}{w^2+\sigma^2}}\langle\phi_*, \tilde{\phi}\rangle_p,$$

where $\phi_*(x) = e^{-\frac{(x-\mu)^2}{2(w^2+\sigma^2)}}$.

Shi et al. (2009) and Williams and Seeger (2000) derived the eigenvalue and eigenfunction of the Gaussian kernel operator $\mathcal{K}_p$ as follows:

$$\lambda_i = \sqrt{\frac{2}{(1+\beta+\sqrt{1+2\beta})}}\left(\frac{\beta}{1+\beta+\sqrt{1+2\beta}}\right)^{i-1}$$

$$\phi_i(x) = \frac{(1+2\beta)^{1/8}}{\sqrt{2^{i-1}(i-1)!}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\frac{\sqrt{1+2\beta}-1}{2}\right)H_{i-1}\left(\left(\frac{1}{4}+\frac{\beta}{2}\right)^{1/4}\frac{x-\mu}{\sigma}\right),$$

for $i = 1, 2, \cdots$, where $\beta = 2\sigma^2/w^2$, and H_i is the ith order Hermite polynomial. In particular, the first eigenfunction is

$$\phi_1(x) = (1+4\sigma^2/w^2)^{1/8}e^{-\frac{(x-\mu)^2}{4\sigma^2}(\sqrt{1+\frac{4\sigma^2}{w^2}}-1)}.$$

If ϕ_* is approximately proportional to ϕ_1, then $\langle\phi_*, \phi_i\rangle_p \approx 0$ for $i = 2, 3, \cdots$ and $\tilde{\mathcal{K}}_p\phi_i \approx \mathcal{K}_p\phi_i = \lambda_i\phi_i$ for $i = 2, 3, \cdots$. The supposed relation between ϕ_* and ϕ_1 may provide a plausible explanation for the observed shift in the spectrum after centering the Gaussian kernel. We can show that ϕ_* and ϕ_1 are somehow similar by proving that the inner product of those two functions is approximately 1 for certain range of w.

To verify the hypothesis, we compute $\langle \phi_1, \frac{\phi_*}{\|\phi_*\|}\rangle_p$, which represents the cosine of the angle between the two functions. A simple calculation shows that the norm of ϕ_* is $\|\phi_*\| = \sqrt{\|\phi_*\|^2} = \sqrt{\int \phi_*^2(x)p(x)dx} = (\frac{w^2+\sigma^2}{w^2+3\sigma^2})^{1/4} = (\frac{w^2+1}{w^2+3})^{1/4}$ given $\sigma^2 = 1$ and the inner product of $\langle \phi_1, \frac{\phi_*}{\|\phi_*\|}\rangle$ is

$$\begin{aligned}\langle \phi_1, \frac{\phi_*}{\|\phi_*\|}\rangle &= \int \phi_1(x)\frac{\phi_*}{\|\phi_*\|}p(x)dx \\ &= \frac{(1+\frac{4}{w^2})^{1/8}}{\sqrt{2\pi}(\frac{w^2+1}{w^2+3})^{1/4}} \int e^{-\frac{(x-\mu)^2}{2}[\frac{\sqrt{1+4/w^2}-1}{2}+\frac{1}{w^2+1}+1]}dx \\ &= \frac{(1+\frac{4}{w^2})^{1/8}}{(\frac{w^2+1}{w^2+3})^{1/4}\sqrt{\triangle}},\end{aligned}$$

where $\triangle = \frac{\sqrt{1+4/w^2}-1}{2} + \frac{1}{w^2+1} + 1$. Figure 4.7 shows the values of the inner product for a range of the values of the bandwidth w from 0.1 to 3. As w increases, the inner product increases from the value close to .96 to 1. Since the inner product of 1 means that the two functions are identical, this suggests that $\phi_* \propto \phi_1$ approximately for a wide range of w, and thus $\langle \phi_*, \phi_i\rangle_p \approx 0$ for $i = 2, 3, \cdots$

Connection with our theory

The preliminary explanation hints that the key to more elaborate analysis lies in the expansion of $E_Y[K(x,Y)]$ in terms of ϕ_i. From Mercer's theorem, we have $K(x,y) = \sum_k \lambda_k\phi_k(x)\phi_k(y)$, and $E_Y[K(x,Y)] = E_Y[\sum_k \lambda_k\phi_k(x)\phi_k(Y)] = \sum_k \lambda_k\mu_k\phi_k(x)$.

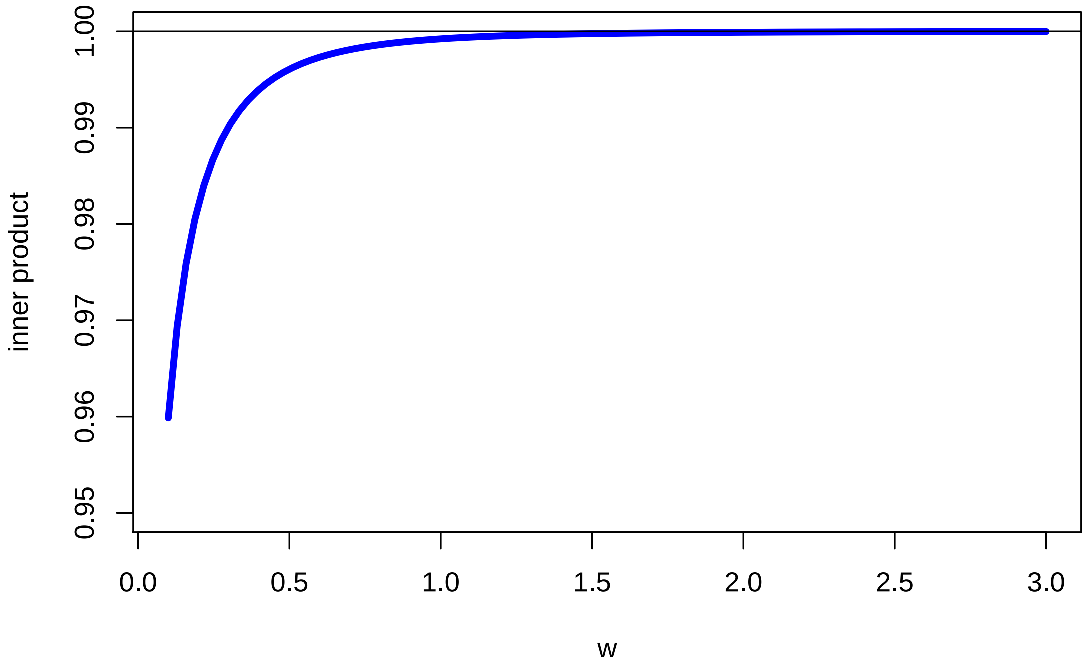


Figure 4.7: The inner product $\langle \phi_1, \dfrac{\phi_*}{\|\phi_*\|} \rangle_p$ versus the value of w.

Suppose $\mu_k \approx 0$ for $k \notin I$ and $|I| < \infty$. Then $E_Y K(x, Y) \approx \sum_{k \in I} \lambda_k \mu_k \phi_k(x)$. Accordingly, we can approximate the ijth entry of matrix $\tilde{M}$ as

$$\tilde{M}_{ij} \approx \begin{cases} \sqrt{\lambda_i}\sqrt{\lambda_j}\delta_{ij} & i \notin I \text{ or } j \notin I, \\ \sqrt{\lambda_i}\sqrt{\lambda_j}(\delta_{ij} - \mu_i\mu_j) & i \in I \text{ and } j \in I \end{cases}$$

$\tilde{M}$ is a matrix with block structure which can be written as the direct sum of A and B, $A \bigoplus B$, where A has its ijth entry $\sqrt{\lambda_i}\sqrt{\lambda_j}(\delta_{ij} - \mu_i\mu_j)$ for $i, j \in I$ while B is a diagonal matrix with values λ_i for $i \notin I$.

Clearly, for $i \notin I$, the eigenvectors of $\tilde{M}$ are $e_i's$ with eigenvalues $\tilde{\lambda}_i = \lambda_i$.

Further from spectral decomposition of A, $Av = \lambda v$, with an eigenvector $v \in R^{|I|}$ of A, we have

$$\left[\begin{array}{c:c} A & 0 \\ \hdashline 0 & B \end{array}\right] \begin{bmatrix} v \\ 0 \end{bmatrix} = \lambda \begin{bmatrix} v \\ 0 \end{bmatrix}.$$

Therefore, the corresponding eigenfunctions for the centered kernel operator have the following form approximately:

$$\begin{aligned} &\tilde{\phi}_i \propto \sqrt{\lambda_i}\bar{\phi}_i, i \notin I \\ \text{and} \quad &\tilde{\phi}_i \propto \sum_{k \in I} v_{ik}\sqrt{\lambda_k}\bar{\phi}_k, i \in I, \end{aligned}$$

where $v_i, i = 1, \cdots, |I|$ are the eigenvectors of A. Note that depending on the eigenvalues of A, the spectrum of $\mathcal{K}_p$ may need to be reordered.

In the following numerical analysis, we apply the approximation of $E_Y[K(x,Y)] = \sum_k \lambda_k \mu_k \phi_k(x)$ to the univariate normal example keeping the eigenfunctions with dominating $\lambda_k \mu_k$. Using the explicit eigenfunctions and eigenvalues of the uncentered Gaussian kernel operator, we have the eigenvalues and expected values of the first five dominating eigenfunctions as follows:

$$\begin{aligned} &\lambda_1 = .75, \lambda_2 = .1875, \lambda_3 = .0469, \lambda_4 = .0117, \lambda_5 = .0029 \\ &\mu_1 = .984, \mu_2 = 0, \mu_3 = -.1305, \mu_4 = 0, \mu_5 = .0212. \end{aligned}$$

Notice that $\mu_2 = \mu_4 = 0$ and in fact, we can show that $\mu_j = 0$ for even j in this case. If we simply write the eigenfunction of the uncentered kernel operator as

$$\phi_j(x) = C_1 \exp(-C_2 \frac{(x-\mu)^2}{2\sigma^2}) H_{j-1}(C_3 \frac{x-\mu}{\sigma}),$$

for some constants C_1, C_2 and C_3. We have

$$\phi_j(x+\mu) = C_1 \exp(-C_2 \frac{x^2}{2\sigma^2}) H_{j-1}(C_3 \frac{x}{\sigma}),$$

which is odd for even j. Then

$$\mu_j = E[\phi_j(X)] = \int \phi_j(x) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

is the integral of an odd function for even j. Therefore, we have $\mu_j = E[\phi_j(X)] = 0$ for even j.

With the approximation of $E_Y[K(x, Y)]$ using the first five eigenfunctions ϕ_k, we can express the 5×5 top left block of $\tilde{M}$, where the diagonal values are $\lambda_i(1-\mu_i)^2$ and the non-diagonal values are $-\sqrt{\lambda_i}\sqrt{\lambda_j}\mu_i\mu_j$,

$$A = \begin{bmatrix} .0238 & 0 & .0241 & 0 & .00097 \\ 0 & .1875 & 0 & 0 & 0 \\ .0241 & 0 & .0461 & 0 & 0 \\ 0 & 0 & 0 & .0117 & 0 \\ .00097 & 0 & 0 & 0 & .0029 \end{bmatrix}.$$

From the eigen-decomposition of this matrix, we can identify the eigenfunctions of the centered Gaussian kernel operator as follows:

$$\begin{aligned} \tilde{\phi}_1(x) &\propto -.4330\phi_2(x) \\ \tilde{\phi}_2(x) &\propto -.4665(\phi_1(x) - .984) - .1825(\phi_3(x) + .1305) \\ \tilde{\phi}_3(x) &\propto .1082\phi_4(x) \\ \tilde{\phi}_4(x) &\propto .7214(\phi_1(x) - .984) + -.1157(\phi_3(x) + .1305) + .0078(\phi_5(x) - .0212) \\ \tilde{\phi}_5(x) &\propto -.1091(\phi_1(x) - .984) + .0152(\phi_3(x) + .1305) + .0539(\phi_5(x) - .0212). \end{aligned} \tag{4.9}$$

The first five eigenvalues of the centered Gaussian kernel operator are $\tilde{\lambda}_1 = .1875$, $\tilde{\lambda}_2 = .0615$, $\tilde{\lambda}_3 = .0117$, $\tilde{\lambda}_4 = .0085$, $\tilde{\lambda}_5 = .0028$.

