The Geometry of Nonlinear Embeddings in Kernel Discriminant Analysis

Jiae Kim* Yoonkyung Lee† Zhiyu Liang‡

Abstract

Fisher's linear discriminant analysis is a classical method for classification, yet it is limited to capturing linear features only. Kernel discriminant analysis as an extension is known to successfully alleviate the limitation through a nonlinear feature mapping. We study the geometry of nonlinear embeddings in discriminant analysis with polynomial kernels and Gaussian kernel by identifying the population-level discriminant function that depends on the data distribution and the kernel. In order to obtain the discriminant function, we solve a generalized eigenvalue problem with between-class and within-class covariance operators. The polynomial discriminants are shown to capture the class difference through the population moments explicitly. For approximation of the Gaussian discriminant, we use a particular representation of the Gaussian kernel by utilizing the exponential generating function for Hermite polynomials. We also show that the Gaussian discriminant can be approximated using randomized projections of the data. Our results illuminate how the data distribution and the kernel interact in determination of the nonlinear embedding for discrimination, and provide a guideline for choice of the kernel and its parameters.

Keywords: Discriminant analysis, Feature map, Gaussian kernel, Polynomial kernel, Rayleigh quotient, Spectral analysis

1 Introduction

Kernel methods have been widely used in statistics and machine learning for pattern recognition and analysis (Hofmann et al. 2008, Schölkopf and Smola 2002, Shawe-Taylor and Cristianini 2004). They can be described in a unified framework with a special class of functions called *kernels* encoding pairwise similarities between data points. Such kernels enable nonlinear extensions of linear methods seamlessly and allow us to deal with general types of data such as vectors, text documents, graphs, and images. Combined with problem-specific evaluation criteria typically in the form of a loss function or a spectral norm of a kernel matrix, this kernel-based framework can produce a variety of learning algorithms for regression, classification, ranking, clustering, and dimension reduction. Popular kernel methods include smoothing splines (Wahba 1990), support vector machines (Vapnik 1995), kernel Fisher discriminant analysis (Baudat and Anouar 2000, Mika et al. 1999), ranking SVM (Joachims 2002), spectral clustering (Scott and Longuet-Higgins 1990, von Luxburg 2007), and kernel principal component analysis (Schölkopf et al. 1998).

This paper regards the geometry of kernel discriminant analysis (KDA). KDA is a nonlinear generalization of Fisher's linear discriminant analysis (LDA), which is a standard multivariate technique for classification. Intrinsically as a dimension reduction method, KDA looks for discriminants that embed multivariate

*Jiae Kim is Graduate Student, Department of Statistics, The Ohio State University, Columbus, OH 43210 (Email: kim.3887@osu.edu).

†Yoonkyung Lee is Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210 (Email: yklee@stat.osu.edu).

‡Zhiyu Liang is Senior Staff Data Scientist, Kohl's, Milpitas, CA 95035 (Email: zhiyu.liang@kohls.com).

data into a real line so that decisions can be made easily in a low dimensional space. For simplicity of exposition, we focus on the case of two classes. Fisher's linear discriminant projects data along the direction that maximizes separation between classes. Extending this geometric idea, kernel discriminant analysis finds a data embedding that maximizes the ratio of the between-class variation to within-class variation measured in the feature space specified by a kernel. To determine the embedding as a discriminant, we solve a generalized eigenvalue problem involving kernel-dependent covariance matrices.

We examine the kernel discriminant at the population level to illuminate the interplay between the kernel and the probability distribution for data. Of particular interest is how the kernel discriminant captures the difference between two classes geometrically, and how the choice of a kernel and associated kernel parameters affect the discriminant in connection with salient features of the underlying distribution. As a continuous analogue of the kernel-dependent covariance matrices, we define the between-class and within-class covariance operators first and state the population version of the eigenvalue problem using those operators which depend on both the data distribution and the kernel. For some kernels, we can obtain explicit solutions and determine the corresponding population kernel discriminants.

Similar population-level analyses have been done for kernel PCA and spectral clustering (Liang and Lee 2013, Shi et al. 2009, Zhu et al. 1998) to gain insights into the interplay between the kernel and distributional features on low dimensional embeddings for data visualization and clustering. The population analyses of kernel PCA, spectral clustering, and KDA require a spectral analysis of kernel operators of different forms depending on the method. They help us examine the dependence of eigenfunctions and eigenvalues of the kernel operators on the data distribution, which can guide applications of those methods in practice.

The population discriminants with polynomial kernels admit a closed-form expression due to their finite dimensional feature map. Analogous to the geometric interpretation of Fisher's linear discriminant that it projects data along the mean difference direction after whitening the within-class covariance, the polynomial discriminants are characterized by the difference in the population moments between classes. By contrast, the Gaussian kernel does not allow a simple closed-form expression for the discriminant because its feature map and associated function space are infinite-dimensional. We provide approximations to the Gaussian discriminant instead using two representations of the kernel. These approximations shed some light on the workings of KDA with the Gaussian kernel. By using a deterministic representation of the Gaussian kernel with the Hermite polynomial generating function, we approximate the population Gaussian discriminant with polynomial discriminants of degree as high as desired for the accuracy of approximation. This implies that the Gaussian discriminant captures the difference between classes through the *entirety* of the moments. Alternatively, using a stochastic representation of the Gaussian kernel through Fourier features of random projections (Rahimi and Recht 2008a), we can also view the Gaussian discriminant as an embedding that combines the expected differences in sinusoidal features of randomly projected data from two classes.

How are the forms of these population kernel discriminants related to the task of minimizing classification error? To attain the least possible error rate, the optimal decision rule assigns a data point $\mathbf{x} \in \mathbb{R}^p$ to the most probable class by comparing the likelihood of one class, say $p_1(\mathbf{x})$, versus the other, $p_2(\mathbf{x})$, given $\mathbf{x}$. In other words, the ideal data embedding for discrimination of two classes should be based on the likelihood ratio $p_1(\mathbf{x})/p_2(\mathbf{x})$ or $\log[p_1(\mathbf{x})/p_2(\mathbf{x})]$. As a simple example, when the population distribution for each class is multivariate normal with a common covariance matrix, $\log[p_1(\mathbf{x})/p_2(\mathbf{x})]$ is linear in $\mathbf{x}$, and it coincides with the population version of Fisher's linear discriminant. Difference in the covariance brings additional quadratic terms to the log likelihood ratio requiring a quadratic discriminant for the lowest error. As the distributions further deviate from elliptical scatter patterns exemplified by normal distributions, the ideal data embedding according to $\log[p_1(\mathbf{x})/p_2(\mathbf{x})]$ will involve nonlinear terms beyond quadratic. The basic fact that each distribution can be identified with its moment-generating function or characteristic function, i.e., its Fourier transform, implies that any difference between two distributions can be described in terms of the moments or expected Fourier features in general. Our population analysis of kernel discriminants indicates that the Gaussian kernel treats the distributional difference as a whole, including both global

and local (or low and high frequency) characteristics, while the polynomial kernels focus on differences in more global characteristics represented by low-order moments. The ideal choice of a kernel in KDA will inevitably depend on the mode of class difference mathematically expressed through the log likelihood ratio, $\log[p_1(\mathbf{x})/p_2(\mathbf{x})]$.

The rest of the paper is organized as follows. Section 2 provides a brief review of kernel discriminant analysis and describes its population version by introducing two kernel covariance operators for measuring the between-class variation and within-class variation in the feature space. Section 3 presents a population-level discriminant analysis using two types of polynomial kernels and Gaussian kernel and provides an explicit form of population kernel discriminants. Numerical examples are given in Section 4 to illustrate the geometry of kernel discriminants in relation to the data distribution. Section 5 concludes the paper with discussions.

2 Preliminaries

This section provides a technical background for kernel discriminant analysis. After reviewing kernel functions, corresponding function spaces, and feature mappings in Section 2.1, we briefly describe Fisher's linear discriminant analysis and its extension using kernels in Section 2.2 and further extend the sample-dependent description of kernel discriminant analysis to its population version in Section 2.3.

2.1 Kernel

Let the input domain for data be denoted by $\mathcal{X}$. A kernel $K(\cdot,\cdot)$ is defined as a positive semi-definite function from $\mathcal{X}\times\mathcal{X}$ to $\mathbb{R}$. As a positive semi-definite function, K is symmetric: $K(\mathbf{x},\mathbf{u}) = K(\mathbf{u},\mathbf{x})$ for all $\mathbf{x},\mathbf{u}\in\mathcal{X}$, and for each $n\in\mathbb{N}$ and for all choices of $\mathbf{x}_1,\ldots,\mathbf{x}_n\in\mathcal{X}$, $K_n = [K(\mathbf{x}_i,\mathbf{x}_j)]$ as an $n\times n$ matrix is positive semi-definite.

Given K, there is a unique function space $\mathcal{H}_K$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$ corresponding to the kernel such that for every $\mathbf{x}\in\mathcal{X}$ and $f\in\mathcal{H}_K$, (i) $K(\mathbf{x},\cdot)\in\mathcal{H}_K$, and (ii) $f(\mathbf{x}) = \langle f, K(\mathbf{x},\cdot)\rangle_{\mathcal{H}_K}$. The second property is called the reproducing property of K, and it entails the following identity: $K(\mathbf{x},\mathbf{u}) = \langle K(\mathbf{x},\cdot), K(\mathbf{u},\cdot)\rangle_{\mathcal{H}_K}$. Such a function space with reproducing kernel is called a reproducing kernel Hilbert space (RKHS). See Aronszajn (1950), Wahba (1990) and Gu (2002) for reference.

Alternatively, kernels can be characterized as those functions that arise as a result of the dot product of feature vectors. This is a common viewpoint in machine learning in the use of kernels for nonlinear generalization of linear methods. To capture nonlinear features often desired for data analysis, consider a mapping $\boldsymbol{\phi}$ from the input space $\mathcal{X}$ to a feature space $\mathcal{F} = \mathbb{R}^D$, $\boldsymbol{\phi}:\mathcal{X}\to\mathcal{F}$, which is called a feature map. The feature vector $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}),\ldots,\phi_D(\mathbf{x}))^T$ consists of D features, and for expressiveness of the features, the dimension of the feature space is often much higher than the input dimension, and possibly infinite. Through the dot product of feature vectors, we can define a valid kernel K on $\mathcal{X}\times\mathcal{X}$ as $K(\mathbf{x},\mathbf{u}) = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{u})$. When $D=\infty$, the dot product is to be interpreted in the sense of ℓ_2 inner product. More general treatment of kernels with a general inner product for the feature space is feasible, but for brevity, we confine our description to the dot product only. Using a feature map, we can generalize a linear method by applying it in the feature space, which amounts to replacing the dot product for the original features, $\mathbf{x}^T\mathbf{u}$, in the linear method with a kernel, $K(\mathbf{x},\mathbf{u})$. This substitution is called the "kernel trick" in machine learning. For general description of kernel methods, an explicit form of a feature map is not needed nor the feature map for a given kernel is unique. See Schölkopf and Smola (2002) for general properties of kernels.

In this paper, we focus on the following kernels that are commonly used in practice with $\mathcal{X} = \mathbb{R}^p$:

- Homogeneous polynomial kernel of degree d: $K_d(\mathbf{x},\mathbf{u}) = (\mathbf{x}^T\mathbf{u})^d$

- Inhomogeneous polynomial kernel of degree d: $\tilde{K}_d(\mathbf{x}, \mathbf{u}) = (1 + \mathbf{x}^T\mathbf{u})^d$
- Gaussian kernel with bandwidth parameter ω: $K_\omega(\mathbf{x}, \mathbf{u}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{u}\|^2}{2\omega^2}\right)$.

Consideration of their explicit feature maps will be useful for the analyses presented in Section 3. For instance, the homogeneous polynomial kernel of degree 2 on $\mathcal{X} = \mathbb{R}^2$, $K_2(\mathbf{x}, \mathbf{u}) = (x_1u_1 + x_2u_2)^2$, can be described with a feature map $\boldsymbol{\phi}(\mathbf{x}) = \left(x_1^2,\ \sqrt{2}x_1x_2,\ x_2^2\right)^T \in \mathbb{R}^3$. The Gaussian kernel on $\mathbb{R}$ with bandwidth parameter 1 admits $\mathcal{F} = \mathbb{R}^\infty$ with ℓ_2 inner product as a feature space and a feature map of

$$\boldsymbol{\phi}(x) = e^{-\frac{x^2}{2}}\left(1,\ x,\ \frac{x^2}{\sqrt{2!}},\ \frac{x^3}{\sqrt{3!}},\ \ldots\right)^T.$$

2.2 Kernel Discriminant Analysis

Kernel discriminant analysis (KDA) is a nonlinear extension of Fisher's linear discriminant analysis using kernels. For description of KDA, we start with a classification problem. Suppose we have data from two classes labeled 1 and 2: $\{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathcal{X} \text{ and } y_i \in \{1, 2\} \text{ for } i = 1, \ldots, n\}$. For simplicity, assume that the data points are ordered so that the first n_1 observations are from class 1 and the rest ($n_2 = n - n_1$) are from class 2.

2.2.1 Fisher's Linear Discriminant Analysis

As a classical approach to classification, Fisher's linear discriminant analysis (LDA) looks for linear combinations of the original variables called *linear discriminants* that can separate observations from different classes effectively. It can be viewed as a dimension reduction technique for classification.

When $\mathcal{X} = \mathbb{R}^p$, a linear discriminant is of the form, $f(\mathbf{x}) = \mathbf{v}^T\mathbf{x}$, with a coefficient vector $\mathbf{v} \in \mathbb{R}^p$. For the discriminant $\mathbf{v}^T\mathbf{x}$ as a univariate measurement, we define the between-class variation as

$$(\mathbf{v}^T\overline{\mathbf{x}}_1 - \mathbf{v}^T\overline{\mathbf{x}}_2)^2 = \mathbf{v}^T(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^T\mathbf{v}$$

and the within-class variation as its pooled sample variance:

$$\frac{n_1}{n}\mathbf{v}^T S_1\mathbf{v} + \frac{n_2}{n}\mathbf{v}^T S_2\mathbf{v} = \mathbf{v}^T\left(\frac{n_1}{n}S_1 + \frac{n_2}{n}S_2\right)\mathbf{v},$$

where $\overline{\mathbf{x}}_j$ and S_j are the sample mean vector and sample covariance matrix of $\mathbf{x}$ for class j. Letting $S_B = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^T$ and $S_W = \frac{n_1}{n}S_1 + \frac{n_2}{n}S_2$ (the pooled covariance matrix), we can express the two variations succinctly as quadratic forms of $\mathbf{v}^T S_B\mathbf{v}$ and $\mathbf{v}^T S_W\mathbf{v}$, respectively. Note that both forms are shift-invariant.

To find the best direction that gives the maximum separation between two classes measured relative to the within-class variance in LDA, we maximize the ratio of the between-class variation to the within-class variation with respect to $\mathbf{v}$:

$$\underset{\mathbf{v}\in\mathbb{R}^p}{\text{maximize}}\frac{\mathbf{v}^T S_B\mathbf{v}}{\mathbf{v}^T S_W\mathbf{v}}.$$

This ratio is also known as the *Rayleigh quotient* and taken as a measure of the signal-to-noise ratio in classification along the direction $\mathbf{v}$. This maximization problem leads to the following generalized eigenvalue problem:

$$S_B\mathbf{v} = \lambda S_W\mathbf{v}$$

and the solution is given by the leading eigenvector. More explicitly, $\hat{\mathbf{v}} = S_W^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ defined only up to a normalization constant, and $\hat{\lambda} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^T S_W^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ is the corresponding eigenvalue. Since S_B has

rank 1, $\hat{\lambda}$ is the only positive eigenvalue. The resulting linear discriminant, $\hat{f}(\mathbf{x}) = \hat{\mathbf{v}}^T\mathbf{x}$, together with an appropriately chosen threshold c yields a classification boundary of the form $\{\mathbf{x} \in \mathbb{R}^p \mid \hat{\mathbf{v}}^T\mathbf{x} = c\}$, which is linear in the input space. When $S_W \approx I_p$, $\hat{\mathbf{v}} \approx \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2$ (mean difference) provides the best direction for projection. Re-expression of the linear discriminant as $\hat{f}(\mathbf{x}) = \hat{\mathbf{v}}^T\mathbf{x} = \left[S_W^{-\frac{1}{2}}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)\right]^T S_W^{-\frac{1}{2}}\mathbf{x}$ further reveals that LDA projects data onto the mean difference direction after *whitening* the variables via $S_W^{-\frac{1}{2}}$. This interpretation also implies the invariance of $\hat{f}(\cdot)$ under variable scaling.

2.2.2 Nonlinear Generalization

Using the aforementioned kernel trick, Mika et al. (1999) proposed a nonlinear extension of linear discriminant analysis, which can be useful when the optimal classification boundary is not linear. Conceptually, by mapping the data into a feature space using a kernel, kernel discriminant analysis finds the best direction for discrimination and corresponding linear discriminant in the feature space, which then defines a nonlinear discriminant function in the input space.

Given kernel K, let $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F}$ be a feature map. Then using the feature vector $\boldsymbol{\phi}(\mathbf{x})$, we can define the sample means and between-class and within-class covariance matrices in the feature space analogously. These matrices are denoted by $S_B^{\boldsymbol{\phi}}$ and $S_W^{\boldsymbol{\phi}}$. KDA aims to find $\mathbf{v}$ in the feature space that maximizes

$$\frac{\mathbf{v}^T S_B^{\boldsymbol{\phi}} \mathbf{v}}{\mathbf{v}^T S_W^{\boldsymbol{\phi}} \mathbf{v}}. \tag{1}$$

When $\mathbf{v}$ is in the span of all training feature vectors $\boldsymbol{\phi}(\mathbf{x}_i)$, it can be expressed as $\mathbf{v} = \sum_{i=1}^n \alpha_i \boldsymbol{\phi}(\mathbf{x}_i)$ for some $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n$. When we plug $\mathbf{v}$ of the form into the numerator and denominator of the ratio in (1) and expand both in terms of α_i using the kernel identity $K(\mathbf{x}, \mathbf{u}) = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{u})$, we have

$$\mathbf{v}^T S_B^{\boldsymbol{\phi}} \mathbf{v} = \boldsymbol{\alpha}^T B_n \boldsymbol{\alpha} \quad \text{and} \quad \mathbf{v}^T S_W^{\boldsymbol{\phi}} \mathbf{v} = \boldsymbol{\alpha}^T W_n \boldsymbol{\alpha},$$

where B_n and W_n are the $n \times n$ matrices defined through the kernel that reflect between-class variation and within-class variation, respectively. To describe B_n and W_n precisely, start with the kernel matrix $K_n = [K(\mathbf{x}_i, \mathbf{x}_j)]$. It can be partitioned into $[K_1\ K_2]$ with $n \times n_1$ matrix of K_1 and $n \times n_2$ matrix of K_2, according to the class label y_i. Using this partition of K_n, we can show that $B_n = (\bar{K}_1 - \bar{K}_2)(\bar{K}_1 - \bar{K}_2)^T$ with $\bar{K}_j = \frac{1}{n_j} K_j 1_{n_j}$ and

$$W_n = \frac{n_1}{n} K_1 \left(\frac{1}{n_1} I_{n_1} - \frac{1}{n_1^2} J_{n_1}\right) K_1^T + \frac{n_2}{n} K_2 \left(\frac{1}{n_2} I_{n_2} - \frac{1}{n_2^2} J_{n_2}\right) K_2^T,$$

where 1_{n_j} is the n_j vector of ones, and $J_{n_j} = 1_{n_j} 1_{n_j}^T$ ($n_j \times n_j$ matrix of ones).

In order to find the best discriminant direction $\mathbf{v} = \sum_{i=1}^n \alpha_i \boldsymbol{\phi}(\mathbf{x}_i)$, we maximize $\dfrac{\boldsymbol{\alpha}^T B_n \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T W_n \boldsymbol{\alpha}}$ with respect to $\boldsymbol{\alpha} \in \mathbb{R}^n$ instead. The solution is again given by the leading eigenvector of the generalized eigenvalue problem:

$$B_n \boldsymbol{\alpha} = \lambda W_n \boldsymbol{\alpha}. \tag{2}$$

Further, the estimated direction $\hat{\mathbf{v}} = \sum_{i=1}^n \hat{\alpha}_i \boldsymbol{\phi}(\mathbf{x}_i)$ results in the discriminant function of the form:

$$\hat{f}(\mathbf{x}) = \hat{\mathbf{v}}^T \boldsymbol{\phi}(\mathbf{x}) = \sum_{i=1}^n \hat{\alpha}_i \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}) = \sum_{i=1}^n \hat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}). \tag{3}$$

Obviously $\hat{f}(\cdot)$ is in the span of $K(\mathbf{x}_i, \cdot)$, $i = 1, \ldots, n$, and belongs to the reproducing kernel Hilbert space $\mathcal{H}_K$. As with Fisher's linear discriminant, the kernel discriminant function is determined only up to a normalization constant. To specify a decision rule completely, we need to choose an appropriate threshold for the discriminant function.

2.3 Population Version of Kernel Discriminant Analysis

To understand the effects of the data distribution, geometrical difference between two classes, in particular, and the kernel on the resulting discriminant function, we consider a population version of KDA. For proper description of the population version, we first assume that $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in the dataset is a random sample of $\mathbf{X}$ from a mixture of two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ with population proportions of π_1 and $\pi_2 (= 1 - \pi_1)$ for two classes, or $\mathbb{P} = \pi_1 \mathbb{P}_1 + \pi_2 \mathbb{P}_2$.

To illustrate how the sample version of KDA extends to the population version under this assumption, we begin with the eigenvalue problem in (2). Suppose λ_n and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ are a pair of eigenvalue and eigenvector satisfying (2). After scaling both sides of (2) by the sample size n, we have

$$\frac{1}{n} \sum_{j=1}^{n} B_n(i,j)\alpha_j = \frac{\lambda_n}{n} \sum_{j=1}^{n} W_n(i,j)\alpha_j \qquad \text{for} \quad i = 1, \ldots, n. \tag{4}$$

As a continuous population analogue of B_n and W_n, we can define the following bivariate functions on $\mathcal{X} \times \mathcal{X}$:

$$B_K(\mathbf{x}, \mathbf{u}) = \Big\{ \mathbb{E}_1[K(\mathbf{x}, \mathbf{X})] - \mathbb{E}_2[K(\mathbf{x}, \mathbf{X})] \Big\} \Big\{ \mathbb{E}_1[K(\mathbf{u}, \mathbf{X})] - \mathbb{E}_2[K(\mathbf{u}, \mathbf{X})] \Big\} \tag{5}$$

$$W_K(\mathbf{x}, \mathbf{u}) = \pi_1 \text{Cov}_1[K(\mathbf{x}, \mathbf{X}), K(\mathbf{u}, \mathbf{X})] + \pi_2 \text{Cov}_2[K(\mathbf{x}, \mathbf{X}), K(\mathbf{u}, \mathbf{X})], \tag{6}$$

where $\mathbb{E}_j$ and Cov_j indicate that the expectation and covariance are taken with respect to $\mathbb{P}_j$. The matrices B_n and W_n can be viewed as a sample version of $B_K(\cdot, \cdot)$ and $W_K(\cdot, \cdot)$ evaluated at all pairs of data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Further treating $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ as a discrete version of a function $\alpha(\cdot)$ at the data points, i.e., $\boldsymbol{\alpha} = (\alpha(\mathbf{x}_1), \ldots, \alpha(\mathbf{x}_n))^T$, and taking the sample class proportion, (n_j/n), as an estimate of the population proportion, π_j, and λ_n as a sample version of the population eigenvalue λ, we arrive at the following integral counterpart of (4):

$$\int_{\mathcal{X}} B_K(\mathbf{x}, \mathbf{u})\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}) = \lambda \int_{\mathcal{X}} W_K(\mathbf{x}, \mathbf{u})\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}) \quad \text{for every } \mathbf{x} \in \mathcal{X}. \tag{7}$$

This eigenvalue problem involves two integral operators: (i) the between-class covariance operator defined as

$$\mathcal{B}[\alpha(\mathbf{x})] = \int_{\mathcal{X}} B_K(\mathbf{x}, \mathbf{u})\alpha(\mathbf{u}) d\mathbb{P}(\mathbf{u}),$$

and (ii) the within-class covariance operator defined as

$$\mathcal{W}[\alpha(\mathbf{x})] = \int_{\mathcal{X}} W_K(\mathbf{x}, \mathbf{u})\alpha(\mathbf{u}) d\mathbb{P}(\mathbf{u}).$$

The form of the sample discriminant function in (3) with scaling of $1/n$ suggests that using the solution to equation (7), $\alpha(\cdot)$, we define the population discriminant function as

$$f(\mathbf{x}) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{u})\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}). \tag{8}$$

Clearly, the eigenfunction $\alpha(\cdot)$ depends on the kernel K and probability distribution $\mathbb{P}$, and so does the kernel discriminant function with $\alpha(\cdot)$ as a coefficient function. Hence, identification of the solution to the generalized eigenvalue problem in (7) will give us better understanding of kernel discriminants in relation to the data distribution and the choice of the kernel. The correspondence between the pattern of class difference and the nature of the resulting discriminant is of particular interest.

3 Kernel Discriminant Analysis with Covariance Operators

In this section, we carry out a population-level discriminant analysis with two types of polynomial kernels and Gaussian kernel and derive an explicit form of population discriminant functions. Section 3.1 covers the case with polynomial kernels in $\mathbb{R}^p$. Section 3.2 extends it to the Gaussian kernel using two types of kernel representations.