The fourth row of Figure 4.6 shows the first five eigenfunctions for the centered kernel operator in (4.9). We notice that the first three eigenfunctions match well the eigenvectors of the centered kernel matrix. The third eigenfunction picks up the peak suggested in the third eigenvector of the centered kernel matrix. In general, numerical discrepancy between eigenvectors and eigenfunctions is expected especially for the pairs with smaller eigenvalues.

To make closer comparison between eigenfunctions and eigenvectors, we approximate the eigenfunctions for the centered kernel operator in an intermediate way such that they are linear combination of eigenvectors instead of eigenfunctions of the uncentered kernel operator, with the same coefficients as in equations 4.9. This is shown in the third row of Figure 4.6. The corresponding figures show more agreement in terms of shape with the second row compared to the fourth row.

The fifth row of Figure 4.6 shows the theoretical eigenfunctions of the uncentered kernel operator, which can be compared directly with the first row of Figure 4.6. While the first three pairs are pretty similar, the fourth and fifth differ in terms of shape to some extent, however, with the information of the peaks retained.

It should be noted that the derivation of the eigenfunctions of the centered kernel operator involves approximation. We selected the first five eigenfunctions with relatively large coefficients compared to other eigenfunctions in this analysis. The result would change as we include more eigenfunctions in the approximation.

4.3.3 Mixture normal distribution

Simulation

We investigate the mixture distribution of two normal components $N(2, 1^2)$ and $N(-2, 1^2)$. The mixture proportions are set to $\pi_1 = .6$ and $\pi_2 = .4$. Shi et al. (2009) also analyzed the spectrum of the uncentered kernel operator for mixture normal distributions and found that when there is enough separation among the components, the leading eigenfunctions of the kernel operator come from the leading ones of the kernel operator for different components and the order is determined by the mixture proportions and eigenvalues.

When the data distribution is $\pi_1 N(\mu_1, \sigma^2) + \pi_2 N(\mu_2, \sigma^2)$, we have the mixture density function as

$$p(x) = \pi_1 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} + \pi_2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}.$$

Therefore, each part of the centered kernel function will be given as follows,

$$E_X[K(X,y)] = \pi_1 \sqrt{\frac{w^2}{w^2+\sigma^2}} e^{-\frac{(y-\mu_1)^2}{2(w^2+\sigma^2)}} + \pi_2 \sqrt{\frac{w^2}{w^2+\sigma^2}} e^{-\frac{(y-\mu_2)^2}{2(w^2+\sigma^2)}}$$

and

$$E_Y[K(x,Y)] = \pi_1 \sqrt{\frac{w^2}{w^2+\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2(w^2+\sigma^2)}} + \pi_2 \sqrt{\frac{w^2}{w^2+\sigma^2}} e^{-\frac{(x-\mu_2)^2}{2(w^2+\sigma^2)}}$$

and

$$E_X E_Y[K(X,Y)] = (\pi_1^2 + \pi_2^2) \sqrt{\frac{w^2}{w^2+2\sigma^2}} + 2\pi_1\pi_2 \sqrt{\frac{w^2}{w^2+2\sigma^2}} e^{-\frac{(\mu_1-\mu_2)^2}{2(w^2+2\sigma^2)}}.$$

For numerical illustration, we took a random sample of size 1000 from the distribution $0.6N(2, 1^2) + 0.4N(-2, 1^2)$. Then the centered kernel matrix was constructed using the Gaussian kernel with the bandwidth $w = 1.5$ and the eigenvectors of the matrix were obtained. The first and second rows of Figure 4.8 compare the first four leading eigenvectors of the uncentered Gaussian kernel matrix and its centered counterpart. The two leading eigenvectors of the uncentered kernel matrix capture the two components of the mixture normal distribution separately, with peaks located at -2 and 2. On the other hand, the leading eigenvector of the centered matrix combines the two components, revealing the information on the mixture differently.

Theoretical analysis

Similar to the analysis for the one-component normal example, we consider the approximation of $E_Y[K(x, Y)]$ for a mixture normal setting, Shi et al. (2009) show that the eigenfunctions of the Gaussian kernel operator are given by the eigenfunctions of the kernel operators of mixture components, with their order determined by the product of mixture proportion and eigenvalues, see Property 1 in the paper.

For the two-component mixture example, the following are the products of the five leading eigenvalues and the mixture proportion for the two components:

$$\pi_1\lambda_1^1 = .45, \pi_1\lambda_2^1 = .1125, \pi_1\lambda_3^1 = .0281, \pi_1\lambda_4^1 = .00702, \pi_1\lambda_5^1 = .00174$$

$$\pi_2\lambda_1^2 = .3, \pi_2\lambda_2^2 = .075, \pi_2\lambda_3^2 = .0188, \pi_2\lambda_4^2 = .00468, \pi_2\lambda_5^2 = .00116,$$

where λ_i^g for $g = 1, 2$ refers to the ith eigenvalue of component g. Those values are ordered as follows:

$$\lambda_1 = .45, \lambda_2 = .3, \lambda_3 = .1125, \lambda_4 = .075, \lambda_5 = .0281,$$

$$\lambda_6 = .0188, \lambda_7 = .00702, \lambda_8 = .00468, \lambda_9 = .00174, \lambda_{10} = .00116.$$

According to the order, we have the corresponding eigenfunctions for the uncentered kernel operator:

$$\phi_1 = \phi_1^1, \phi_2 = \phi_1^2, \phi_3 = \phi_2^1, \phi_4 = \phi_2^2, \phi_5 = \phi_3^1$$

$$\phi_6 = \phi_3^2, \phi_7 = \phi_4^1, \phi_8 = \phi_4^2, \phi_9 = \phi_5^1, \phi_{10} = \phi_5^2.$$

where ϕ_i^g for $g = 1, 2$ refers to the ith eigenfunction of component g. Therefore, based on the theoretical eigenfunction form and the mixture density function, we obtain the means for the eigenfunctions as follows:

$$\mu_1 = .6437, \mu_2 = .4735, \mu_3 = -.1032, \mu_4 = .1547, \mu_5 = .0559$$

$$\mu_6 = .1491, \mu_7 = -.1342, \mu_8 = .2013, \mu_9 = .1209, \mu_{10} = .1707.$$

Using the first ten eigenfunctions, we can approximate $E_Y[K(x,Y)]$ and construct the 10×10 block of $\tilde{M}$. With the eigenvalues and eigenvectors of the matrix A, we have approximate eigenfunctions given as follows:

$$\begin{aligned}
\tilde{\phi}_1(x) &\propto \; .4494(\phi_1(x) - .6437) - .4043(\phi_2(x) - .4735) \\
\tilde{\phi}_2(x) &\propto \; .3330(\phi_3(x) + .1032) - .0640(\phi_2(x) - .4735) - .0150(\phi_1(x) - .6437) \\
\tilde{\phi}_3(x) &\propto \; .2676(\phi_4(x) - .1547) - .0889(\phi_1(x) - .6437) - .0875(\phi_2(x) - .4735) \\
\tilde{\phi}_4(x) &\propto \; - .1670(\phi_5(x) - .0559) + .0305(\phi_1(x) - .6437) + .0261(\phi_2(x) - .4735) \\
&\quad + .0089(\phi_4(x) - .1547) - .0057(\phi_6(x) - .1496),
\end{aligned}$$

with the corresponding eigenvalues $\tilde{\lambda}_1 = .1833$, $\tilde{\lambda}_2 = .1383$, $\tilde{\lambda}_3 = .0579$, $\tilde{\lambda}_4 = .0253$.

The third row of Figure 4.8 shows the approximate eigenfunctions. The first two eigenfunctions have strong agreement with the two leading eigenvectors of the centered kernel matrix. As in the one-component case, accuracy of the approximation depends on that of the expansion of the adjustment term $E_Y[K(x,Y)]$.

Figure 4.8 indicates that while leading eigenvectors of the uncentered kernel matrix reveal the clusters separately, the leading eigenvector of the centered kernel matrix combines the two clusters, which is also evident in the expression of $\tilde{\phi}_1$. For clustering, centering a Gaussian kernel seems to be counter-effective as it convolutes the spectrum of the original kernel with clear correspondence to clusters.

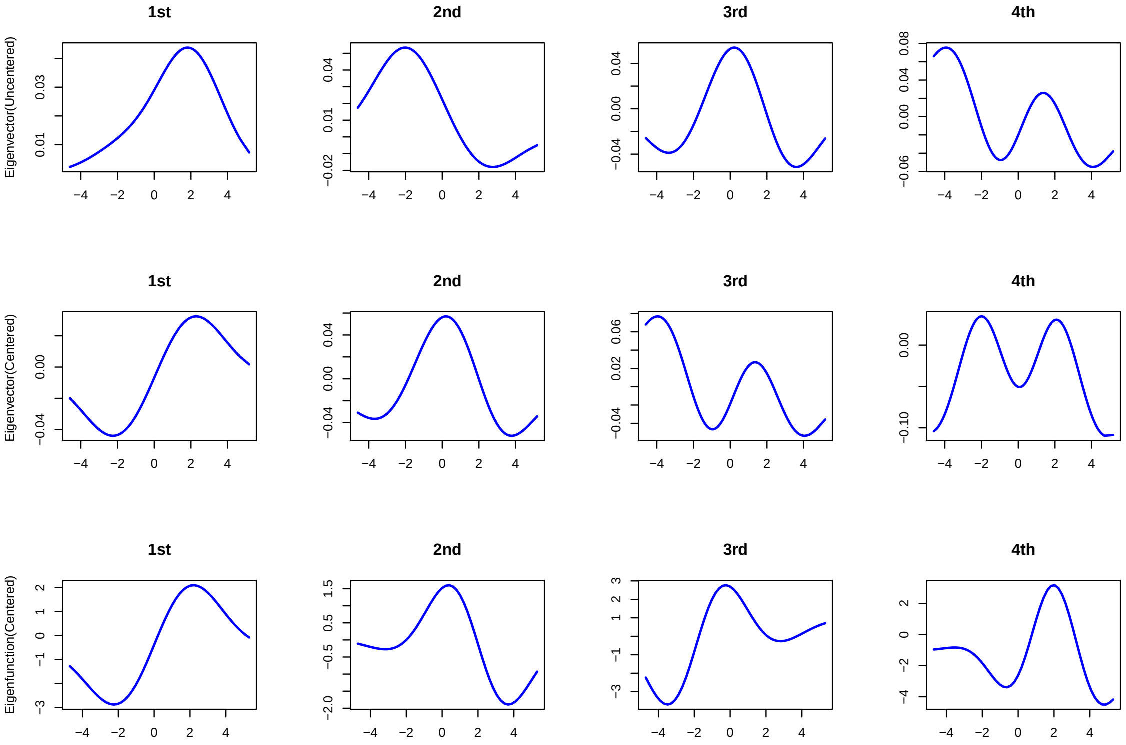


Figure 4.8: The top row shows five leading eigenvectors of a Gaussian kernel matrix of data sampled from a mixture normal distribution $0.6N(2, 1^2) + 0.4N(-2, 1^2)$, the second row shows the five leading eigenvectors of the uncentered kernel matrix. The third row shows the theoretical eigenfunctions of the centered kernel operator we obtained.

CHAPTER 5

EIGEN-ANALYSIS OF KERNEL OPERATORS FOR NONLINEAR DISCRIMINATION

As was mentioned in Section 2.2.1, PCA is mainly used for explaining the variance of data and extracting features for describing data is different from extracting features for classification. Classical algorithms which discriminate between the two classes include linear and quadratic discriminant analysis. Fisher's linear discriminant analysis finds the direction maximizing the between-class variance relative to the within-class variance. However, a linear discriminant function may not be flexible enough to capture the optimal discriminant direction. Kernel LDA extends the classical LDA by mapping data from the input space into a feature space, which allows us to find a nonlinear discriminant direction. Kernel LDA has been used successfully in a wide range of applications. In this chapter, we describe how to extend the current framework for the analysis of kernel PCA to kernel LDA. We characterize the theoretical discriminant function that maximizes the between-class variation relative to within-class variation by solving the general eigen-problem associated with the population version of kernel LDA. Through the theoretical analysis of the method from the population point of view, we will have better understanding of the kernel LDA projections in relation to the underlying data distribution. We also conduct simulation studies to explore the connection between empirical and theoretical discriminant functions.