3.1 Polynomial Discriminant

Starting with $\mathcal{X} = \mathbb{R}^2$, we lay out steps necessary for a population version of discriminant analysis with homogeneous polynomial kernel and derive the population kernel discriminant function in Section 3.1.1. We then extend the results to a multi-dimensional setting with homogeneous polynomial kernel in Section 3.1.2 and inhomogeneous polynomial in Section 3.1.3.

3.1.1 Homogeneous Polynomial Kernel in Two-Dimensional Setting

The homogeneous polynomial kernel of degree d in $\mathbb{R}^2$ is

$$K_d(\mathbf{x}, \mathbf{u}) = (x_1u_1 + x_2u_2)^d = \sum_{i=0}^{d} \binom{d}{i} (x_1u_1)^{d-i}(x_2u_2)^i = \sum_{i=0}^{d} \binom{d}{i} \left(x_1^{d-i}x_2^i\right)\left(u_1^{d-i}u_2^i\right). \quad (9)$$

The simple form of K_d allows us to obtain the between-class variation function $B_K(\mathbf{x}, \mathbf{u})$ in (5) and within-class variation function $W_K(\mathbf{x}, \mathbf{u})$ in (6) explicitly in terms of the population parameters.

For $B_K(\mathbf{x}, \mathbf{u})$, we begin with

$$\begin{aligned}
& \mathbb{E}_1[K_d(\mathbf{x}, \mathbf{X})] - \mathbb{E}_2[K_d(\mathbf{x}, \mathbf{X})] \\
= \; & \mathbb{E}_1\left[\sum_{i=0}^{d} \binom{d}{i} \left(x_1^{d-i}x_2^i\right)\left(X_1^{d-i}X_2^i\right)\right] - \mathbb{E}_2\left[\sum_{i=0}^{d} \binom{d}{i} \left(x_1^{d-i}x_2^i\right)\left(X_1^{d-i}X_2^i\right)\right] \\
= \; & \sum_{i=0}^{d} \binom{d}{i} \left(x_1^{d-i}x_2^i\right)\left(\mathbb{E}_1[X_1^{d-i}X_2^i] - \mathbb{E}_2[X_1^{d-i}X_2^i]\right),
\end{aligned}$$

which depends on the difference in the moments of total degree d between two classes. Letting $\Delta_i = \mathbb{E}_1[X_1^{d-i}X_2^i] - \mathbb{E}_2[X_1^{d-i}X_2^i]$, the difference in moments, we can express $B_K(\mathbf{x}, \mathbf{u})$ as

$$\begin{aligned}
B_K(\mathbf{x}, \mathbf{u}) &= \left\{\mathbb{E}_1[K_d(\mathbf{x}, \mathbf{X})] - \mathbb{E}_2[K_d(\mathbf{x}, \mathbf{X})]\right\}\left\{\mathbb{E}_1[K_d(\mathbf{u}, \mathbf{X})] - \mathbb{E}_2[K_d(\mathbf{u}, \mathbf{X})]\right\} \\
&= \left\{\sum_{i=0}^{d} \binom{d}{i} \left(x_1^{d-i}x_2^i\right)\Delta_i\right\}\left\{\sum_{j=0}^{d} \binom{d}{j} \left(u_1^{d-j}u_2^j\right)\Delta_j\right\} \\
&= \sum_{i=0}^{d}\sum_{j=0}^{d} \binom{d}{i}\binom{d}{j}\Delta_i\Delta_j \left(x_1^{d-i}x_2^i\right)\left(u_1^{d-j}u_2^j\right).
\end{aligned}$$

Similarly, for $W_K(\mathbf{x}, \mathbf{u})$, using the form of K_d, we first derive the covariance for each class ($l = 1, 2$)

$$\begin{aligned}
\mathrm{Cov}_l[K_d(\mathbf{x}, \mathbf{X}), K_d(\mathbf{u}, \mathbf{X})] &= \mathrm{Cov}_l\left[\sum_{i=0}^{d} \binom{d}{i} \left(x_1^{d-i}x_2^i\right)\left(X_1^{d-i}X_2^i\right), \sum_{j=0}^{d} \binom{d}{j} \left(u_1^{d-j}u_2^j\right)\left(X_1^{d-j}X_2^j\right)\right] \\
&= \sum_{i=0}^{d}\sum_{j=0}^{d} \binom{d}{i}\binom{d}{j} \left(x_1^{d-i}x_2^i\right)\left(u_1^{d-j}u_2^j\right) \mathrm{Cov}_l[X_1^{d-i}X_2^i, X_1^{d-j}X_2^j].
\end{aligned}$$

Letting $W_{i,j} = \pi_1 \text{Cov}_1[X_1^{d-i}X_2^i, X_1^{d-j}X_2^j] + \pi_2 \text{Cov}_2[X_1^{d-i}X_2^i, X_1^{d-j}X_2^j]$, the within-class covariance of a pair of polynomial features of degree d, we can express the within-class variation function as

$$\begin{aligned} W_K(\mathbf{x}, \mathbf{u}) &= \pi_1 \text{Cov}_1[K_d(\mathbf{x}, \mathbf{X}), K_d(\mathbf{u}, \mathbf{X})] + \pi_2 \text{Cov}_2[K_d(\mathbf{x}, \mathbf{X}), K_d(\mathbf{u}, \mathbf{X})] \\ &= \sum_{i=0}^{d}\sum_{j=0}^{d} \binom{d}{i}\binom{d}{j} W_{i,j} \left(x_1^{d-i}x_2^i\right)\left(u_1^{d-j}u_2^j\right). \end{aligned}$$

Using these two functions for K_d, we obtain the between-class covariance operator as

$$\begin{aligned} &\int_{\mathbb{R}^2} B_K(\mathbf{x}, \mathbf{u})\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}) \\ &\quad= \int_{\mathbb{R}^2} \sum_{i=0}^{d}\sum_{j=0}^{d} \binom{d}{i}\binom{d}{j} \Delta_i\Delta_j \left(x_1^{d-i}x_2^i\right)\left(u_1^{d-j}u_2^j\right)\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}) \\ &\quad= \sum_{i=0}^{d}\sum_{j=0}^{d} \binom{d}{i}\binom{d}{j} \Delta_i\Delta_j \left(x_1^{d-i}x_2^i\right)\int_{\mathbb{R}^2}\left(u_1^{d-j}u_2^j\right)\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}) \end{aligned}$$

and the within-class covariance operator as

$$\begin{aligned} &\int_{\mathbb{R}^2} W_K(\mathbf{x}, \mathbf{u})\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}) \\ &\quad= \int_{\mathbb{R}^2} \sum_{i=0}^{d}\sum_{j=0}^{d} \binom{d}{i}\binom{d}{j} W_{i,j} \left(x_1^{d-i}x_2^i\right)\left(u_1^{d-j}u_2^j\right)\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}) \\ &\quad= \sum_{i=0}^{d}\sum_{j=0}^{d} \binom{d}{i}\binom{d}{j} W_{i,j} \left(x_1^{d-i}x_2^i\right)\int_{\mathbb{R}^2}\left(u_1^{d-j}u_2^j\right)\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}). \end{aligned}$$

Given $\alpha(\mathbf{u})$, $\int_{\mathbb{R}^2}\left(u_1^{d-j}u_2^j\right)\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u})$ is a constant. Thus, letting $\nu_j = \binom{d}{j}\int_{\mathbb{R}^2}\left(u_1^{d-j}u_2^j\right)\alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u})$, we arrive at the following eigenvalue problem from (7) for identification of $\alpha(\cdot)$:

$$\sum_{i=0}^{d}\sum_{j=0}^{d} \binom{d}{i} \Delta_i\Delta_j\nu_j \left(x_1^{d-i}x_2^i\right) = \lambda \sum_{i=0}^{d}\sum_{j=0}^{d} \binom{d}{i} W_{i,j}\nu_j \left(x_1^{d-i}x_2^i\right),$$

which should hold for all $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$. Rearranging the terms in the polynomial equation, we have

$$\sum_{i=0}^{d}\left\{\binom{d}{i}\Delta_i \sum_{j=0}^{d}\Delta_j\nu_j\right\}\left(x_1^{d-i}x_2^i\right) = \lambda\sum_{i=0}^{d}\left\{\binom{d}{i}\sum_{j=0}^{d} W_{i,j}\nu_j\right\}\left(x_1^{d-i}x_2^i\right).$$

Matching the coefficients of $x_1^{d-i}x_2^i$ on both sides of the equation leads to the following system of linear equations for $\boldsymbol{\nu} = (\nu_0, \ldots, \nu_d)^T$:

$$\boldsymbol{\Delta}\boldsymbol{\Delta}^T\boldsymbol{\nu} = \lambda W\boldsymbol{\nu}, \tag{10}$$

where $\boldsymbol{\Delta} = (\Delta_0, \Delta_1, \ldots, \Delta_d)^T$ is a vector of the mean differences of $X_1^{d-i}X_2^i$ for $i = 0, \ldots, d$, and $W = [W_{i,j}]$ is a weighted average of their covariance matrices.

When $d = 1$, K_d becomes a linear kernel, and the features are just X_1 and X_2. Thus, $\boldsymbol{\Delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ (population mean difference) and $W = \pi_1\Sigma_1 + \pi_2\Sigma_2$ (pooled population covariance matrix). Clearly, the eigenvalue problem in (10) reduces to that for the population version of Fisher's LDA when $d = 1$.

Assuming that W^{-1} exists, we can show that the largest eigenvalue satisfying equation (10) is $\lambda^* = \boldsymbol{\Delta}^T W^{-1} \boldsymbol{\Delta}$ with eigenvector of $\boldsymbol{\nu}^* = W^{-1}\boldsymbol{\Delta}$. Given the best direction $\boldsymbol{\nu}^* = (\nu_0, \ldots, \nu_d)^T$, the population discriminant function $f(\cdot)$ in (8) with homogenous polynomial kernel of degree d is

$$\begin{aligned} f(\mathbf{x}) &= \int_{\mathbb{R}^2} K_d(\mathbf{x}, \mathbf{u})\, \alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}) = \int_{\mathbb{R}^2} \sum_{j=0}^{d} \binom{d}{j} x_1^{d-j} x_2^j u_1^{d-j} u_2^j\, \alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u}) \\ &= \sum_{j=0}^{d} x_1^{d-j} x_2^j \underbrace{\binom{d}{j} \int_{\mathbb{R}^2} u_1^{d-j} u_2^j\, \alpha(\mathbf{u})\, d\mathbb{P}(\mathbf{u})}_{\nu_j} = \sum_{j=0}^{d} \nu_j x_1^{d-j} x_2^j. \end{aligned}$$

We see that this polynomial discriminant is expressed as a linear combination of the corresponding polynomial features and their coefficients are determined through the mean differences and variances of the features.

3.1.2 Homogeneous Polynomial Kernel in Multi-Dimensional Setting

We extend the result in $\mathcal{X} = \mathbb{R}^2$ to general $\mathbb{R}^p$. The homogeneous polynomial kernel of degree d in $\mathbb{R}^p$ is given as

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

As a function of $\mathbf{x}$, it involves polynomials in p variables of total degree d. To facilitate similar derivations as in $\mathbb{R}^2$, we will use a multi-index for the polynomial features.

Let $\mathbf{j}_d$ denote a p-tuple multi-index with non-negative integer entries that sum up to d. That is, $\mathbf{j}_d \in \mathbf{S}_d := \{(j_1, \ldots, j_p) \mid j_i \in \mathbb{N} \cup \{0\}, \sum_{i=1}^{p} j_i = d\}$ with cardinality of $\binom{d+p-1}{d}$. We will omit the subscript d from $\mathbf{j}_d$ for brevity whenever it is clear from the context. For $\mathbf{j} = (j_1, \ldots, j_p) \in \mathbf{S}_d$, we abbreviate the multinomial coefficient $\binom{d}{j_1, \ldots, j_p}$ to $\binom{d}{\mathbf{j}}$, and let $|\mathbf{j}| = j_1 + \cdots + j_p$ and $\mathbf{j}! = \prod_{k=1}^{p} j_k!$. For $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{j} \in \mathbf{S}_d$, let $\mathbf{x}^{\mathbf{j}} = x_1^{j_1} \cdots x_p^{j_p}$, and for $a \in \mathbb{R}$, $a^{\mathbf{j}}$ means $a^{j_1} \cdots a^{j_p} = a^{|\mathbf{j}|}$. For convenience, we will use $\mathbf{j} \in \mathbf{S}_d$ and $|\mathbf{j}| = d$ interchangeably.