5.1 The Population Version of Kernel LDA

Kernel LDA aims to find the direction which maximizes the between-class variation relative to within-class variation in the feature space. In order to find the best discriminant direction in the feature space $\mathcal{H}$, Mika et al. (1999) maximize the equation (2.1), which leads to the following equation:

$$B\alpha = \lambda W\alpha,$$

where B and W are the between-class variation matrix and within-class variation matrix of the feature vectors $\Phi(X)$. Explicitly, the (i,j)th entry in the between-class variation matrix B is $[(B_1^*)_i - (B_0^*)_i][(B_1^*)_j - (B_0^*)_j]$, given that $(B_l^*)_j = \frac{1}{n_l}\sum_{k=1}^{n_l} K(x_j, x_k^l)$. Hence, we have the (i,j)th entry of B as

$$\left[\frac{1}{n_1}\sum_{k=1}^{n_1} K(x_i, x_k^1) - \frac{1}{n_0}\sum_{k=1}^{n_0} K(x_i, x_k^0)\right]\left[\frac{1}{n_1}\sum_{k=1}^{n_1} K(x_j, x_k^1) - \frac{1}{n_0}\sum_{k=1}^{n_0} K(x_j, x_k^0)\right].$$

The within-class variation matrix W is

$$W := K_1(I - 1_{n_1})K_1^T + K_2(I - 1_{n_0})K_2^T,$$

where K_l is a $n \times n_l$ matrix with its entries $(K_l)_{nk} = K(x_n, x_k^l)$, 1_{n_l} is the matrix with all entries $1/n_l$. Therefore, the (i,j)th entry of W is of the form,

$$\begin{aligned}&\sum_{k=1}^{n_1} K(x_i, x_k^1)K(x_j, x_k^1) - \frac{1}{n_1}\sum_{k=1}^{n_1} K(x_i, x_k^1)\sum_{k=1}^{n_1} K(x_j, x_k^1)\\ &+\sum_{k=1}^{n_0} K(x_i, x_k^0)K(x_j, x_k^0) - \frac{1}{n_0}\sum_{k=1}^{n_0} K(x_i, x_k^0)\sum_{k=1}^{n_0} K(x_j, x_k^0).\end{aligned}$$

The above sample version of of B and W leads to the following population version $B(x^*, x)$ and $W(x^*, x)$,

$$\begin{aligned} B(x^*, x) =& \{E_X [K(x^*, X)|Y = 1] - E_X [K(x^*, X)|Y = 0]\} \times \\ & \{E_X [K(x, X)|Y = 1] - E_X [K(x, X)|Y = 0]\} \\ W(x^*, x) =& \pi_1 \text{cov}_X (K(x^*, X), K(x, X)|Y = 1) + \pi_0 \text{cov}_X (K(x^*, X), K(x, X)|Y = 0) . \end{aligned}$$

Analogous to the eigen-analysis of kernel PCA, we consider the following eigen problem given in terms of between-class and within-class variation operators for kernel LDA.

$$\int_{\mathcal{X}} B(x^*, x)\alpha(x)p(x)dx = \lambda \int_{\mathcal{X}} W(x^*, x)\alpha(x)p(x)dx \tag{5.1}$$

for all $x^* \in \mathcal{X}$.

The solution $\alpha(x)$ to the equation (5.1) depends on the kernel K and the probability distribution p and identification of the solution will give us better understanding of kernel LDA projections in relation to probability distribution of the data.

Just as the empirical kernel embedding of a new point x is of the form $\sum_i \alpha_i K(x_i, x)$ in kernel LDA algorithm for discrimination, the corresponding discriminant function is

$$f(x) = \int \alpha(x^*)K(x^*, x)p(x^*)dx^*. \tag{5.2}$$

In the following section, we focus on the polynomial kernel and solve for the theoretical discriminant function $f(x)$.

5.2 Eigen-analysis of the Polynomial Kernel Operator

5.2.1 Two-dimensional setting

For simple illustration, we consider a two-dimensional setting, with two classes. For the polynomial kernel with degree d, we obtain the between-class variation function $B(x^*, x)$ and within-class variation function $W(x^*, x)$ explicitly.

For $B(x^*, x)$, observe that

$$\begin{aligned} & E_X\left[K(x^*, X)|Y=1\right] = E_X\left[(x_1^* X_1 + x_2^* X_2)^d|Y=1\right] \\ =& E_X\left[\sum_{j=0}^{d}\binom{d}{j}(x_1^* X_1)^{d-j}(x_2^* X_2)^j|Y=1\right] \\ =& \sum_{j=0}^{d}\binom{d}{j}E_X\left[X_1^{d-j}X_2^j|Y=1\right]x_1^{*(d-j)}x_2^{*j}. \end{aligned}$$

Similarly,

$$E_X\left[K(x^*, X)|Y=0\right] = \sum_{j=0}^{d}\binom{d}{j}E_X\left[X_1^{d-j}X_2^j|Y=0\right]x_1^{*(d-j)}x_2^{*j}.$$

Letting $\mu^1_{d-j,j}$ denote $E_X\left[X_1^{d-j}X_2^j|Y=1\right]$ and $\mu^0_{d-j,j}$ denote $E_X\left[X_1^{d-j}X_2^j|Y=0\right]$ respectively. We have

$$\begin{aligned} & E_X\left[K(x^*, X)|Y=1\right] - E_X\left[K(x^*, X)|Y=0\right] \\ =& \sum_{j=0}^{d}\binom{d}{j}\left\{\mu^1_{d-j,j} - \mu^0_{d-j,j}\right\}x_1^{*(d-j)}x_2^{*j}. \end{aligned}$$

$B(x^*, x)$ can thus be expressed as

$$\sum_{j=0}^{d}\sum_{k=0}^{d}\binom{d}{j}\binom{d}{k}\left\{\mu^1_{d-j,j} - \mu^0_{d-j,j}\right\}\left\{\mu^1_{d-k,k} - \mu^0_{d-k,k}\right\}x_1^{*(d-j)}x_2^{*j}x_1^{d-k}x_2^{k}.$$

Therefore, the left side of the equation (5.1) is given by

$$\begin{aligned}
&\int B(x^*, x)\alpha(x)p(x)dx \\
=&\sum_{j=0}^{d}\sum_{k=0}^{d}\binom{d}{j}\binom{d}{k}\left\{\mu^1_{d-j,j}-\mu^0_{d-j,j}\right\}\left\{\mu^1_{d-k,k}-\mu^0_{d-k,k}\right\}x_1^{*(d-j)}x_2^{*j}\int x_1^{d-k}x_2^{k}\alpha(x)p(x)dx.
\end{aligned}$$

On the other hand, observe that

$$\begin{aligned}
&\operatorname{cov}_X\left(K(x^*,X),K(x,X)\quad|Y=1\right) \\
&=\operatorname{cov}_X\left(\sum_{j=0}^{d}\binom{d}{j}(x_1^*X_1)^{d-j}(x_2^*X_2)^j,\sum_{k=0}^{d}\binom{d}{k}(x_1X_1)^{d-k}(x_2X_2)^k\quad|Y=1\right) \\
&=\operatorname{cov}_X\left(\sum_{j=0}^{d}\binom{d}{j}x_1^{*(d-j)}x_2^{*j}X_1^{d-j}X_2^j,\sum_{k=0}^{d}\binom{d}{k}x_1^{d-k}x_2^kX_1^{d-k}X_2^k\quad|Y=1\right) \\
&=\sum_{j=0}^{d}\sum_{k=0}^{d}\binom{d}{j}\binom{d}{k}\operatorname{cov}_X\left(X_1^{d-j}X_2^j,X_1^{d-k}X_2^k\quad|Y=1\right)x_1^{*(d-j)}x_2^{*j}x_1^{d-k}x_2^k.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\operatorname{cov}_X\left(K(x^*,X),K(x,X)\quad|Y=0\right) \\
=&\sum_{j=0}^{d}\sum_{k=0}^{d}\binom{d}{j}\binom{d}{k}\operatorname{cov}_X\left(X_1^{d-j}X_2^j,X_1^{d-k}X_2^k\quad|Y=0\right)x_1^{*(d-j)}x_2^{*j}x_1^{d-k}x_2^k
\end{aligned}$$

Letting $\sigma^1_{(d-j,j),(d-k,k)}$ denote $\operatorname{cov}_X\left(X_1^{d-j}X_2^j,X_1^{d-k}X_2^k\quad|Y=1\right)$ and $\sigma^0_{(d-j,j),(d-k,k)}$ denote $\operatorname{cov}_X\left(X_1^{d-j}X_2^j,X_1^{d-k}X_2^k\quad|Y=0\right)$, $W(x^*,x)$ can then be expressed as the weighted sum of those parts as follows:

$$\sum_{j=0}^{d}\sum_{k=0}^{d}\binom{d}{j}\binom{d}{k}\left\{\pi_1\sigma^1_{(d-j,j),(d-k,k)}+\pi_0\sigma^0_{(d-j,j),(d-k,k)}\right\}x_1^{*(d-j)}x_2^{*j}x_1^{d-k}x_2^k.$$

The right side of equation (5.1) is then given by

$$\begin{aligned}
&\int W(x^*, x)\alpha(x)p(x)dx \\
=&\sum_{j=0}^{d}\sum_{k=0}^{d}\binom{d}{j}\binom{d}{k}\left\{\pi_1\sigma^1_{(d-j,j),(d-k,k)}+\pi_0\sigma^0_{(d-j,j),(d-k,k)}\right\}x_1^{*(d-j)}x_2^{*j}\int x_1^{d-k}x_2^{k}\alpha(x)p(x)dx.
\end{aligned}$$

Let $C_k = \displaystyle\int x_1^{d-k} x_2^k \alpha(x) p(x) dx$ for $k = 0, \cdots, d$. Then the equation (5.1) is simplified as

$$\begin{aligned}
&\sum_{j=0}^{d}\sum_{k=0}^{d} C_k \binom{d}{j}\binom{d}{k} \left\{\mu^1_{d-j,j} - \mu^0_{d-j,j}\right\} \left\{\mu^1_{d-k,k} - \mu^0_{d-k,k}\right\} x_1^{*(d-j)} x_2^{*j} \\
=&\lambda \sum_{j=0}^{d}\sum_{k=0}^{d} C_k \binom{d}{j}\binom{d}{k} \left\{\pi_1 \sigma^1_{(d-j,j),(d-k,k)} + \pi_0 \sigma^0_{(d-j,j),(d-k,k)}\right\} x_1^{*(d-j)} x_2^{*j}.
\end{aligned}$$

Notice that this equation holds for all $x^* \in R^2$. Hence for $j = 0, \cdots, d$, the coefficients for $x_1^{*(d-j)} x_2^{*j}$ should match on both sides of the equation. For example, when $j = 0$, matching the coefficients leads to the following equation:

$$\begin{aligned}
&\left\{\mu^1_{d,0} - \mu^0_{d,0}\right\} \sum_{k=0}^{d} C_k \binom{d}{k} \left\{\mu^1_{d-k,k} - \mu^0_{d-k,k}\right\} \\
=&\lambda \sum_{k=0}^{d} C_k \binom{d}{k} \left\{\pi_1 \sigma^1_{(d,0),(d-k,k)} + \pi_0 \sigma^0_{(d,0),(d-k,k)}\right\}.
\end{aligned}$$

Equations can be set up for $j = 1$ through d similarly, where each equation is linear in the coefficients C_k. Define B^* as a $(d+1) \times (d+1)$ matrix and W^* also a $(d+1) \times (d+1)$ matrix, with the ij-th entry of B^* and W^* given by,

$$\begin{aligned}
B^*_{ij} &= \binom{d}{i}\binom{d}{j} \left\{\mu^1_{d-i,i} - \mu^0_{d-i,i}\right\} \left\{\mu^1_{d-j,j} - \mu^0_{d-j,j}\right\} \qquad (5.3) \\
W^*_{ij} &= \binom{d}{i}\binom{d}{j} \left\{\pi_1 \sigma^1_{(d-i,i),(d-j,j)} + \pi_0 \sigma^0_{(d-i,i),(d-j,j)}\right\}.
\end{aligned}$$

Then the set of equations to solve for λ and $\mathbf{C} = (C_0, C_1, \cdots, C_d)^t$ is expressed as

$$B^* \mathbf{C} = \lambda W^* \mathbf{C},$$

which is a generalized eigenvalue problem. Since W^* is symmetric and non-negative definite, we can write W^* as $W^{*\frac{1}{2}} W^{*\frac{1}{2}}$. Multiplying both sides of the equation by $W^{*-\frac{1}{2}}$, we have $W^{*-\frac{1}{2}} B \mathbf{C} = \lambda W^{*\frac{1}{2}} \mathbf{C}$. This equation is the same as

$$W^{*-\frac{1}{2}} B W^{*-\frac{1}{2}} W^{*\frac{1}{2}} \mathbf{C} = \lambda W^{*\frac{1}{2}} \mathbf{C}.$$

Let $\mathbf{d} = W^{*\frac{1}{2}}\mathbf{C}$, then λ and $\mathbf{d}$ is a eigenvalue-eigenvector pair of the matrix $W^{*-\frac{1}{2}}BW^{*-\frac{1}{2}}$. Therefore, given the eigenvector $\mathbf{d}$, $\mathbf{C} = W^{*-\frac{1}{2}}\mathbf{d}$.

Given C_j, we can verify that the theoretical discriminant function $f(x)$ in equation (5.2) for the polynomial kernel function is

$$\begin{aligned} f(x) &= \int \alpha(x^*)K(x^*x)p(x^*)dx^* \\ &= \int \alpha(x^*)\left[\sum_{j=0}^{d}\binom{d}{j}x_1^{d-j}x_2^j x_1^{*(d-j)}x_2^{*j}\right]p(x^*)dx^* \\ &= \sum_{j=0}^{d}\binom{d}{j}x_1^{d-j}x_2^j\int x_1^{*(d-j)}x_2^{*j}\alpha(x^*)p(x^*)dx^* \\ &= \sum_{j=0}^{d}C_j\binom{d}{j}x_1^{d-j}x_2^j. \end{aligned}$$

This theoretical discriminant function $f(x)$ can be viewed as the population version of the empirical kernel LDA embeddings. The statements so far lead to the following theorem.

Theorem 4. *Suppose that $\pi_1 = P(Y=1), \pi_0 = P(Y=0)$ and given Y, the conditional distribution of $(X_1, X_2)^t \in \mathbb{R}^2$ has finite moments, $\mu^l_{d-j,j} = E(X_1^{d-j}X_2^j|Y = l)$ and $\sigma^l_{(d-j,j),(d-k,k)} = cov(X_1^{d-j}X_2^j, X_1^{d-k}X_2^k|Y=l)$, $j,k = 0,\cdots,d$ for each class, $l = 0,1$. For polynomial kernel of degree d, $K(\mathbf{x},\mathbf{u}) = (\mathbf{x}^T\mathbf{u})^d$,*

(i) The discriminant function maximizing the ratio of between-class variation relative to within-class variation is of the form

$$f(x) = \sum_{j=0}^{d}C_j\binom{d}{j}x_1^{d-j}x_2^j.$$

(ii) The coefficients $\mathbf{C} = (C_0,\cdots,C_d)^t$ determining the discriminant function

and λ *satisfy the equation* $B^*\mathbf{C} = \lambda W^*\mathbf{C}$*, where the* ij*th entries of matrices* B^* *and* W^* *are*

$$B^*_{ij} = \binom{d}{i}\binom{d}{j}\left\{\mu^1_{d-i,i} - \mu^0_{d-i,i}\right\}\left\{\mu^1_{d-j,j} - \mu^0_{d-j,j}\right\} \quad \textit{and}$$

$$W^*_{ij} = \binom{d}{i}\binom{d}{j}\left\{\pi_1\sigma^1_{(d-i,i),(d-j,j)} + \pi_0\sigma^0_{(d-i,i),(d-j,j)}\right\}.$$

5.2.2 Multi-dimensional setting

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

$$(\mathbf{x}^t\mathbf{u})^d = \left(\sum_{k=1}^{p} x_k u_k\right)^d = \sum_{j_1+\cdots+j_p=d}\binom{d}{j_1,\cdots,j_p}\prod_{k=1}^{p}(x_k u_k)^{j_k}.$$

Using the expansion, we have

$$E_X\left[K(x^*,X)|Y=1\right] = E_X\left[\sum_{j_1+\cdots+j_p=d}\binom{d}{j_1,\cdots,j_p}\prod_{k=1}^{p}(x^*_k X_k)^{j_k}|Y=1\right]$$

$$= \sum_{j_1+\cdots+j_p=d}\binom{d}{j_1,\cdots,j_p}E_X\left[\prod_{k=1}^{p}X_k^{j_k}|Y=1\right]\prod_{k=1}^{p}x_k^{*j_k}$$

$$= \sum_{j_1+\cdots+j_p=d}\binom{d}{j_1,\cdots,j_p}\mu^1_{j_1,j_2,\cdots,j_p}\prod_{k=1}^{p}x_k^{*j_k},$$

and

$$E_X\left[K(x^*,X)|Y=0\right] = \sum_{j_1+\cdots+j_p=d}\binom{d}{j_1,\cdots,j_p}\mu^0_{j_1,j_2,\cdots,j_p}\prod_{k=1}^{p}x_k^{*j_k},$$

where $\mu^1_{j_1,j_2,\cdots,j_p} = E_X\left[\prod_{k=1}^{p}X_k^{j_k}|Y=1\right]$ and $\mu^0_{j_1,j_2,\cdots,j_p} = E_X\left[\prod_{k=1}^{p}X_k^{j_k}|Y=0\right]$. Thus, the first part of $B(x^*,x)$ is

$$E_X\left[K(x^*,X)|Y=1\right] - E_X\left[K(x^*,X)|Y=0\right]$$

$$= \sum_{j_1+\cdots+j_p=d}\binom{d}{j_1,\cdots,j_p}(\mu^1_{j_1,j_2,\cdots,j_p} - \mu^0_{j_1,j_2,\cdots,j_p})\prod_{k=1}^{p}x_k^{*j_k},$$

while the second part of $B(x^*, x)$ is

$$\begin{aligned}
&E_X\left[K(x,X)|Y=1\right] - E_X\left[K(x,X)|Y=0\right] \\
=&\sum_{i_1+\cdots+i_p=d} \binom{d}{i_1,\cdots,i_p} (\mu^1_{i_1,i_2,\cdots,i_p} - \mu^0_{i_1,i_2,\cdots,i_p}) \prod_{k=1}^p x_k^{i_k}.
\end{aligned}$$

Similar as in the two-dimensional setting, $B(x^*, x)$ can be expressed as

$$\begin{aligned}
&\sum_{j_1+\cdots+j_p=d} \sum_{i_1+\cdots+i_p=d} \binom{d}{j_1,\cdots,j_p} \binom{d}{i_1,\cdots,i_p} (\mu^1_{j_1,j_2,\cdots,j_p} - \mu^0_{j_1,j_2,\cdots,j_p}) \times \\
&(\mu^1_{i_1,i_2,\cdots,i_p} - \mu^0_{i_1,i_2,\cdots,i_p}) \prod_{k=1}^p x_k^{*j_k} x_k^{i_k}.
\end{aligned}$$

For $W(x^*, x)$, note that

$$\begin{aligned}
&\text{cov}_X\left(K(x^*,X), K(x,X) \quad |Y=1\right) \\
=&\text{cov}_X\left(\sum_{j_1+\cdots+j_p=d} \binom{d}{j_1,\cdots,j_p} \prod_{k=1}^p (x_k^* X_k)^{j_k}, \sum_{i_1+\cdots+i_p=d} \binom{d}{i_1,\cdots,i_p} \prod_{k=1}^p (x_k X_k)^{i_k} \quad |Y=1\right) \\
=&\sum_{j_1+\cdots+j_p=d} \sum_{i_1+\cdots+i_p=d} \binom{d}{j_1,\cdots,j_p} \binom{d}{i_1,\cdots,i_p} \text{cov}_X\left(\prod_{k=1}^p X_k^{j_k}, \prod_{k=1}^p X_k^{i_k} \quad |Y=1\right) \prod_{k=1}^p x_k^{*j_k} x_k^{i_k} \\
=&\sum_{j_1+\cdots+j_p=d} \sum_{i_1+\cdots+i_p=d} \binom{d}{j_1,\cdots,j_p} \binom{d}{i_1,\cdots,i_p} \sigma^1_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)} \prod_{k=1}^p x_k^{*j_k} x_k^{i_k}.
\end{aligned}$$

where $\sigma^1_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)} = \text{cov}_X\left(\prod_{k=1}^p X_k^{j_k}, \prod_{k=1}^p X_k^{i_k} \quad |Y=1\right)$. Therefore, $W(x^*, x)$ can be expressed as

$$\begin{aligned}
&\sum_{j_1+\cdots+j_p=d} \sum_{i_1+\cdots+i_p=d} \binom{d}{j_1,\cdots,j_p} \binom{d}{i_1,\cdots,i_p} \times \\
&\{\pi_1 \sigma^1_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)} + \pi_0 \sigma^0_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)}\} \prod_{k=1}^p x_k^{*j_k} x_k^{i_k}.
\end{aligned}$$

Letting $C_{i_1,\cdots,i_p} = \int \prod_{k=1}^{p} x_k^{i_k}\alpha(x)p(x)dx$, we have the equation(5.1) given as

$$\sum_{j_1+\cdots+j_p=d}\sum_{i_1+\cdots+i_p=d} C_{i_1,\cdots,i_p}\binom{d}{j_1,\cdots,j_p}\binom{d}{i_1,\cdots,i_p}\times$$

$$(\mu^1_{j_1,j_2,\cdots,j_p}-\mu^0_{j_1,j_2,\cdots,j_p})(\mu^1_{i_1,i_2,\cdots,i_p}-\mu^0_{i_1,i_2,\cdots,i_p})\prod_{k=1}^{p}x_k^{*j_k}$$

$$=\lambda\sum_{j_1+\cdots+j_p=d}\sum_{i_1+\cdots+i_p=d} C_{i_1,\cdots,i_p}\binom{d}{j_1,\cdots,j_p}\binom{d}{i_1,\cdots,i_p}\times$$

$$\{\pi_1\sigma^1_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)}+\pi_0\sigma^0_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)}\}\prod_{k=1}^{p}x_k^{*j_k}.$$

Define B^* and W^* as matrices of dimension $\binom{d+p-1}{d}$ with entries given by

$$\binom{d}{j_1,\cdots,j_p}\binom{d}{i_1,\cdots,i_p}(\mu^1_{j_1,j_2,\cdots,j_p}-\mu^0_{j_1,j_2,\cdots,j_p})(\mu^1_{i_1,i_2,\cdots,i_p}-\mu^0_{i_1,i_2,\cdots,i_p})$$

$$\text{and}\quad\binom{d}{j_1,\cdots,j_p}\binom{d}{i_1,\cdots,i_p}\{\pi_1\sigma^1_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)}+\pi_0\sigma^0_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)}\}.$$

Then the equation $B^*\mathbf{C} = \lambda W^*\mathbf{C}$ gives the values of $C_{i_1,\cdots,i_p}$ through the generalized eigen-problem described for the two-dimensional setting. The theoretical discriminant function $f(x)$ is thus in the form of

$$f(x) = \sum_{j_1+\cdots+j_p=d} C_{j_1,\cdots,j_p}\binom{d}{j_1,\cdots,j_p}x_1^{j_1}\cdots x_p^{j_p}.$$

We arrive at the following theorem.