Using this multi-index, we rewrite the homogeneous polynomial kernel in $\mathbb{R}^p$ simply as

$$K_d(\mathbf{x}, \mathbf{u}) = \sum_{|\mathbf{j}|=d} \binom{d}{\mathbf{j}} \mathbf{x}^{\mathbf{j}} \mathbf{u}^{\mathbf{j}}, \tag{11}$$

which can be viewed as a multi-dimensional extension of the expression in (9). Further, we can derive the between-class and within-class variation functions similarly:

$$\begin{aligned} B_K(\mathbf{x}, \mathbf{u}) &= \sum_{|\mathbf{i}|=d} \sum_{|\mathbf{j}|=d} \binom{d}{\mathbf{i}} \binom{d}{\mathbf{j}} \Delta_{\mathbf{i}} \Delta_{\mathbf{j}} \mathbf{x}^{\mathbf{i}} \mathbf{u}^{\mathbf{j}} \\ W_K(\mathbf{x}, \mathbf{u}) &= \sum_{|\mathbf{i}|=d} \sum_{|\mathbf{j}|=d} \binom{d}{\mathbf{i}} \binom{d}{\mathbf{j}} W_{\mathbf{i},\mathbf{j}} \mathbf{x}^{\mathbf{i}} \mathbf{u}^{\mathbf{j}} \end{aligned}$$

with $\Delta_{\mathbf{i}} = \mathbb{E}_1[\mathbf{X}^{\mathbf{i}}] - \mathbb{E}_2[\mathbf{X}^{\mathbf{i}}]$ and $W_{\mathbf{i},\mathbf{j}} = \pi_1 \mathrm{Cov}_1[\mathbf{X}^{\mathbf{i}}, \mathbf{X}^{\mathbf{j}}] + \pi_2 \mathrm{Cov}_2[\mathbf{X}^{\mathbf{i}}, \mathbf{X}^{\mathbf{j}}]$ for $\mathbf{i}, \mathbf{j} \in \mathbf{S}_d$. As an example, when the degree d is 2 in $\mathbb{R}^3$, $\mathbf{S}_2 = \{(2,0,0), (1,1,0), (1,0,1), (0,2,0), (0,1,1), (0,0,2)\}$. For $\mathbf{i} = (1,1,0)$ and $\mathbf{j} = (0,1,1)$, $\mathbf{X}^{\mathbf{i}} = X_1 X_2$ and $\mathbf{X}^{\mathbf{j}} = X_2 X_3$, and thus we have $\Delta_{\mathbf{i}} = \mathbb{E}_1[X_1 X_2] - \mathbb{E}_2[X_1 X_2]$ and $W_{\mathbf{i},\mathbf{j}} = \pi_1 \mathrm{Cov}_1[X_1 X_2, X_2 X_3] + \pi_2 \mathrm{Cov}_2[X_1 X_2, X_2 X_3]$. Due to the same structure, we can easily extend the between-class and within-class covariance operators.

To identify the population discriminant function in this setting, we define $\boldsymbol{\Delta} = (\Delta_{\mathbf{i}})_{\mathbf{i}\in\mathbf{S}_d}^T$, and $\mathbf{W} = [W_{\mathbf{i},\mathbf{j}}]_{\mathbf{i},\,\mathbf{j}\in\mathbf{S}_d}$ analogously. Letting $\nu_{\mathbf{j}} = \binom{d}{\mathbf{j}} \int_{\mathbb{R}^p} \mathbf{u}^{\mathbf{j}} \alpha(\mathbf{u})\, \mathbb{P}(\mathbf{u})$ given a kernel coefficient function $\alpha(\cdot)$, we solve the generalized eigenvalue problem in (10) for $\boldsymbol{\nu} = (\nu_{\mathbf{j}})_{\mathbf{j}\in\mathbf{S}_d}^T$, and determine the population-level discriminant function as

$$f(\mathbf{x}) = \sum_{|\mathbf{j}|=d} \nu_{\mathbf{j}} \mathbf{x}^{\mathbf{j}}.$$

Note that the size of $\boldsymbol{\Delta}$ and $\mathbf{W}$ is $|\mathbf{S}_d| = \binom{d+p-1}{d}$, and while ordering of the indices in $\mathbf{S}_d$ does not matter, the elements in $\boldsymbol{\Delta}$ and $\mathbf{W}$ should be consistently indexed for specification of the eigenvalue problem. The following theorem summarizes the results so far.

Theorem 3.1. *Suppose that for each class, the distribution of $\mathbf{X} \in \mathbb{R}^p$ has finite moments, $\mathbb{E}_l[\mathbf{X}^{\mathbf{i}}]$ and $Cov_l[\mathbf{X}^{\mathbf{i}}, \mathbf{X}^{\mathbf{j}}]$ for all $\mathbf{i}, \mathbf{j} \in \mathbf{S}_d$. For the homogeneous polynomial kernel of degree d, $K_d(\mathbf{x}, \mathbf{u}) = (\mathbf{x}^T\mathbf{u})^d$,*

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

$$f_d(\mathbf{x}) = \sum_{|\mathbf{j}|=d} \nu_{\mathbf{j}} \mathbf{x}^{\mathbf{j}}. \tag{12}$$

(ii) The coefficients, $\boldsymbol{\nu} = (\nu_{\mathbf{i}})_{\mathbf{i}\in\mathbf{S}_d}^T$, for the discriminant function satisfy the eigen-equation with $\lambda > 0$:

$$\boldsymbol{\Delta}\boldsymbol{\Delta}^T\boldsymbol{\nu} = \lambda \mathbf{W}\boldsymbol{\nu}, \tag{13}$$

where $\boldsymbol{\Delta} = (\Delta_{\mathbf{i}})_{\mathbf{i}\in\mathbf{S}_d}^T$ is a vector of moment differences, $\Delta_{\mathbf{i}} = \mathbb{E}_1[\mathbf{X}^{\mathbf{i}}] - \mathbb{E}_2[\mathbf{X}^{\mathbf{i}}]$, and $\mathbf{W} = [W_{\mathbf{i},\mathbf{j}}]_{\mathbf{i},\,\mathbf{j}\in\mathbf{S}_d}$ is a matrix of pooled covariances, $W_{\mathbf{i},\mathbf{j}} = \pi_1 Cov_1[\mathbf{X}^{\mathbf{i}}, \mathbf{X}^{\mathbf{j}}] + \pi_2 Cov_2[\mathbf{X}^{\mathbf{i}}, \mathbf{X}^{\mathbf{j}}]$.

Alternatively, the discriminant function can be derived using an explicit feature map for the kernel. The expression of K_d in (11) suggests $\phi(\mathbf{x}) = \left(\binom{d}{\mathbf{j}}^{\frac{1}{2}} \mathbf{x}^{\mathbf{j}} \right)_{\mathbf{j}\in\mathbf{S}_d}^T$ as a feature vector, and it can be shown that a direct application of LDA to the between-class and within-class variance matrices of $\phi(\mathbf{X})$ leads to the same kernel discriminant. This result indicates that employing homogeneous polynomial kernels in discriminant analysis has the same effect as using the polynomial features of given degree in LDA.

3.1.3 Inhomogeneous Polynomial Kernel

The inhomogeneous polynomial kernel of degree d in $\mathbb{R}^p$ can be expanded as

$$\tilde{K}_d(\mathbf{x}, \mathbf{u}) = (1 + \mathbf{x}^T\mathbf{u})^d = \sum_{m=0}^{d} \binom{d}{m} (\mathbf{x}^T\mathbf{u})^m = \sum_{m=0}^{d} \binom{d}{m} \sum_{|\mathbf{j}|=m} \binom{m}{\mathbf{j}} \mathbf{x}^{\mathbf{j}}\mathbf{u}^{\mathbf{j}},$$

which is a sum of all homogeneous polynomial kernels of degree up to d. Since $\binom{d}{\mathbf{j}} = \binom{d}{m}\binom{m}{\mathbf{j}}$ for $\mathbf{j} \in \mathbf{S}_m$, $m = 0, \ldots, d$, and the term with $m = 0$ is 1, we can rewrite the kernel as

$$\tilde{K}_d(\mathbf{x}, \mathbf{u}) = 1 + \sum_{m=1}^{d} \sum_{|\mathbf{j}|=m} \binom{d}{\mathbf{j}} \mathbf{x}^{\mathbf{j}}\mathbf{u}^{\mathbf{j}} = 1 + \sum_{|\mathbf{j}|=1}^{d} \binom{d}{\mathbf{j}} \mathbf{x}^{\mathbf{j}}\mathbf{u}^{\mathbf{j}}.$$

Note that $\sum_{m=1}^{d} \sum_{|\mathbf{j}|=m}$ is abbreviated to $\sum_{|\mathbf{j}|=1}^{d}$. This kernel has the same structure as the homogenous polynomial kernel. Using the relation, we can find the population kernel discriminant function similarly. Recognizing that $\tilde{K}_d$ involves expanded polynomial features in p variables of total degree 0 to d:

$1, \mathbf{x}, (\mathbf{x}^{\mathbf{j}})_{|\mathbf{j}|=2}, \ldots, (\mathbf{x}^{\mathbf{j}})_{|\mathbf{j}|=d}$, we define a vector of the mean differences of those features (excluding the constant 1) and a block matrix of their pooled covariances as follows:

$$\tilde{\boldsymbol{\Delta}} = \begin{pmatrix} \boldsymbol{\Delta}_1 \\ \vdots \\ \boldsymbol{\Delta}_d \end{pmatrix}, \quad \text{and} \quad \tilde{\mathbf{W}} = \begin{bmatrix} \mathbf{W}_{1,1} & \cdots & \mathbf{W}_{1,d} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{d,1} & \cdots & \mathbf{W}_{d,d} \end{bmatrix},$$

where $\boldsymbol{\Delta}_m = (\Delta_{\mathbf{i}})_{\mathbf{i}\in\mathbf{S}_m}^T$ and $\mathbf{W}_{m,l} = [W_{\mathbf{i},\mathbf{j}}]_{\mathbf{i}\in\mathbf{S}_m,\mathbf{j}\in\mathbf{S}_l}$ for all $m,\ l = 1, \ldots, d$. That is, $\tilde{\boldsymbol{\Delta}}$ contains all the difference of the moments of degree 1 to d, and $\tilde{\mathbf{W}}$ has the covariances between all the monomials of degree 1 to d. Thus, the size of the eigenvalue problem to solve becomes $\sum_{m=1}^{d} \binom{m+p-1}{m} = \binom{p+d}{d} - 1$. The following theorem states similar results for the discriminant function with inhomogeneous polynomial kernel.

Theorem 3.2. *Suppose that for each class, the distribution of $\mathbf{X} \in \mathbb{R}^p$ has finite moments, $\mathbb{E}_l\left[\mathbf{X}^{\mathbf{i}}\right]$ and $Cov_l\left[\mathbf{X}^{\mathbf{i}}, \mathbf{X}^{\mathbf{j}}\right]$ for all $\mathbf{i} \in \mathbf{S}_m$, $\mathbf{j} \in \mathbf{S}_l$ and $m, l = 1, \ldots, d$. For the inhomogeneous polynomial kernel of degree d, $\tilde{K}_d(\mathbf{x}, \mathbf{u}) = (1 + \mathbf{x}^T\mathbf{u})^d$,*

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

$$\tilde{f}_d(\mathbf{x}) = \sum_{|\mathbf{j}|=1}^{d} \tilde{\nu}_{\mathbf{j}} \mathbf{x}^{\mathbf{j}}. \tag{14}$$

(ii) The coefficients, $\tilde{\boldsymbol{\nu}} = (\tilde{\nu}_{\mathbf{j}})_{1\le|\mathbf{j}|\le d}^T$, for the discriminant function satisfy the eigen-equation with $\lambda > 0$:

$$\tilde{\boldsymbol{\Delta}}\tilde{\boldsymbol{\Delta}}^T\tilde{\boldsymbol{\nu}} = \lambda\tilde{\mathbf{W}}\tilde{\boldsymbol{\nu}}. \tag{15}$$

3.2 Gaussian Discriminant

We extend the discriminant analysis with polynomial kernels in the previous section to the Gaussian kernel. For the extension, we use two representations for the Gaussian kernel: a deterministic representation in Section 3.2.1 and a randomized feature representation in Section 3.2.2.

3.2.1 Deterministic Representation of Gaussian Kernel

We have seen so far that derivation of the population discriminant function with polynomial kernels is aided by their expansion, or equivalently, their explicit feature maps. Taking a similar approach to the Gaussian kernel, we could use the Maclaurin series of e^x to express it as

$$K_\omega(\mathbf{x}, \mathbf{u}) \quad = \quad \exp\left(-\frac{\|\mathbf{x}-\mathbf{u}\|^2}{2\omega^2}\right) = \sum_{|\mathbf{j}|=0}^{\infty} \phi_{\mathbf{j}}(\mathbf{x})\phi_{\mathbf{j}}(\mathbf{u}),$$

with $\phi_{\mathbf{j}}(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}\|^2}{2\omega^2}\right)\frac{1}{\sqrt{\mathbf{j}!}}\frac{\mathbf{x}^{\mathbf{j}}}{\omega^{\mathbf{j}}}$. While the structure of K_ω in this representation would permit similar derivations as before for the discriminant function, the result will depend on the expectations and covariances of $\phi_{\mathbf{j}}(\mathbf{X})$ which may not be easy to obtain analytically in general.