Theorem 5. *Suppose that* $\pi_1 = P(Y=1), \pi_0 = P(Y=0)$ *and given* Y*, the conditional distribution of* $(X_1, X_2, \cdots, X_p)^t \in \mathbb{R}^p$ *has finite moments,* $\mu^l_{j_1,j_2,\cdots,j_p} = E_X\left[\prod_{k=1}^{p}X_k^{j_k}|Y=l\right]$, $\sigma^l_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)} = cov_X\left(\prod_{k=1}^{p}X_k^{j_k}, \prod_{k=1}^{p}X_k^{i_k} \quad |Y=l\right)$, *for* $j_1+\cdots+j_p = d, i_1+\cdots+i_p = d$ *for each class,* $l = 0, 1$. *For the polynomial kernel of degree* d, $K(\mathbf{x},\mathbf{u}) = (\mathbf{x}^t\mathbf{u})^d$, *(i) The discriminant function maximizing the ratio of between-class variation relative to within-class variation is of the form*

$$f(x) = \sum_{j_1+\cdots+j_p=d} C_{j_1,\cdots,j_p} \binom{d}{j_1,\cdots,j_p} x_1^{j_1} \cdots x_p^{j_p}.$$

(ii) The coefficients $\mathbf{C}$ *determining the discriminant function and* λ *satisfy the equation* $B^*\mathbf{C} = \lambda W^*\mathbf{C}$*, where* B^* *and* W^* *are matrices of dimension* $\binom{d+p-1}{d}$ *with entries given by*

$$\binom{d}{j_1,\cdots,j_p}\binom{d}{i_1,\cdots,i_p}(\mu^1_{j_1,j_2,\cdots,j_p} - \mu^0_{j_1,j_2,\cdots,j_p})(\mu^1_{i_1,i_2,\cdots,i_p} - \mu^0_{i_1,i_2,\cdots,i_p})$$
$$\text{and} \quad \binom{d}{j_1,\cdots,j_p}\binom{d}{i_1,\cdots,i_p}\{\pi_1\sigma^1_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)} + \pi_0\sigma^0_{(j_1,\cdots,j_p),(i_1,\cdots,i_p)}\}.$$

5.3 Simulation Studies

In this section, we present a simulation study to investigate the empirical discriminant function from a kernel LDA algorithm and the theoretical discriminant function which could be obtained from the population perspective. We consider a bivariate normal example with two classes:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \quad \text{for Y} = 1 \tag{5.4}$$

and

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 2 \\ -2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \quad \text{for Y} = 0.$$

We examine the explicit form of discriminant functions with kernel LDA and comment on the effect of the degree of polynomial kernels on the discriminant functions.

A sample of size 400 is drawn from the distribution and the mixture proportion for both classes is .5. Figure 5.1 shows the scatterplot of the simulated data and the contours of the probability density function.

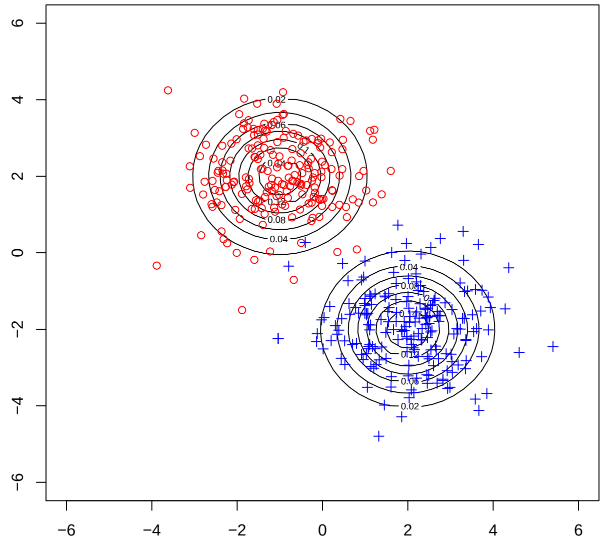


Figure 5.1: The contours of the probability density function of a mixture normal example with two classes. The red circles and blue crosses show the data points generated from the distribution.

For the two classes, we have the following moments necessary for the construction of the theoretical discriminant function later,

$$
\begin{aligned}
&\text{Class 1}:\\
&E(X_1) = -1, E(X_1^2) = 2, E(X_1^3) = -4, E(X_1^4) = 10, E(X_1^5) = -26, E(X_1^6) = 76\\
&E(X_2) = 2, E(X_2^2) = 5, E(X_2^3) = 14, E(X_2^4) = 43, E(X_2^5) = 142, E(X_2^6) = 499,\\
&\text{Class 0}:\\
&E(X_1) = 2, E(X_1^2) = 5, E(X_1^3) = 14, E(X_1^4) = 43, E(X_1^5) = 142, E(X_1^6) = 499\\
&E(X_2) = -2, E(X_2^2) = 5, E(X_2^3) = -14, E(X_2^4) = 43, E(X_2^5) = -142, E(X_2^6) = 499.
\end{aligned}
$$

We investigate the explicit theoretical discriminant function $f(x)$ for the polynomial kernel with $d = 1, 2, 3$ separately in the following.

When degree is 1

For the first-order polynomial kernel $K(\mathbf{x}, \mathbf{u}) = x_1u_1 + x_2u_2$, the between-class matrix B^* and within-class matrix W^* in (5.3) are given by

$$B^* = \begin{bmatrix} 9 & -12 \\ -12 & 16 \end{bmatrix}$$

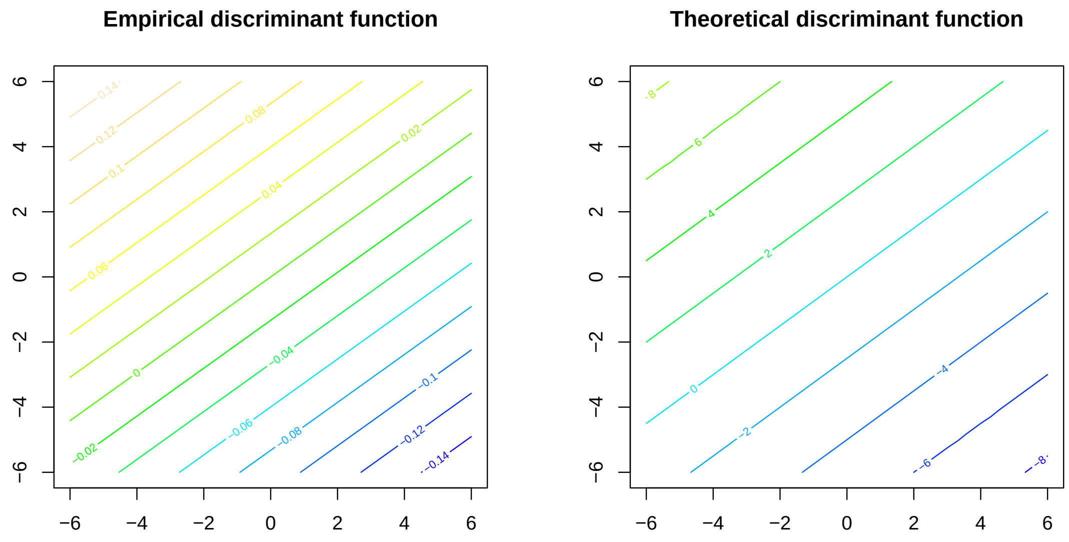
and

$$W^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

$\mathbf{d}$ is the leading eigenvector for matrix $W^{*-\frac{1}{2}}B^*W^{*-\frac{1}{2}}$, which is B^* itself in this example. Since $W^{*-\frac{1}{2}}$ is the identity matrix, $\mathbf{C}$ is the same as $\mathbf{d}$ and it equals $(-.6, .8)^t$.

Therefore, $f(x) = C_0 x_1 + C_1 x_2 = -.6x_1 + .8x_2$. We notice that this discriminant function is proportional to Fisher's linear discriminant function, $-3x_1 + 4x_2$, obtained from the direction of the difference between two class centers. With the linear kernel, we are essentially doing linear discriminant analysis.

Figure 5.2 compares the contours of the theoretical discriminant function and its empirical version from kernel LDA algorithm with the sample. It can be seen that the directions of the change in contour lines point to the direction of mean difference, which naturally maximizes the between-class variation.



Empirical discriminant function

Theoretical discriminant function

Figure 5.2: The left panel shows the contours of the empirical discriminant function from kernel LDA algorithm with linear kernel; the right panel shows the contours of theoretical discriminant function.

When degree is 2

Now, for the polynomial kernel of degree $d = 2$, where $K(\mathbf{x}, \mathbf{u}) = (x_1u_1 + x_2u_2)^2$, the between-class matrix B^* and within-class matrix W^* are given by

$$B^* = \begin{bmatrix} 9 & -12 & 0 \\ -12 & 16 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$W^* = \begin{bmatrix} 12 & -12 & 0 \\ -12 & 30 & -12 \\ 0 & -12 & 18 \end{bmatrix}.$$

From the leading eigenvector $\mathbf{d}$ of the matrix $W^{*-\frac{1}{2}}B^*W^{*-\frac{1}{2}}$, we obtain $\mathbf{C} = W^{*-\frac{1}{2}}\mathbf{d} = (.1627, -.1085, -.0723)^t$.

Therefore, the discriminant function $f(x) = C_0x_1^2+2C_1x_1x_2+C_2x_2^2 = .1627x_1^2 - .217x_1x_2 - .0723x_2^2$. The contours of the empirical kernel LDA embedding and the discriminant function $f(x)$ are given in Figure 5.3. The left panel is for the empirical embedding from kernel LDA algorithm while the right panel is for the theoretical discriminant function $f(x)$. The two plots for this mixture normal example show great similarity in terms of shape and change in the contours. Both embeddings are symmetric around the optimal linear discriminant boundary that is orthogonal to the mean difference. Hence, information for discrimination of the two classes is completely lost in the embeddings. Therefore, the polynomial kernel with $d = 2$ would not be useful for classification in this example.

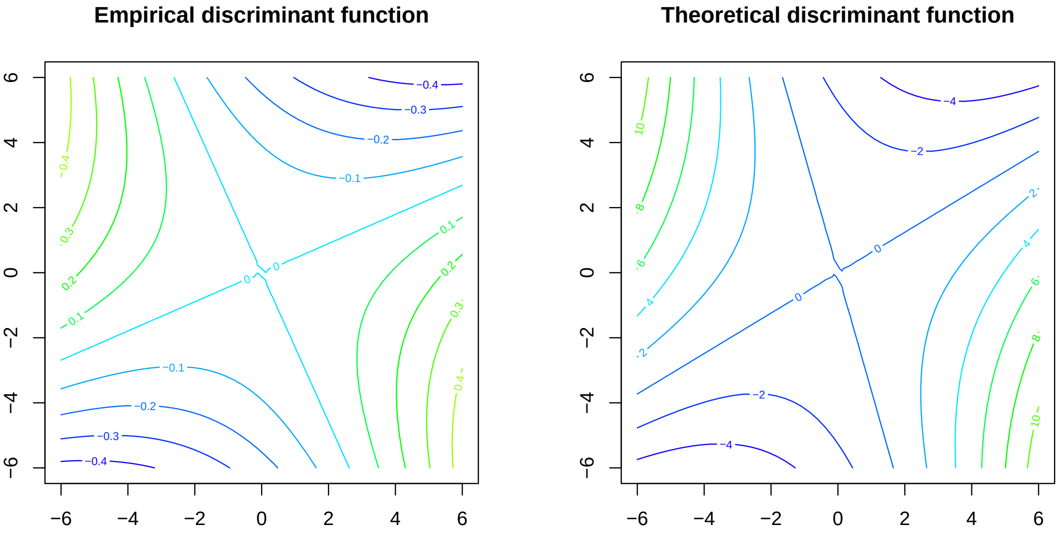


Figure 5.3: The left panel shows the contours of the empirical discriminant function from kernel LDA algorithm; the right panel shows the contours of theoretical discriminant function ($d = 2$).

When degree is 3

The between-class variation matrix B^* and within-class variation matrix W^* for the polynomial kernel of degree $d = 3$ are the 4×4 matrices and given explicitly as

$$B^* = \begin{bmatrix} 324 & -756 & 810 & -504 \\ -756 & 1764 & -1890 & 1176 \\ 810 & -1890 & 2025 & -1260 \\ -504 & 1176 & -1260 & 784 \end{bmatrix}$$

and

$$W^* = \begin{bmatrix} 181.5 & -270 & 157.5 & 0 \\ -270 & 670.5 & -594 & 157.5 \\ 157.5 & -594 & 792 & -324 \\ 0 & 157.5 & -324 & 303 \end{bmatrix}.$$

The leading eigenvector of $W^{*-\frac{1}{2}} B^* W^{*-\frac{1}{2}}$ is $\mathbf{d} = (0.4376 - 0.49240.3994 - 0.6376)^t$, so $\mathbf{C}$ is given as $W^{*-\frac{1}{2}}\mathbf{d} = (0.0414, -0.0071, -0.0052, -0.0460)^t$, which results in the discriminant function, $f(x) = C_0 x_1^3 + 3C_1 x_1^2 x_2 + 3C_2 x_1 x_2^2 + C_3 x_2^3 = .0414x_1^3 - .0142x_1^2 x_2 - .0104 x_1 x_2^2 - .0460 x_2^3$.