Alternatively, we consider a representation of the kernel in the form that allows a direct use of polynomial features in much the same way as polynomial kernels. We start with a one-dimensional case and then extend it to a multi-dimensional case. The one-dimensional Gaussian kernel with bandwidth ω can be written as

$$K_\omega(x, u) = \exp\left(-\frac{(x-u)^2}{2\omega^2}\right) = \exp\left(-\frac{x^2}{2\omega^2}\right)\sum_{m=0}^{\infty} He_m\left(\frac{x}{\omega}\right)\frac{u^m}{m!\,\omega^m}. \tag{16}$$

$He_m(x)$ are referred to as the probabilist's Hermite polynomials and defined as

$$He_m(x) \quad = \quad (-1)^m(\phi(x))^{-1}\frac{d^m}{dx^m}\phi(x),$$

where ϕ is the density function of the standard normal distribution. The representation of K_ω in (16) comes from the Hermite polynomial generating function:

$$\exp\left(xu - \frac{1}{2}u^2\right) = \sum_{m=0}^{\infty} He_m(x)\frac{u^m}{m!}. \tag{17}$$

It can be extended to a multivariate case using the vector-valued Hermite polynomials introduced in Holmquist (1996).

For $\mathbf{x} \in \mathbb{R}^p$ and $m \in \mathbb{N}$, the p-variate vector-valued Hermite polynomial of order m is defined as

$$\mathbf{H}_m(\mathbf{x}) = (-1)^m(\Phi(\mathbf{x}))^{-1}\partial_{\mathbf{x}}^{\langle m\rangle}\Phi(\mathbf{x}),$$

where $\partial_{\mathbf{x}}^{\langle m\rangle} = \partial_{\mathbf{x}} \otimes \partial_{\mathbf{x}} \otimes \cdots \otimes \partial_{\mathbf{x}}$ (m-times) is a Kronecker product of the differential operator $\partial_{\mathbf{x}} = (\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_p})^T$ and Φ is the product of p univariate standard normal densities. Thus the components of $\mathbf{H}_m(\mathbf{x})$ are a product of univariate Hermite polynomials whose total degree is m: $\mathbf{H}_m(\mathbf{x}) = (H_{\mathbf{j}}(\mathbf{x}))_{\mathbf{j}\in\mathbf{S}_m}^T$, where $H_{\mathbf{j}}(\mathbf{x}) = He_{j_1}(x_1)\cdots He_{j_p}(x_p)$ for each $\mathbf{j} \in \mathbf{S}_m$.

Using this notation, a multivariate version of the generating function (17) can be written as

$$\exp\left(\mathbf{x}^T\mathbf{u} - \frac{1}{2}\mathbf{u}^T\mathbf{u}\right) = \sum_{m=0}^{\infty}\frac{1}{m!}\langle\mathbf{H}_m(\mathbf{x}), \mathbf{u}^{\langle m\rangle}\rangle,$$

where $\mathbf{u}^{\langle m\rangle} = \mathbf{u} \otimes \mathbf{u} \otimes \cdots \otimes \mathbf{u}$ (m-times) and

$$\langle\mathbf{H}_m(\mathbf{x}), \mathbf{u}^{\langle m\rangle}\rangle = \sum_{\mathbf{j}\in\mathbf{S}_m}\binom{m}{\mathbf{j}}He_{j_1}(x_1)\cdots He_{j_p}(x_p)u_1^{j_1}\cdots u_p^{j_p} = \sum_{\mathbf{j}\in\mathbf{S}_m}\binom{m}{\mathbf{j}}H_{\mathbf{j}}(\mathbf{x})\mathbf{u}^{\mathbf{j}}.$$

Using the generating function for $\mathbf{H}_m$ and letting $\mathbf{x}_\omega = (1/\omega)\mathbf{x}$ with bandwidth ω, we get the following expansion for the multivariate Gaussian kernel:

$$\begin{aligned} K_\omega(\mathbf{x}, \mathbf{u}) &= \exp\left(-\frac{\|\mathbf{x}-\mathbf{u}\|^2}{2\omega^2}\right) = \exp\left(-\frac{\|\mathbf{x}_\omega\|^2}{2}\right)\exp\left(\mathbf{x}_\omega^T\mathbf{u}_\omega - \frac{1}{2}\mathbf{u}_\omega^T\mathbf{u}_\omega\right) \\ &= \exp\left(-\frac{\|\mathbf{x}_\omega\|^2}{2}\right)\sum_{|\mathbf{j}|=0}^{\infty}\frac{1}{\mathbf{j}!}H_{\mathbf{j}}(\mathbf{x}_\omega)\mathbf{u}_\omega^{\mathbf{j}}. \end{aligned}$$

Further with the definition of $\tilde{H}_{\mathbf{j}}(\mathbf{x}_\omega) = \frac{1}{\mathbf{j}!\omega^{\mathbf{j}}}\exp\left(-\frac{\|\mathbf{x}_\omega\|^2}{2}\right)H_{\mathbf{j}}(\mathbf{x}_\omega)$, the kernel is represented as

$$K_\omega(\mathbf{x}, \mathbf{u}) = \sum_{|\mathbf{j}|=0}^{\infty}\tilde{H}_{\mathbf{j}}(\mathbf{x}_\omega)\mathbf{u}^{\mathbf{j}}. \tag{18}$$

Although this representation is asymmetric in $\mathbf{x}$ and $\mathbf{u}$, it facilitates similar derivations of the generalized eigenvalue problem and population kernel discriminant as with polynomial kernels, but using the entirety of polynomial features.

With this representation, it is easy to show that

$$\begin{aligned}\mathbb{E}_1\left[K_\omega(\mathbf{x},\mathbf{X})\right]-\mathbb{E}_2[K_\omega(\mathbf{x},\mathbf{X})] &= \sum_{|\mathbf{j}|=0}^{\infty}\tilde{H}_{\mathbf{j}}(\mathbf{x}_\omega)\left\{\mathbb{E}_1[\mathbf{X}^{\mathbf{j}}]-\mathbb{E}_2[\mathbf{X}^{\mathbf{j}}]\right\}\\ &= \sum_{|\mathbf{j}|=0}^{\infty}\tilde{H}_{\mathbf{j}}(\mathbf{x}_\omega)\Delta_{\mathbf{j}}=\sum_{|\mathbf{j}|=1}^{\infty}\tilde{H}_{\mathbf{j}}(\mathbf{x}_\omega)\Delta_{\mathbf{j}},\end{aligned}$$

which involves the moments of the distribution rather than the expectations of $\tilde{H}_{\mathbf{j}}(\mathbf{X}_\omega)$. Note that the last equality is due to $\Delta_{\mathbf{0}}=0$ for $\mathbf{X}^{\mathbf{0}}=1$. Thus the between-class variation function is given as

$$B_K(\mathbf{x},\mathbf{u}) = \sum_{|\mathbf{i}|=1}^{\infty}\sum_{|\mathbf{j}|=1}^{\infty}\Delta_{\mathbf{i}}\Delta_{\mathbf{j}}\tilde{H}_{\mathbf{i}}(\mathbf{x}_\omega)\tilde{H}_{\mathbf{j}}(\mathbf{u}_\omega).$$

Similarly the within-class variation function is given as

$$W_K(\mathbf{x},\mathbf{u}) = \sum_{|\mathbf{i}|=1}^{\infty}\sum_{|\mathbf{j}|=1}^{\infty}W_{\mathbf{i},\mathbf{j}}\tilde{H}_{\mathbf{i}}(\mathbf{x}_\omega)\tilde{H}_{\mathbf{j}}(\mathbf{u}_\omega).$$

Therefore, the eigenvalue problem in (7) with the Gaussian kernel is given by

$$\sum_{|\mathbf{i}|=1}^{\infty}\sum_{|\mathbf{j}|=1}^{\infty}\Delta_{\mathbf{i}}\Delta_{\mathbf{j}}\nu_{\mathbf{j}}\tilde{H}_{\mathbf{i}}(\mathbf{x}_\omega)=\lambda\sum_{|\mathbf{i}|=1}^{\infty}\sum_{|\mathbf{j}|=1}^{\infty}\mathbf{W}_{\mathbf{i},\mathbf{j}}\nu_{\mathbf{j}}\tilde{H}_{\mathbf{i}}(\mathbf{x}_\omega), \tag{19}$$

where $\nu_{\mathbf{j}}=\int_{\mathcal{X}}\tilde{H}_{\mathbf{j}}(\mathbf{u}_\omega)\,\alpha(\mathbf{u})\,d\mathbb{P}(\mathbf{u})$.

To find $\nu_{\mathbf{j}}$ satisfying (19) for every $\mathbf{x}_\omega$, the coefficients of $\tilde{H}_{\mathbf{i}}(\mathbf{x}_\omega)$ on both sides must equal for all $\mathbf{i}\in\mathbf{S}_m$, $m\in\mathbb{N}$. This entails the following system of an infinite number of linear equations for $\nu_{\mathbf{j}}$:

$$\Delta_{\mathbf{i}}\sum_{|\mathbf{j}|=1}^{\infty}\Delta_{\mathbf{j}}\nu_{\mathbf{j}}=\lambda\sum_{|\mathbf{j}|=1}^{\infty}W_{\mathbf{i},\mathbf{j}}\nu_{\mathbf{j}},\quad \mathbf{i}\in\mathbf{S}_m,\ m\in\mathbb{N}, \tag{20}$$

and the resulting discriminant function of the form: $f(\mathbf{x})=\sum_{|\mathbf{j}|=1}^{\infty}\nu_{\mathbf{j}}\mathbf{x}^{\mathbf{j}}$.

For a finite dimensional approximation of the population discriminant function, we may consider truncation of the kernel representation in (18) at $|\mathbf{j}|=N$:

$$K_N(\mathbf{x},\mathbf{u}) = \sum_{|\mathbf{j}|=0}^{N}\tilde{H}_{\mathbf{j}}(\mathbf{x}_\omega)\mathbf{u}^{\mathbf{j}}.$$

This approximation brings the corresponding truncation of the system of linear equations for the generalized eigenvalue problem in (20). As a result, the eigenvalue equation coincides with that for the inhomogeneous polynomial kernel of degree N in Theorem 3.2, and so does the truncated discriminant function. As more polynomial features are added or N increases, the largest eigenvalue satisfying equation (15) increases. Adding subscript N to λ, $\tilde{\mathbf{\Delta}}$ and $\tilde{\mathbf{W}}$ to indicate the degree clearly, let $\lambda_N=\max_{\boldsymbol{\nu}}\dfrac{\boldsymbol{\nu}^T\tilde{\mathbf{\Delta}}_N\tilde{\mathbf{\Delta}}_N^T\boldsymbol{\nu}}{\boldsymbol{\nu}^T\tilde{\mathbf{W}}_N\boldsymbol{\nu}}$. The moment difference vector $\tilde{\mathbf{\Delta}}_N$ and the within-class covariance matrix $\tilde{\mathbf{W}}_N$ expand with N, including all the elements up to degree N. This nesting structure produces an increasing sequence of λ_N. It is because maximization of the ratio for degree N amounts to that for degree $N+1$ with a limited space for $\boldsymbol{\nu}$. In Section 4.1, we will study the relation between polynomial and Gaussian discriminants numerically under various scenarios and discuss the effect of N on the quality of the discriminant function.

3.2.2 Fourier Feature Representation of Gaussian kernel

In addition to the polynomial approximation presented in the previous section, a stochastic approximation to the Gaussian kernel can be used for population analysis. Rahimi and Recht (2008a) examined approximation of shift-invariant kernels in general using random Fourier features for fast large-scale optimization with kernels. They proposed the following representation for the Gaussian kernel using random features of the form $z_{\mathbf{w}}(\mathbf{x}) = (\cos(\mathbf{w}^T\mathbf{x}), \sin(\mathbf{w}^T\mathbf{x}))^T$:

$$K_\omega(\mathbf{x}, \mathbf{u}) = \mathbb{E}_{\mathbf{w}}\left[z_{\mathbf{w}}(\mathbf{x})^T z_{\mathbf{w}}(\mathbf{u})\right], \tag{21}$$

where $\mathbf{w}$ is a random vector from a multivariate normal distribution with mean zero and covariance matrix $\frac{1}{\omega^2} I_p$. This representation comes from Bochner's theorem (Rudin 2017), which describes the correspondence between a positive definite shift-invariant kernel and the Fourier transform of a nonnegative measure. The feature map $z_{\mathbf{w}}(\cdot)$ projects $\mathbf{x}$ onto a random direction $\mathbf{w}$ first and then takes sinusoidal transforms. Their frequency depends on the norm of $\mathbf{w}$. A large bandwidth ω for the Gaussian kernel implies realization of $\mathbf{w}$ with a small norm on average, which generally entails a low frequency for the sinusoids.