Figure 5.4 shows the empirical kernel LDA embedding and the theoretical discriminant function. We can see that both plots are similar in shape and direction of change in the contours. The change in the contour lines points to the direction of the mean difference between the two classes, which indicates that the polynomial kernel with $d = 3$ is effective in classification.

For this simulation example, the odd degrees $d = 1$ and $d = 3$ show successful separation of two classes, while the even degree $d = 2$ fails to do so. As in kernel PCA with polynomial kernel, the degree of polynomial kernel has different effects on the data embeddings in kernel LDA as well.

5.4 Effect of Degree

The mixture normal example with two classes we have investigated makes the polynomial kernels with odd degree effective in kernel LDA. For comparison, we give another example where the polynomial kernel of even degree would perform

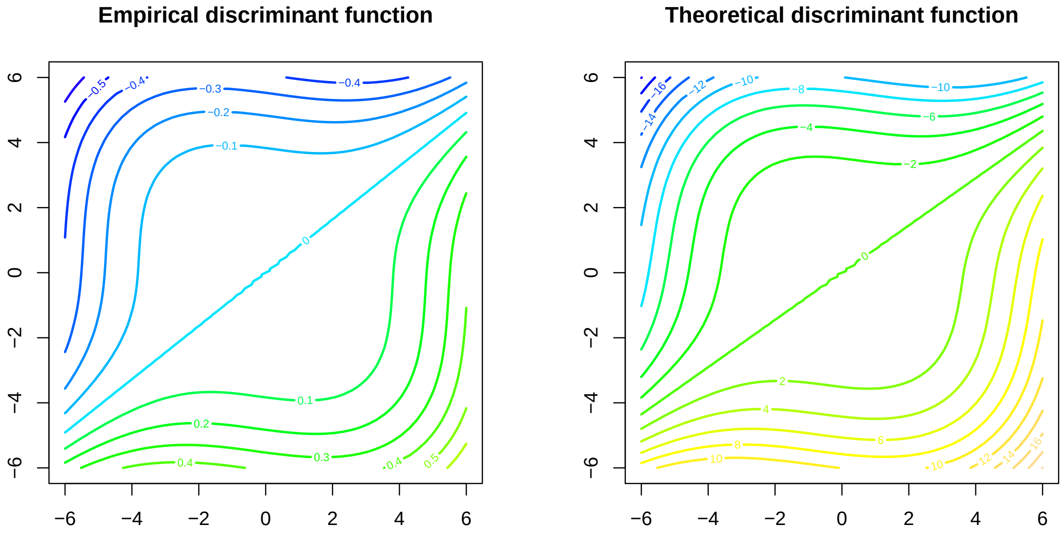


Figure 5.4: The left panel shows the embeddings of kernel LDA algorithm; the right panel shows the contours of the discriminant function ($d = 3$).

better than that of odd degree for discrimination. We revisit the "wheel" data discussed in kernel PCA.

Figure 5.5 shows the original wheel data and its associated theoretical discriminant functions for the polynomial kernel with degree $d = 1$ up to $d = 3$. Since the data set wasn't generated under a formal probability model, we used the sample moments as an intermediate step to derive the theoretical discriminant functions. For comparison with the theoretical functions, Figure 5.6 shows the empirical kernel LDA discriminant functions.

As we can see from the figures, $d = 2$ leads to successful separation of the two classes with contours of the discriminant functions forming concentric circles around the origin. In comparison, the theoretical discriminant functions with odd degree polynomial kernels do not provide such ideal embeddings for discrimination of the two classes.

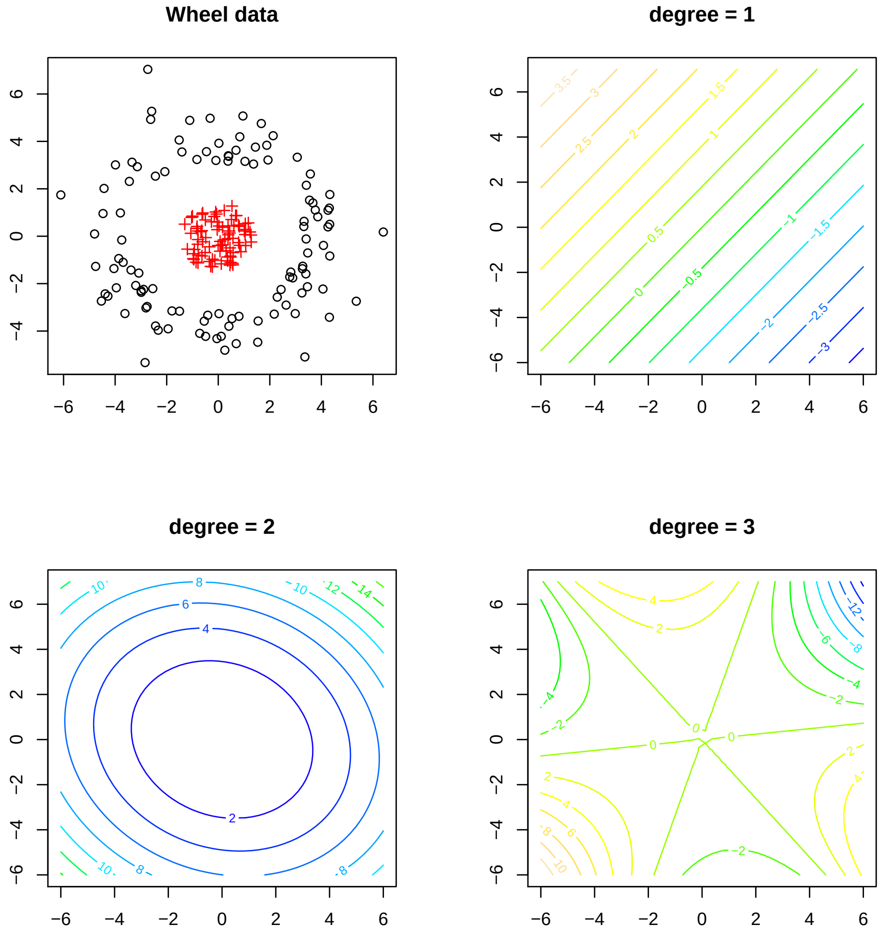


Figure 5.5: “Wheel” data and contours of the theoretical discriminant functions of kernel LDA with polynomial kernel of varying degrees based on the sample moments.

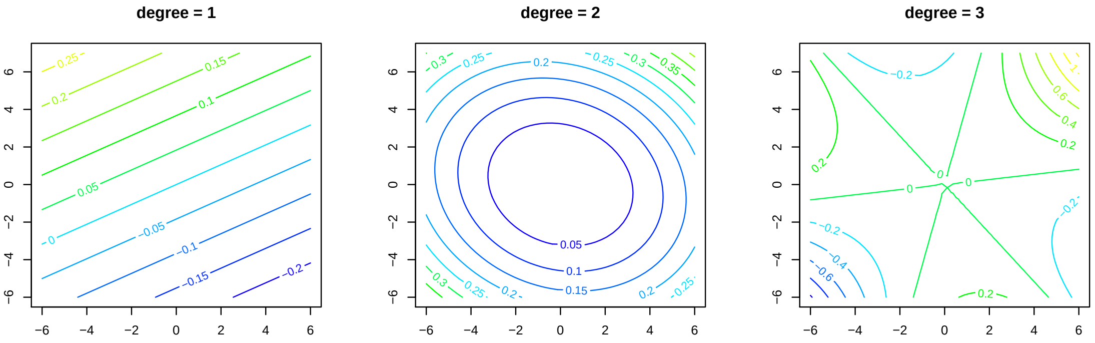


Figure 5.6: Contours of the empirical discriminant functions for “wheel” data.

Since we have the feature mapping for the polynomial kernel, the effect of degree can be seen through the explicit features accordingly. Figure 5.7 shows the plot of two explicit features x_1^3 vs $\sqrt{3}x_1^2x_2$ (polynomial kernel with $d = 3$) for the bivariate normal example in (5.4). Figure 5.8 gives the scatterplot of two features x_1^2 vs x_2^2 (polynomial kernel with $d = 2$) for the “wheel” data. We see that the two classes in both data examples are separated well as expected given the appropriate degree of polynomial kernel, which explains the effect of degree on the embedding.

The analysis of the wheel data set in Section 3.2 in the context of kernel PCA suggests that the first principal component can be used to distinguish the two classes when $d = 2$, as shown in the top right panel of Figure 3.5. This brings an interesting comparison between two nonlinear embeddings given by kernel PCA and kernel LDA. To make direct comparison between kernel PCA and kernel LDA, Figure 5.9 shows the contours of the leading eigenfunction from kernel PCA based on the sample moment matrix and the discriminant function from kernel LDA. The

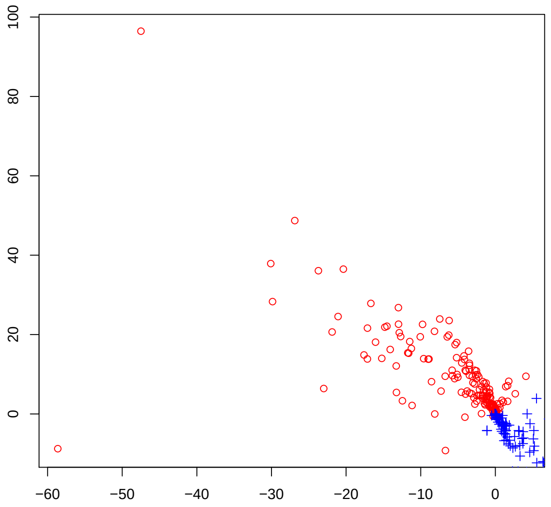


Figure 5.7: The scatterplot of two explicit features x_1^3 vs $\sqrt{3}x_1^2x_2$ (polynomial kernel with $d = 3$) for the bivariate normal example in (5.4). Red circles and blue crosses represent two classes.

contours from both methods are ellipses close to circles, which could discriminate two classes. However, the ellipses differ in their their major and minor axes. As we have found in the simulation study in kernel PCA section, the contours of the leading eigenfunction of polynomial kernel operator of degree 2 form ellipses with the minor axis aligned with the direction of the largest variation in the data, which happens to be roughly $x_2 = -x_1$ line, which is observed in the left panel of Figure 5.9. By contrast, kernel LDA picks $x_2 = x_1$ line as the direction of the minor axis of the ellipses, which captures the largest between-class variation relative to within-class variation.

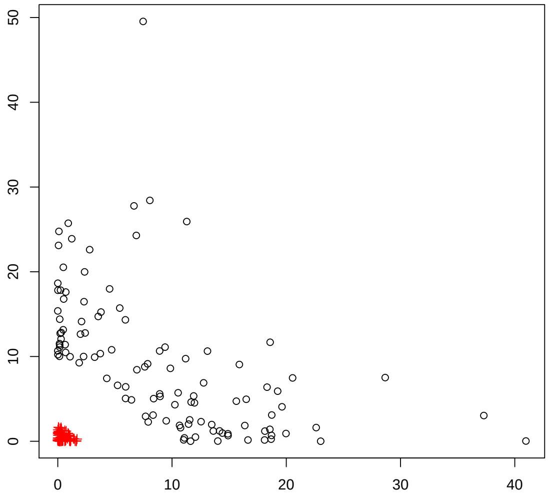


Figure 5.8: The scatterplot of two features x_1^2 vs x_2^2 (polynomial kernel with $d = 2$) for the "wheel" data. Black circles and red crosses represent the outer circle and inner cluster in the original data.