The representation in (21) suggests a Monte Carlo approximation of the kernel. Suppose that $\mathbf{w}_i$, $i = 1, \ldots, D$ are randomly generated from $N_p(\mathbf{0}, \frac{1}{\omega^2} I_p)$. Defining random Fourier features $z_{\mathbf{w}}(\mathbf{x})$ with $\mathbf{w} = \mathbf{w}_i$, we can approximate the Gaussian kernel using a sample average as follows:

$$K_\omega(\mathbf{x}, \mathbf{u}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{u}\|^2}{2\omega^2}\right) \approx \frac{1}{D}\sum_{i=1}^{D} z_{\mathbf{w}_i}(\mathbf{x})^T z_{\mathbf{w}_i}(\mathbf{u}).$$

This average can be taken as an unbiased estimate of the kernel, and its precision is controlled by D. Concatenating these D random components $z_{\mathbf{w}_i}(\mathbf{x})$, we can also see that the stochastic approximation above amounts to defining

$$\underset{\sim}{\mathbf{Z}}_D(\mathbf{x}) = \frac{1}{\sqrt{D}}(z_{\mathbf{w}_1}(\mathbf{x})^T, \ldots, z_{\mathbf{w}_D}(\mathbf{x})^T)^T$$

as a randomized feature map for the kernel.

Using the random Fourier features, we approximate the between-class variation function $B_K(\mathbf{x}, \mathbf{u})$ and within-class variation function $W_K(\mathbf{x}, \mathbf{u})$ as follows:

$$\begin{aligned} B_K(\mathbf{x}, \mathbf{u}) &\approx \frac{1}{D^2}\sum_{i=1}^{D}\sum_{j=1}^{D} z_{\mathbf{w}_i}(\mathbf{x})^T \Delta_{\mathbf{w}_i} \Delta_{\mathbf{w}_j}^T z_{\mathbf{w}_j}(\mathbf{u}) \\ W_K(\mathbf{x}, \mathbf{u}) &\approx \frac{1}{D^2}\sum_{i=1}^{D}\sum_{j=1}^{D} z_{\mathbf{w}_i}(\mathbf{x})^T W_{\mathbf{w}_i, \mathbf{w}_j} z_{\mathbf{w}_j}(\mathbf{u}), \end{aligned}$$

where $\Delta_{\mathbf{w}_i} = \mathbb{E}_1\left[z_{\mathbf{w}_i}(\mathbf{X})\right] - \mathbb{E}_2\left[z_{\mathbf{w}_i}(\mathbf{X})\right]$ and $W_{\mathbf{w}_i, \mathbf{w}_j} = \pi_1 \mathrm{Cov}_1\left[z_{\mathbf{w}_i}(\mathbf{X}), z_{\mathbf{w}_j}(\mathbf{X})\right] + \pi_2 \mathrm{Cov}_2\left[z_{\mathbf{w}_i}(\mathbf{X}), z_{\mathbf{w}_j}(\mathbf{X})\right]$. Then we can define a randomized version of the eigenvalue problem in (7) with these approximations. Let $\hat{\alpha}(\cdot)$ denote the solution to the problem with $\lambda > 0$ and define $\boldsymbol{\nu}_i = \int z_{\mathbf{w}_i}(\mathbf{u})\hat{\alpha}(\mathbf{u})\, d\mathbb{P}(\mathbf{u})$. Similar arguments as before lead to the following generalized eigenvalue problem to determine $\boldsymbol{\nu} = (\boldsymbol{\nu}_i^T)^T$:

$$\hat{\boldsymbol{\Delta}}\hat{\boldsymbol{\Delta}}^T\boldsymbol{\nu} = \lambda\hat{\mathbf{W}}\boldsymbol{\nu},$$

where $\hat{\boldsymbol{\Delta}} = (\Delta_{\mathbf{w}_i}^T)^T$ and $\hat{\mathbf{W}} = [W_{\mathbf{w}_i, \mathbf{w}_j}]$ for $i, j = 1, \ldots, D$. Given $\boldsymbol{\nu}$, the approximate Gaussian discriminant obtained via random Fourier features is

$$f_D(\mathbf{x}) = \frac{1}{D}\sum_{i=1}^{D} \boldsymbol{\nu}_i^T z_{\mathbf{w}_i}(\mathbf{x}). \tag{22}$$

Rather than sine and cosine pairs, we could also use phase-shifted cosine features only to approximate the Gaussian kernel as suggested in Rahimi and Recht (2008a) and Rahimi and Recht (2008b). Let $z_{\mathbf{w},b}(\mathbf{x}) = \sqrt{2}\cos(\mathbf{w}^T\mathbf{x}+b)$ with an additional phase parameter b which is independent of $\mathbf{w}$ and distributed uniformly on $(0, 2\pi)$. Then using a trigonometric identity, we can verify that

$$K_\omega(\mathbf{x}, \mathbf{u}) = \mathbb{E}_{\mathbf{w},b}\left[z_{\mathbf{w},b}(\mathbf{x})z_{\mathbf{w},b}(\mathbf{u})\right] = \mathbb{E}_{\mathbf{w},b}\left[2\cos(\mathbf{w}^T\mathbf{x} + b)\cos(\mathbf{w}^T\mathbf{u} + b)\right].$$

Given $\mathbf{w}$ and b, if $\mathbf{X}$ is distributed with $N_p(\boldsymbol{\mu}, \Sigma)$, we can show that

$$\mathbb{E}_{\mathbf{X}}\left[\cos(\mathbf{w}^T\mathbf{X} + b)\right] = \exp(-\frac{1}{2}\mathbf{w}^T\Sigma\mathbf{w})\cos(\mathbf{w}^T\boldsymbol{\mu} + b).$$

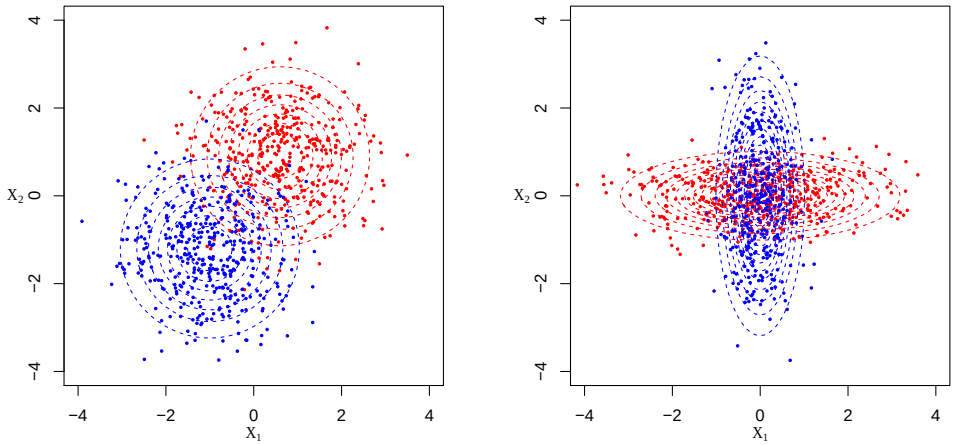
Thus in the classical LDA setting of $\mathbb{P}_j = N(\boldsymbol{\mu}_j, \Sigma)$ for $j = 1, 2$, this Fourier feature lets us focus on the difference in $\cos(\mathbf{w}^T\boldsymbol{\mu}_j + b)$ rather than $\boldsymbol{\mu}_j$.

4 Numerical Studies

This section illustrates the relation between the data distribution and kernel discriminants discussed so far through simulation studies and an application to real data.

4.1 Simulation Study

We numerically study the population discriminant functions in (12), (14), and (22) with both polynomial and Gaussian kernels, and examine their relationship with the underlying data distributions for two classes. For illustration, we consider two scenarios where each class follows a bivariate normal distribution. In Scenario 1, two classes have different means ($\boldsymbol{\mu}_1 = (0.6, 0.9)^T$ and $\boldsymbol{\mu}_2 = (-1.0, -1.2)^T$) but the same covariance ($\Sigma_1 = \Sigma_2 = I_2$), and in Scenario 2, they have the same mean ($\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$) but different covariances ($\Sigma_1 = \text{diag}(2, 0.2)$ and $\Sigma_2 = \text{diag}(0.2, 2)$). Figure 1 shows the scatter plots of samples generated from each scenario with 400 data points in each class (red: class 1 and blue: class 2) under the assumption that two classes are equally likely.



(a) Scenario 1 (b) Scenario 2

Figure 1: Scatterplots of the samples simulated from a mixture of two normal distributions with contours of the probability densities for each class overlaid in two settings: (a) Scenario 1 and (b) Scenario 2.

4.1.1 Polynomial Kernel

Under each scenario, we find the population discriminant functions in (12) and (14) with polynomial kernels of degree 1 to 4 and examine the effect of the degree on the discriminants. To determine $f_d(\mathbf{x})$, we first obtain the population moment differences $\mathbf{\Delta}$ and covariances $\mathbf{W}$ explicitly and solve the eigenvalue problem in (13). Similarly we determine $\tilde{f}_d(\mathbf{x})$ with $\tilde{\mathbf{\Delta}}$ and $\tilde{\mathbf{W}}$. Tables 1 and 2 present the coefficients for the polynomial discriminants $f_d(\mathbf{x})$ and $\tilde{f}_d(\mathbf{x})$ in each scenario, which are the solution $\boldsymbol{\nu}$ or $\tilde{\boldsymbol{\nu}}$ (eigenvector) normalized to unit length.

Table 1: Coefficients for the population polynomial discriminants under Scenario 1.

	Homogeneous polynomial				Inhomogeneous polynomial			
Term	$f_{\mathbf{1}}(\mathbf{x})$	$f_{\mathbf{2}}(\mathbf{x})$	$f_{\mathbf{3}}(\mathbf{x})$	$f_{\mathbf{4}}(\mathbf{x})$	$\tilde{f}_{\mathbf{1}}(\mathbf{x})$	$\tilde{f}_{\mathbf{2}}(\mathbf{x})$	$\tilde{f}_{\mathbf{3}}(\mathbf{x})$	$\tilde{f}_{\mathbf{4}}(\mathbf{x})$
x_1	0.6060	-	-	-	0.6060	0.6060	0.6033	0.6033
x_2	0.7954	-	-	-	0.7954	0.7954	0.7919	0.7919
x_1^2	-	-0.4461	-	-	-	0.0000	-0.0141	-0.0141
x_1x_2	-	-0.8376	-	-	-	0.0000	-0.0369	-0.0369
x_2^2	-	-0.3154	-	-	-	0.0000	-0.0242	-0.0242
x_1^3	-	-	0.6412	-	-	-	-0.0118	-0.0118
$x_1^2x_2$	-	-	0.3105	-	-	-	-0.0465	-0.0465
$x_1x_2^2$	-	-	-0.2277	-	-	-	-0.0610	-0.0610
x_2^3	-	-	0.6637	-	-	-	-0.0267	-0.0267
x_1^4	-	-	-	-0.2575	-	-	-	0.0000
$x_1^3x_2$	-	-	-	-0.6186	-	-	-	0.0000
$x_1^2x_2^2$	-	-	-	0.3860	-	-	-	0.0000
$x_1x_2^3$	-	-	-	-0.6146	-	-	-	0.0000
x_2^4	-	-	-	-0.1563	-	-	-	0.0000

Scenario 1: Fisher's linear discriminant analysis is optimal in this scenario. Since the common covariance matrix is I_2, the linear discriminant is simply determined by the direction of the mean difference, which is $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (1.6, 2.1)^T$. This gives $f^*(\mathbf{x}) = 1.6x_1 + 2.1x_2$ as an optimal linear discriminant defined up to a multiplicative constant. From Table 1, we first notice that the coefficient vector for the population linear discriminant, $f_1(\mathbf{x})$, $\boldsymbol{\nu} = (0.6060, 0.7954)^T$, is a normalized mean difference. Further we observe that the coefficients for the discriminants with inhomogeneous polynomial kernels, $\tilde{f}_{\mathbf{1}}(\mathbf{x})$ and $\tilde{f}_{\mathbf{2}}(\mathbf{x})$, are also proportional to the mean difference.

Figures 2 and 3 display the polynomial discriminants identified in Table 1. The first row of Figure 2 shows contours of the population discriminants with homogenous polynomial kernels. High to low discriminant scores correspond to red to blue contours. The black dashed line is $1.6x_1 + 2.1x_2 = 0.635$, which is the classification boundary from Fisher's linear discriminant analysis. The second row of Figure 2 presents the corresponding sample embeddings obtained by performing a kernel discriminant analysis to the given samples. Figure 3 shows contours of both versions with inhomogeneous polynomial kernels of degree 2 to 4, omitting degree 1 as they are identical to those with the linear kernel in Figure 2.

The population discriminants and sample versions are similar in terms of shape and direction of change in contours. With odd-degree homogeneous polynomial kernels, we observe that the contours change in the direction of the mean difference, indicating that odd degrees are effective in this setting. The even-degree discriminants, however, are of hyperbolic paraboloid shape, varying in a way that masks the class difference completely. By contrast, the degree doesn't affect the major direction of change in the population discriminants with inhomogeneous polynomial kernels. Their variation seems to occur only in the direction

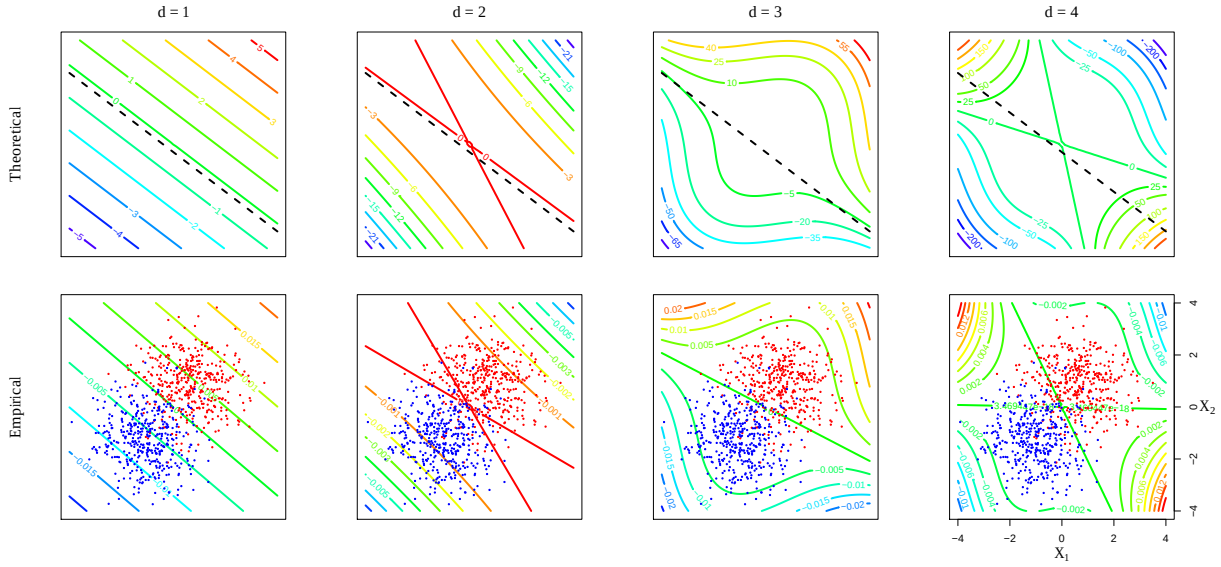


Figure 2: Contours of the population discriminant functions with homogeneous polynomial kernels of degree 1 to 4 (upper panels from left to right) and their corresponding sample counterparts (lower panels) under Scenario 1. The black dashed lines are the optimal classification boundary.

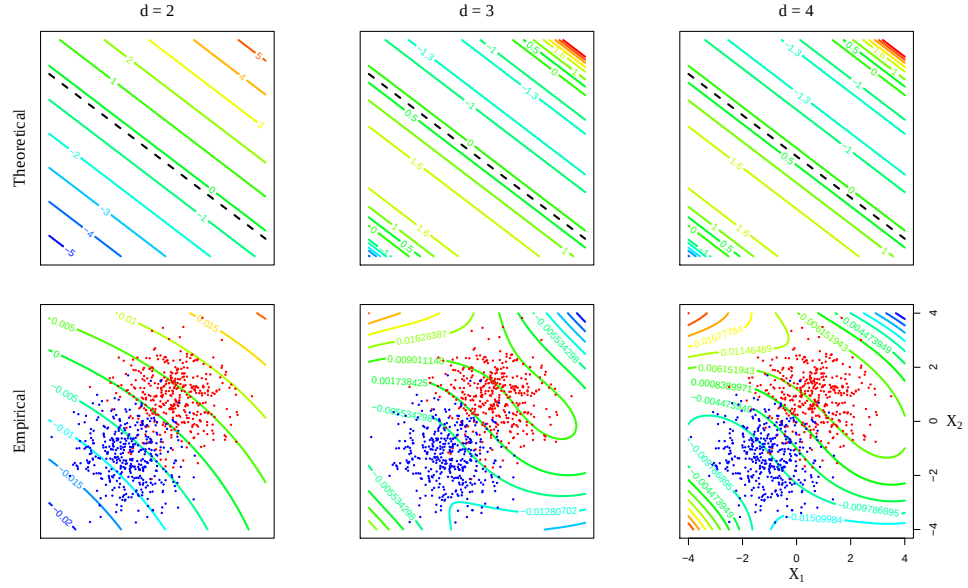


Figure 3: Contours of the population discriminant functions with inhomogeneous polynomial kernels of degree 2 to 4 (upper panels from left to right) and their corresponding sample counterparts (lower panels) under Scenario 1. The black dashed lines are the optimal classification boundary.

of the mean difference. Table 1 confirms that the resulting discriminants $\tilde{f}_d(\mathbf{x})$ are identical for degrees $d = 2k-1$ and $2k$, $k = 1, 2$.

Scenario 2: In this scenario, using the true densities, the optimal decision boundary is found to be $(x_1 + x_2)(x_1 - x_2) = 0$, and the optimal discriminant function is $f^*(\mathbf{x}) = x_1^2 - x_2^2$, which is a homogeneous polynomial of degree 2. In contrast with Scenario 1, even-degree features are discriminative in this setting. Note

Table 2: Coefficients for the population polynomial discriminants under Scenario 2.

	Homogeneous polynomial				Inhomogeneous polynomial			
Term	$f_{\mathbf{1}}(\mathbf{x})$	$f_{\mathbf{2}}(\mathbf{x})$	$f_{\mathbf{3}}(\mathbf{x})$	$f_{\mathbf{4}}(\mathbf{x})$	$\tilde{f}_{\mathbf{1}}(\mathbf{x})$	$\tilde{f}_{\mathbf{2}}(\mathbf{x})$	$\tilde{f}_{\mathbf{3}}(\mathbf{x})$	$\tilde{f}_{\mathbf{4}}(\mathbf{x})$
x_1	0.00	-	-	-	0.00	0.0000	0.0000	0.0000
x_2	0.00	-	-	-	0.00	0.0000	0.0000	0.0000
x_1^2	-	0.7071	-	-	-	0.7071	0.7071	0.7063
x_1x_2	-	0.0000	-	-	-	0.0000	0.0000	0.0000
x_2^2	-	-0.7071	-	-	-	-0.7071	-0.7071	-0.7063
x_1^3	-	-	0.0000	-	-	-	0.0000	0.0000
$x_1^2x_2$	-	-	0.0000	-	-	-	0.0000	0.0000
$x_1x_2^2$	-	-	0.0000	-	-	-	0.0000	0.0000
x_2^3	-	-	0.0000	-	-	-	0.0000	0.0000
x_1^4	-	-	-	0.7071	-	-	-	-0.0335
$x_1^3x_2$	-	-	-	0.0000	-	-	-	0.0000
$x_1^2x_2^2$	-	-	-	0.0000	-	-	-	0.0000
$x_1x_2^3$	-	-	-	0.0000	-	-	-	0.0000
x_2^4	-	-	-	-0.7071	-	-	-	0.0335

that the coefficients of $f_{\mathbf{2}}(\mathbf{x})$, $\tilde{f}_{\mathbf{2}}(\mathbf{x})$ and $\tilde{f}_{\mathbf{3}}(\mathbf{x})$ in Table 2 are proportional to those of $f^*(\mathbf{x})$. Odd-degree homogeneous polynomials produce a degenerate discriminant in this setting. The quadratic discriminant, $f_{\mathbf{2}}(\mathbf{x}) = 0.7071x_1^2 - 0.7071x_2^2$, is a normalized version of $f^*(\mathbf{x})$. With degree 4 homogeneous polynomial kernel, we have $f_{\mathbf{4}}(\mathbf{x}) = 0.7071x_1^4 - 0.7071x_2^4$, which has the optimal discriminant as its factor. Contours of these polynomial discriminants are displayed in the first row of Figure 4. The black dashed lines are the optimal decision boundaries. The second row of Figure 4 presents the corresponding nonlinear kernel embeddings of degree 1 to 4 induced by the samples. Figure 5 shows contours of both versions (theoretical in the first row and empirical in the second row) with inhomogeneous polynomial kernels of degree 2 to 4, omitting the degenerate linear case in Table 2.

Similar to Scenario 1, we observe that the population discriminant functions and their sample counterparts in Figures 4 and 5 exhibit similarity in terms of shape and direction of change in contours. The contours of the population quadratic and quartic discriminants in Figure 4 show symmetry along each variable axis. Quadratic features contain all information necessary for discrimination in this scenario. Even-degree features successfully discriminate the two classes while odd-degree features completely fail as shown in Figure 4. Nonlinear inhomogeneous polynomial kernels with even-degree features enable proper classification as illustrated in Figure 5. Inhomogeneous polynomial kernels of degree $2k+1$ and $2k$ produce identical discriminants in this setting.

4.1.2 Gaussian Kernel

We examine Gaussian discriminant functions under each scenario using two types of approximation to the Gaussian kernel discussed earlier.

Deterministic representation: Truncation of the deterministic representation of the Gaussian kernel at a certain degree leads to the population polynomial discriminant using the inhomogeneous polynomial kernel of the same degree. Thus to approximate the population Gaussian discriminant, we need to choose an appropriate degree for truncation. As the truncation degree N increases, the largest (and only nonzero) eigenvalue λ_N as a measure of class separation naturally increases. We may stop at N where the increment

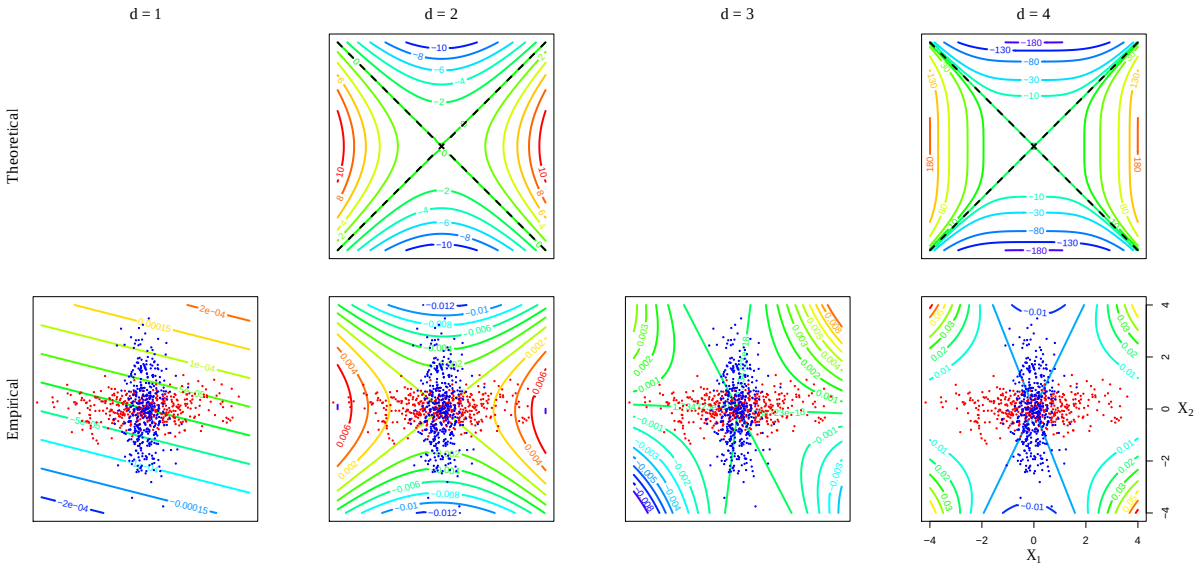


Figure 4: Contours of the population discriminant functions with homogeneous polynomial kernels of degree 1 to 4 (upper panels) and their sample counterparts (lower panels) under Scenario 2. The black dashed lines are the optimal classification boundaries.