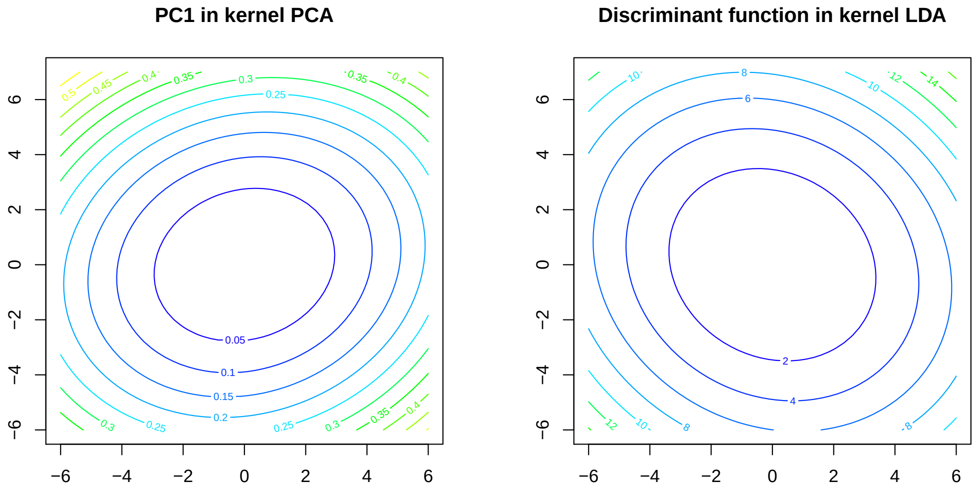


Figure 5.9: Comparison between the first principal component of kernel PCA and the discriminant function of kernel LDA for polynomial kernel with $d = 2$ over "wheel" data.

CHAPTER 6

CONCLUSION AND DISCUSSION

6.1 Conclusion

We have discussed several common kernel methods, as well as the results available regarding the eigen-analysis of the Gaussian kernel operator. 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.

While it is common practice to center the data in standard PCA, it is natural to do centering in the feature space when it comes to kernel PCA. The centered kernel is what's obtained in this process. We compare the eigen-analysis for centered and uncentered kernel matrix through numerical examples. Both centered polynomial kernel function and Gaussian kernel function of the population version are investigated to see the effect of centering kernels on the spectral property of centered kernel operator. Mercer's theorem is used to provide general explanation of the corresponding theoretical eigen-analysis. In general, we do not see much advantage in centering. Our analysis suggests that centering is not recommended when clustering is of our main concern.

As another popular kernel method, kernel LDA has shown its good performance in many applications regarding classification. By generalizing the eigen-analysis of kernel PCA, we conduct similar analysis with respect to kernel LDA. The kernel LDA operator is defined through maximizing between-class variation with respect to within-class variation, which brings the discriminant function our target function. We mainly focus on polynomial kernel function and derive in both two-dimensional setting and multi-dimensional setting. Numerical examples are given to illustrate our theoretical direction in contours, which is compared with the empirical discriminant function based on the kernel LDA algorithm. We comment on the effect of degree on the classification result in relation to the data distribution.

6.2 Discussion

As we have observed in the wheels data plot, the projections tend to be more heavily influenced by the outliers as we increase the degree of polynomial kernel. This might be due to the sensitivity to outlying observations of the classical PCA method. Correspondingly, applying PCA in the feature space through feature mapping might result in the same issue or worse. 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. (2011) and Hubert et al. (2005) could be extended to kernel PCA. Hubert et al. (2005) proposed a ROBPCA method to look for the principal components with the existence of outliers. The constructed algorithm combines the ideas of projection pursuit techniques and robust covariance estimation. Candes et al. (2011) proposed a robust PCA algorithm which focused mainly on the large data matrix which can be decomposed into a low rank component and a sparse component. Such algorithm is useful in the application including video surveillance, face recognition, web search indexing and so on. We hope that the research in robust PCA will serve as a base for future studies on robust kernel PCA.

El Karoui (2010) mentioned that in high dimension settings, kernel matrix essentially can be approximated by some matrix similar to the sample covariance matrix. Hence, the properties of the sample covariance matrix can be extended to the kernel matrix in such setting. This implies that kernel PCA in high dimensions essentially behaves like the linear PCA, which goes against what people originally thought when coming up with this method. In our simulation studies, we have tried kernel PCA in low dimensions, while we also explored the high-dimensional handwritten digit data case. To fill the gap between low dimensional study and high dimensional study, we intend to try kernel PCA in a moderate dimensional setting in an effort to understand kernel PCA from a comprehensive point of view.

As a related but different approach to nonlinear classification with kernels, Suykens et al. (2002), De Brabanter et al. (2010), and Karsmakers et al. (2011) considered a variant of the SVM called the least square support vector machines and used the Nyström method to numerically approximate a feature map from the eigen-analysis of the kernel matrix for large-scale applications. We note that the approximate feature map used in the references has closer ties to the eigenfunctions of the kernel operator for kernel PCA than kernel LDA.

Appendix A

PROOF FOR THE FORM OF LEADING EIGENFUNCTIONS IN A SIMPLE CENTERED DATA EXAMPLE

In Figure 3.2, we observe that the contour plots of eigenfunctions for the first two leading eigenvalues exhibit interesting patterns. As was mentioned in section 3.2.2, it appears that the contours of the leading eigenfunction form ellipses centered at the origin. The minor axes of the ellipses for the leading eigenfunction indicate the direction capturing the largest data variation. 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.

To theoretically illustrate this finding, we start with a simple example.

1. **Simplified setting 1**

   Suppose that (X_1, X_2) follow the bivariate normal distribution

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & k\sigma^2 \end{pmatrix}\right).$$

Given that we apply the polynomial kernel of degree 2, the moment matrix can be obtained as follows based on Theorem 1,

$$M_2^2 = \begin{bmatrix} \mu_{4,0} & \sqrt{2}\mu_{3,1} & \mu_{2,2} \\ \sqrt{2}\mu_{3,1} & 2\mu_{2,2} & \sqrt{2}\mu_{1,3} \\ \mu_{2,2} & \sqrt{2}\mu_{1,3} & \mu_{0,4} \end{bmatrix} = \begin{bmatrix} 3\sigma^4 & 0 & k\sigma^4 \\ 0 & 2k\sigma^4 & 0 \\ k\sigma^4 & 0 & 3k^2\sigma^4 \end{bmatrix}.$$

To compute the eigenvalues of this moment matrix , which also form the eigenvalues of the polynomial kernel operator, we solve $|M_2^2 - \lambda I| = 0$,

$$\begin{aligned} |M_2^2 - \lambda I| &= \begin{vmatrix} 3\sigma^4 - \lambda & 0 & k\sigma^4 \\ 0 & 2k\sigma^4 - \lambda & 0 \\ k\sigma^4 & 0 & 3k^2\sigma^4 - \lambda \end{vmatrix} \\ &= (3\sigma^4 - \lambda)\left[(2k\sigma^4 - \lambda)(3k^2\sigma^4 - \lambda)\right] + k\sigma^4\left[k\sigma^4(\lambda - 2k\sigma^4)\right] \\ &= (\lambda - 2k\sigma^4)\left[(\lambda - 3\sigma^4)(\lambda - 3k^2\sigma^4) - k^2\sigma^8\right] = 0. \end{aligned}$$

Hence, three eigenvalues of this moment matrix are

$$\begin{aligned} \lambda_1 &= 2k\sigma^4, \\ \lambda_2 &= \frac{(3k^2\sigma^4 + 3\sigma^4) + \sqrt{(3k^2\sigma^4 + 3\sigma^4)^2 - 32k^2\sigma^8}}{2}, \\ \lambda_3 &= \frac{(3k^2\sigma^4 + 3\sigma^4) - \sqrt{(3k^2\sigma^4 + 3\sigma^4)^2 - 32k^2\sigma^8}}{2}. \end{aligned}$$

To determine the order of those three eigenvalues, we need to make the comparison among them. Consider λ_2 and λ_1 first. We write $\lambda_2 = \frac{3}{2}k^2\sigma^4 + \frac{3}{2}\sigma^4 + \Delta$, where Δ is a non-negative component. Notice that we have

$$\frac{3}{2}k^2\sigma^4 + \frac{3}{2}\sigma^4 - 2k\sigma^4 = \frac{1}{2}(3k^2 - 4k + 3)\sigma^4 = \frac{1}{2}\left[3(k - \frac{2}{3})^2 + \frac{15}{9}\right]\sigma^4 > 0.$$

Given that λ_2 has a non-negative component added, we can say that λ_2 is the largest eigenvalue. Now we should compare λ_1 and λ_3. We have

$$\lambda_3 - \lambda_1 = \frac{(3k^2+3) - \sqrt{9(k^2+1)^2 - 32k^2}}{2}\sigma^4 - 2k\sigma^4.$$

Comparing λ_3 and λ_1 will lead to the following step to check whether $3k^2 + 3 - \sqrt{9(k^2+1)^2 - 32k^2} >$ or $< 4k$. We have $(3k^2-4k+3)^2 = 9k^4 - 24k^3 + 34k^2 - 24k + 9$ and $9(k^2+1)^2 - 32k^2 = 9k^4 - 14k^2 + 9$, then $(3k^2-4k+3)^2 - [9(k^2+1)^2 - 32k^2] = -24k(k-1)^2 < 0$ (we have $k > 0$, since $k\sigma^2$ is the variance of X_2).

Therefore, λ_1 is larger than λ_3. The order of three eigenvalues are then determined as follows,

$$\begin{aligned}
\lambda_{(1)} &= \frac{(3k^2\sigma^4 + 3\sigma^4) + \sqrt{(3k^2\sigma^4 + 3\sigma^4)^2 - 32k^2\sigma^8}}{2} \\
\lambda_{(2)} &= 2k\sigma^4 \\
\lambda_{(3)} &= \frac{(3k^2\sigma^4 + 3\sigma^4) - \sqrt{(3k^2\sigma^4 + 3\sigma^4)^2 - 32k^2\sigma^8}}{2}.
\end{aligned}$$

Now, we can obtain the eigenfunction corresponding to certain eigenvalue. For each eigenfunction and eigenvalue pairs, we have the following equations

$$\begin{cases}
3\sigma^4 C_0 + k\sigma^4 C_2 & = \lambda C_0 \\
2k\sigma^4 C_1 & = \lambda C_1 \\
k\sigma^4 C_0 + 3k^2\sigma^4 C_2 & = \lambda C_2.
\end{cases}$$

For $\lambda_{(1)}$ and $\lambda_{(3)}$, C_1 is obviously zero. Thus, the leading eigenfunction form would be $\phi_1 = \dfrac{C_0 x_1^2 + C_2 x_2^2}{\lambda_{(1)}}$.

For $\lambda_{(2)}$, by plugging $\lambda_{(2)} = 2k\sigma^4$ into the equations, we can tell that the corresponding pair C_0 and C_2 are both zero. Thus given that C_0, C_1 and

C_2 satisfy the orthonormal condition, we have $C_1 = 1$. The second leading eigenfunction is of the form $\phi_2 = \dfrac{2x_1x_2}{\lambda_{(2)}}$. This form suggests that the contours of the second leading eigenfunction are hyperbolas with the asymptotes in the direction of X_1 and X_2.

Now we want to investigate the shape of the leading eigenfunction corresponding to the eigenvalue $\lambda_{(1)}$. We know for eigenvalue $\lambda_{(1)}$, eigenvector pair (C_0, C_1, C_2) satisfy the condition $3\sigma^4 C_0 + k\sigma^4 C_2 = \lambda_{(1)} C_0$. By plugging $\lambda_{(1)}$ into this equation, we have the following equation

$$3\sigma^4 C_0 + k\sigma^4 C_2 = \frac{(3k^2\sigma^4 + 3\sigma^4) + \sqrt{(3k^2\sigma^4 + 3\sigma^4)^2 - 32k^2\sigma^8}}{2} C_0.$$

Since σ^4 cannot be zero, it can be canceled out from both sides of the equation. The above equation can be simplified as the following form

$$\left(\sqrt{9(k^2+1)^2 - 32k^2} + 3k^2 - 3\right) C_0 = 2kC_2.$$

We can thus determine the magnitude of C_0 and C_2 based on the above equation. By performing the calculation, we can show that when $k^2 > 1$, which means $k > 1$ (since $k > 0$), $\sqrt{9(k^2+1)^2 - 32k^2} - 3k^2 + 3 > 2k$. Notice that $k > 1$ means the variance of X_2 is greater than the variance of X_1. This condition leads to the conclusion that the magnitude of C_0 is smaller than the magnitude of C_2. Based on the form of the first leading eigenfunction, we know that when the magnitude of C_0 is smaller than the magnitude of C_2, the contours form the ellipses with the major axes along the X_1 direction and minor axes along the X_2 direction. On the contrary, when $k < 1$ indicating that the variance of X_2 is smaller than the variance of X_1, the contours are ellipses with the major axes along the X_2 direction.

In summary, we conclude that the contours of the first leading eigenfunction are of ellipses with the minor axes indicating the largest variance direction in the data. The contours of the second leading eigenfunction form hyperbolas with their asymptotes the same as the major and minor axes of the first leading eigenfunction.

2. **Simplified setting 2**

   By rotating this single component, the proof can be generalized to the rotated bivariate normal case.

   Suppose our data is $X = [X_1, X_2]^{'}$ and the variance can be decomposed as follows $\text{var}(X) = \Sigma = V\Lambda V^{'}$, Λ is a matrix only having values along its diagonal. We want to rotate the data so that the covariance matrix has the diagonal form as in the setting 1. To accomplish this, we can pre-multiply X by V^{-1} so that $\text{var}(V^{-1}X) = V^{-1}\text{var}(X)V = V^{-1}V\Lambda V^{-1}V = \Lambda$.

   Now, suppose that $V^{-1} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$ and we thus have

$$V^{-1}X = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix} X = \begin{bmatrix} v_{11}X_1 + v_{12}X_2 \\ v_{21}X_1 + v_{22}X_2 \end{bmatrix}.$$

   Let $[X_1^*, X_2^*]^{'}$ be the rotated data with the covariance matrix of the diagonal form, we have $X_1^* = v_{11}X_1 + v_{12}X_2$ and $X_2^* = v_{21}X_1 + v_{22}X_2$. In part(a), we already have conclusion for the eigenfunction form for $[X_1^*, X_2^*]^{'}$, which is $\phi_1^* = \dfrac{c_1^* x_1^{*2} + c_2^* x_2^{*2}}{\lambda_1^*}$ and $\phi_2^* = \dfrac{2x_1^* x_2^*}{\lambda_2^*}$. Therefore, by plugging X_1^* and X_2^* into the above equation, we can obtain the exact form of the eigenfunctions. Those eigenfunctions are actually the rotated ellipses or hyperbolas.

On the other hand, the relationship between X^* and X can lead to the following,

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \frac{-v_{21}}{v_{11}v_{22}-v_{12}v_{21}} & \frac{v_{11}}{v_{11}v_{22}-v_{12}v_{21}} \\ \frac{v_{22}}{v_{11}v_{22}-v_{12}v_{21}} & \frac{-v_{12}}{v_{11}v_{22}-v_{12}v_{21}} \end{bmatrix} \begin{bmatrix} X_1^* \\ X_2^* \end{bmatrix} = V^* \begin{bmatrix} X_1^* \\ X_2^* \end{bmatrix}.$$

We consider the matrix V^* as the matrix which rotates back to our original data case.

Appendix B

EXAMPLE FOR THE CENTERED DATA WITH CENTERED KERNEL

- $m = 0$

  The moment matrix is

$$M_2^2 = \begin{bmatrix} 3 & 0 & 2 \\ 0 & 4 & 0 \\ 2 & 0 & 12 \end{bmatrix}.$$

  Thus, the three corresponding eigenfunctions are

$$\phi_1(\mathbf{x}) = \frac{-.208x_1^2 - .978x_2^2}{12.424}, \quad \phi_2(\mathbf{x}) = \frac{-\sqrt{2}x_1x_2}{4}, \quad \text{and} \quad \phi_3(\mathbf{x}) = \frac{.978x_1^2 - .208x_2^2}{2.576}.$$

  The minor axis of contours of the first eigenfunction points to the direction of the maximum variation in the data.

- $m = 0.5$

  The corresponding moment matrix is

$$M_2^2 = \begin{bmatrix} 4.5625 & 0 & 2.5 \\ 0 & 5 & 0 \\ 2.5 & 0 & 12 \end{bmatrix}.$$

Thus, the three corresponding eigenfunctions are

$$\phi_1(\mathbf{x}) = \frac{-.292x_1^2 - .957x_2^2}{12.762}, \quad \phi_2(\mathbf{x}) = \frac{-\sqrt{2}x_1x_2}{5}, \quad \text{and} \quad \phi_3(\mathbf{x}) = \frac{.957x_1^2 - .292x_2^2}{3.8}.$$

Similar to $m = 0$ case, contours of the top eigenfunction are ellipses with the major axis along x_1-axis.

- $m = 1$

The corresponding moment matrix is

$$M_2^2 = \begin{bmatrix} 10 & 0 & 4 \\ 0 & 8 & 0 \\ 4 & 0 & 12 \end{bmatrix}.$$

The three corresponding eigenfunctions are

$$\phi_1(\mathbf{x}) = \frac{.615x_1^2 + .788x_2^2}{15.123}, \quad \phi_2(\mathbf{x}) = \frac{-\sqrt{2}x_1x_2}{8}, \quad \text{and} \quad \phi_3(\mathbf{x}) = \frac{.788x_1^2 - .615x_2^2}{6.877}.$$

Although the coefficients of x_1^2 and x_2^2 are quite close to each other, the leading eigenfunction still give ellipses with the major axis along x_1-axis.

- $m = 1.1147$

We pick the value such that the contours of the leading eigenfunction form circles since this exhibits the transition observed in the contours. The moment matrix is given as follows,

$$M_2^2 = \begin{bmatrix} 11.9993 & 0 & 4.4851 \\ 0 & 8.9702 & 0 \\ 4.4851 & 0 & 12 \end{bmatrix}.$$

This moment matrix leads to the following eigenfunctions.

$$\phi_1(\mathbf{x}) = \frac{.707x_1^2 + .707x_2^2}{16.485}, \quad \phi_2(\mathbf{x}) = \frac{-\sqrt{2}x_1x_2}{8.97}, \quad \text{and} \quad \phi_3(\mathbf{x}) = \frac{.707x_1^2 - .707x_2^2}{7.515}.$$

- $m = 2$

  As we move the center of the distribution to (2,0), the moment matrix becomes

  $$M_2^2 = \begin{bmatrix} 43 & 0 & 10 \\ 0 & 20 & 0 \\ 10 & 0 & 12 \end{bmatrix}.$$

  The three eigenfunctions are

  $$\phi_1(\mathbf{x}) = \frac{.959x_1^2 + .283x_2^2}{45.946}, \quad \phi_2(\mathbf{x}) = \frac{\sqrt{2}x_1x_2}{20}, \quad \text{and} \quad \phi_3(\mathbf{x}) = \frac{.283x_1^2 - .959x_2^2}{9.054}.$$

  Contrary to the other cases obtained so far, contours of the top eigenfunction form ellipses with the major axis along x_2-axis.

- $m = 5$

  Moving the center of the data distribution even further to (5,0) leads to the following moment matrix,

  $$M_2^2 = \begin{bmatrix} 778 & 0 & 52 \\ 0 & 104 & 0 \\ 52 & 0 & 12 \end{bmatrix}.$$

  The three eigenfunctions are then given by,

  $$\phi_1(\mathbf{x}) = \frac{.998x_1^2 + .067x_2^2}{781.514}, \quad \phi_2(\mathbf{x}) = \frac{\sqrt{2}x_1x_2}{104}, \quad \text{and} \quad \phi_3(\mathbf{x}) = \frac{.067x_1^2 - .998x_2^2}{8.486}.$$

  Compared to $m = 2$ case, contours of the leading eigenfunction are more elongated with the major axis along x_2-axis. The examples with $m = 2$ and $m = 5$ show a different pattern from other choices of m that the major axis points to the direction of largest variation in the data instead.

Appendix C

REMARKS ON $\mathcal{K}_P$ BEING A VALID MAPPING

When we introduced the kernel operator, we have stated that $\mathcal{K}_p$ is a mapping from $\mathcal{H}_K$ to $\mathcal{H}_K$. In order to show that this mapping is valid with the polynomial kernel function, we provide the following justification.

Based on our analysis in the above section, we have shown that the eigenfunction is of the form

$$f(x) = \sum_{k=0}^{2} \binom{2}{k} C_k x_1^k x_2^{2-k}$$

Since we have kernel function of the form $K(x, y) = (x^t y)^2, x, y \in \mathbb{R}^2$, the following equations hold.

$$\begin{aligned}
\mathcal{K}_p(f) &= \int K(x, y) f(y) p(y) dy \\
&= \int (x^t y)^2 \sum_{k=0}^{2} \binom{2}{k} C_k x_1^k x_2^{2-k} p(y) dy \\
&= \int \sum_{j=0}^{2} \binom{2}{j} (x_1 y_1)^j (x_2 y_2)^{2-j} \sum_{k=0}^{2} \binom{2}{k} C_k x_1^k x_2^{2-k} p(y) dy \\
&= \sum_{j=0}^{2} \binom{2}{j} x_1^j x_2^{2-j} \int y_1^j y_2^{2-j} \sum_{k=0}^{2} C_k y_1^k y_2^{2-k} p(y) dy
\end{aligned}$$

While $\int y_1^j y_2^{2-j} \sum_{k=0}^{2} C_k y_1^k y_2^{2-k} p(y) dy$ should be able to be integrated out as a constant, we can show that $\mathcal{K}_p(f) \in \mathcal{H}_K$ as long as we show that $x_1^2, x_2^2, x_1 x_2 \in \mathcal{H}_K$.

Based on the properties of the reproducing kernel Hilbert space, if x_1^2, x_2^2, x_1x_2 can all be expressed in the form of $\sum_{i=1}^{n} c_i K(x, x_i^*), x_i^* \in \mathbb{R}^2, c_i \in \mathbb{R}$, we can say that they lie in the space $\mathcal{H}_K$. Then the mapping $\mathcal{K}_p(f)$ lies in the space $\mathcal{H}_K$ as well.

$$x_1^2 = (\langle (x_1, x_2), (1, 0)\rangle)^2 = K\left((x_1, x_2), (1, 0)\right)$$

Similarly, we have

$$x_2^2 = (\langle (x_1, x_2), (0, 1)\rangle)^2 = K\left((x_1, x_2), (0, 1)\right)$$

Besides, we have

$$\begin{aligned} x_1x_2 &= \frac{1}{2}(\langle (x_1, x_2), (1, 1)\rangle)^2 - \frac{1}{2}(x_1^2 + x_2^2) \\ &= \frac{1}{2}K((x_1, x_2), (1, 1)) - \frac{1}{2}[K((x_1, x_2), (0, 1)) + K((x_1, x_2), (1, 0))]. \end{aligned}$$

where $c_1 = \frac{1}{2}, c_2 = c_3 = -\frac{1}{2}$ and $x_1^* = (1, 1), x_2^* = (0, 1), x_3^* = (1, 0)$ in the $\sum_{i=1}^{n} c_i K(x, x_i^*)$.

Given what's shown above, we can say that $\mathcal{K}_p$ is a valid mapping from $\mathcal{H}_K$ to $\mathcal{H}_K$.

BIBLIOGRAPHY

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

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.

C.T.H. Baker. *The numerical treatment of integral equations*, volume 13. Clarendon press Oxford, 1977.

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? *Journal of ACM*, 58(3):1–37, 2011.

K. De Brabanter, J. De Brabanter, J. A. K. Suykens, and B. De Moor. Optimized fixed-size kernel models for large data sets. *Computational Statistics Data Analysis*, 54(6):1484–1504, 2010.

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):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.

P. Karsmakers, K. Pelckmans, K. De Brabanter, Van hamme H., and J. A. K. Suykens. Sparse conjugate directions pursuit with application to fixed-size kernel models. *Machine learning*, 85(1):109–148, 2011.

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.

R. Koekoek and R.F. Swarttouw. The askey scheme of hypergeometric orthogonal polynomials and its q-analogue, fac. techn. math. informatics. *Delft University of Technology, Report*, pages 98–17, 1998.

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. Nadler and M. Galun. Fundamental limitations of spectral clustering. *Advances in Neural Information Processing Systems*, 19:1017, 2007.

P. Perona and W. Freeman. A factorization approach to grouping. *Computer VisionECCV'98*, pages 655–670, 1998.

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*, 10:1299–1319, 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.

J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

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

J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

V. Vapnik and A. Chervonenkis. A note on one class of perceptrons. *Automation and Remote Control*, 25(1), 1964.

V. Vapnik and A.J. Lerner. Generalized portrait method for pattern recognition. *Automation and Remote Control*, 24(6):774–780, 1963.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 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. In C. Bishop, editor, *Neural Networks and Machine Learning*, pages 167–184. Springer, Berlin, 1998.