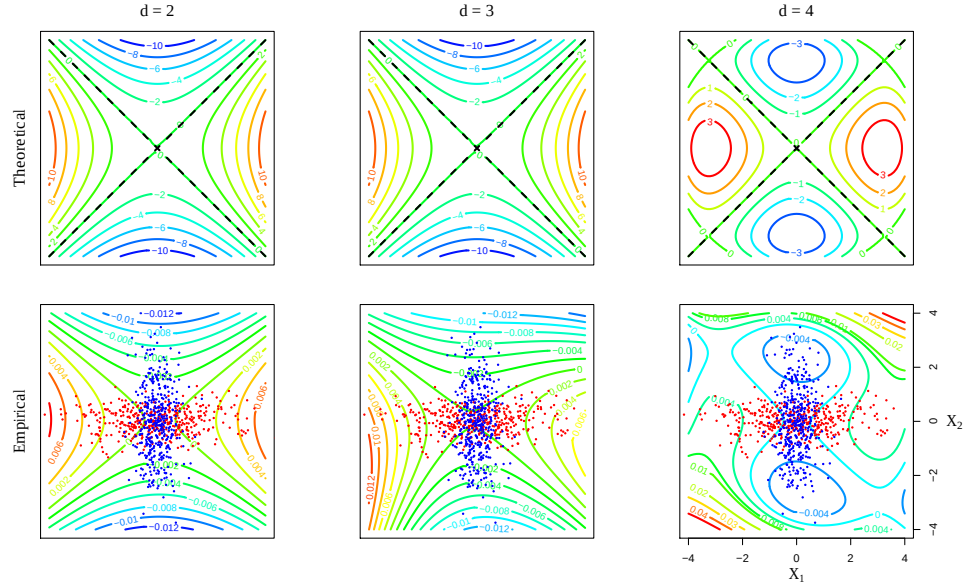
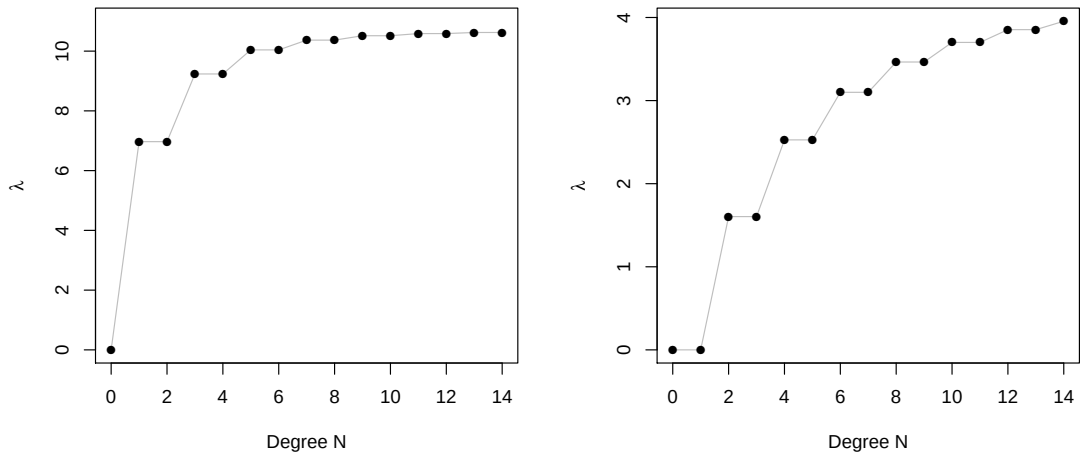


Figure 5: Contours of the population discriminant functions with inhomogeneous polynomial kernels of degree 2 to 4 (upper panels) and their sample counterparts (lower panels) under Scenario 2. The black dashed lines are the optimal classification boundaries.

in λ_N is negligible.

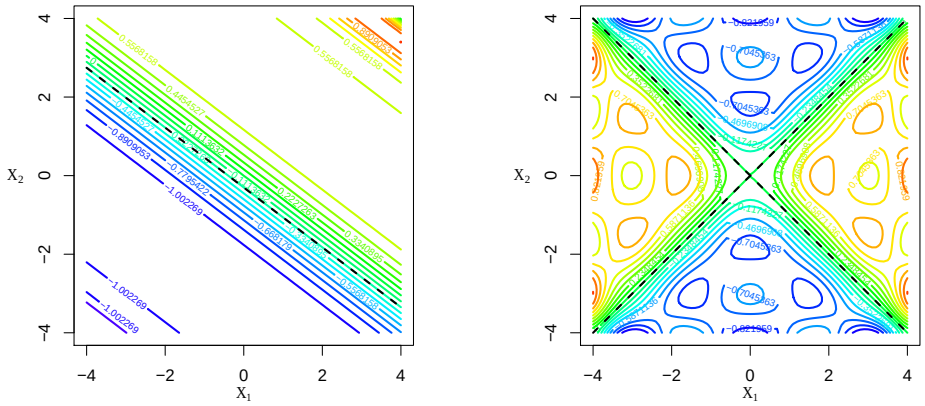
Figure 6 shows how this eigenvalue λ_N changes with degree N for each scenario. In Scenario 1, since a linear component is essential, there is a sharp increase in λ_N at degree 1 followed by a gradual increase as odd features are added. By contrast, in Scenario 2, λ_N steadily increases as even features are added. Overall the magnitude of the maximum ratio of between-class variation to within-class variation (λ_N) indicates



(a) Scenario 1 (b) Scenario 2

Figure 6: The ratio of between-class variation to within-class variation (λ_N) as a function of the truncation degree N under (a) Scenario 1 and (b) Scenario 2.

that Scenario 1 presents an inherently easier problem than Scenario 2. Figure 7 displays some contours of the approximate Gaussian discriminants for each scenario using $N = 14$, which suggest that the Gaussian kernel can capture the difference between classes effectively in both scenarios.



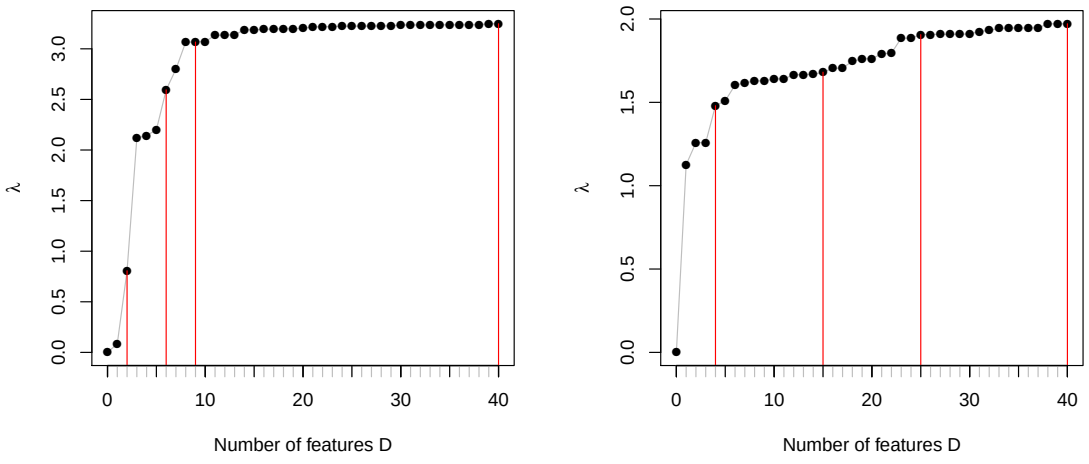
(a) Scenario 1 (b) Scenario 2

Figure 7: Contours of the population Gaussian discriminants approximated by polynomials truncated at degree 14 under (a) Scenario 1 and (b) Scenario 2. The black dashed lines are the optimal classification boundaries.

Random Fourier feature representation: While polynomial features in the deterministic representation are naturally ordered by degree, there is no natural order in random Fourier features. As with degree N for deterministic features, however, the Rayleigh quotient as a measure of class separation or the corresponding

eigenvalue increases as we add more random features. We numerically examine the effect of the number of random features D on the eigenvalue λ_D and monitor the increment in λ_D.

For both scenarios, we randomly generated 40 $\mathbf{w}_i$ from $N_2(\mathbf{0}, I_2)$ and b_i from Uniform$(0, 2\pi)$, and defined phase-shifted cosine features, $z_{\mathbf{w}_i, b_i}(\mathbf{x}) = \sqrt{2}\cos(\mathbf{w}_i^T\mathbf{x} + b_i)$. Figure 8 shows how λ_D changes with D for each scenario. Figure 9 shows how the approximate Gaussian discriminant in (22) changes as the number of random features increases from 2 to 40 under Scenario 1. Figure 10 shows a similar change under Scenario 2. Those snapshots in Figures 9 and 10 are chosen by monitoring the increment in the eigenvalue as more features are added. The number of features used is marked by the red vertical lines in Figure 8 for reference. As D increases, the approximate Gaussian discriminants tend to better approximate the optimal classification boundaries. Compared to the polynomial approximation, the eigenvalues level off quickly with the number of random features D, and the maximum values are far less than their counterparts with polynomial features in both scenarios in part due to the randomness in the choice of $\mathbf{w}_i$ and b_i and the fact that the nature of class difference is not harmonic. In summary, Fourier features are not as effective as polynomial features in these two settings.



(a) Scenario 1 (b) Scenario 2

Figure 8: The ratio of between-class variation to within-class variation (λ_D) as a function of the number of random Fourier features D under (a) Scenario 1 and (b) Scenario 2. The red vertical lines indicate the number of random features used in Figures 9 and 10.

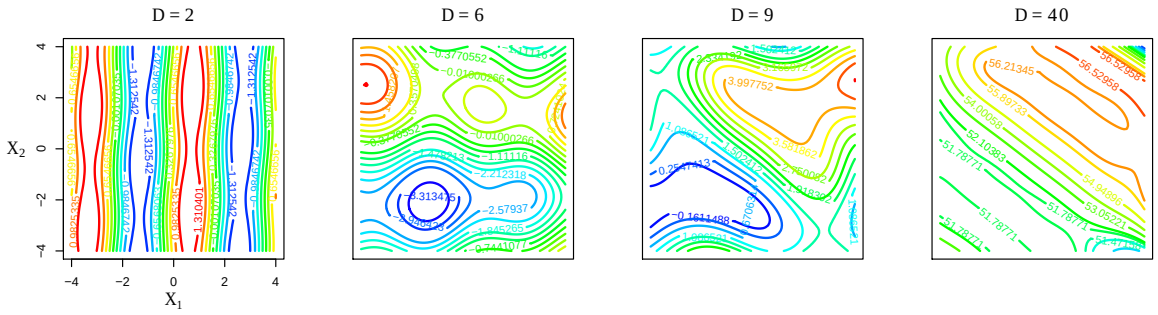


Figure 9: Contours of the approximate discriminant functions using random Fourier features under Scenario 1. The value of D in each panel indicates the number of random Fourier features.

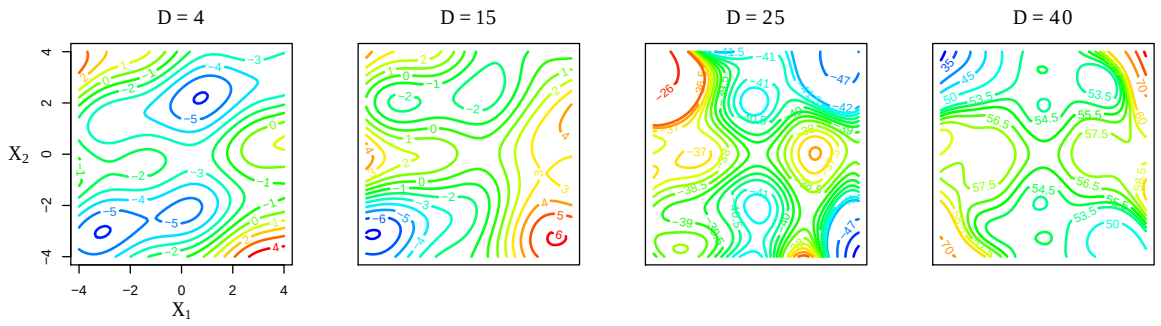


Figure 10: Contours of the approximate discriminant functions using random Fourier features under Scenario 2. The value of D in each panel indicates the number of random Fourier features.

4.2 Real Data Example

In this section, we carry out a kernel discriminant analysis on the spam email data set from the UCI Machine Learning Repository (Dua and Graff 2017). We examine the geometry of sample kernel discriminants with various kernels as in the simulation study, and test the performance of the induced classifiers to see the impact of the kernel choice and kernel parameters.

The data set contains information from 4601 email messages of which 60.6% are regular email and 39.4% spam. The task is to detect whether a given email is regular or spam using 57 predictors available in order to filter out spam. 48 predictors are the percentage of words in the email that match a given word (e.g., credit, you, free), 6 predictors are the percentage of punctuation marks in the email that match a given punctuation mark (e.g., !, $), and additional three predictors are the longest, average, and total length of strings of capital letters in the message.

For ease of illustration, we start with a low dimensional representation of the data using principal components and construct kernel discriminants with those components rather than the individual predictors. We observed that the predictors measuring relative frequencies of words exhibit strong skewness in distribution. To alleviate the skewness, we considered a logit transformation before defining principal components. We also observed a large number of zeros on many predictors as some words do not necessarily appear in every e-mail message. To handle this issue, we replaced zeros with a half of the least nonzero value in each predictor before taking a logit transformation and carried out a principal component analysis on the transformed data using their correlation matrix. We then split the principal component scores into training and test sets of about 60% and 40% each and evaluated the performance of trained classifiers over the test set.

Figure 11 shows the scores on the first two principal components for the training data. The two principal components explain 26% of variation in the original data. The score distributions for two types of email are skewed and substantially overlap with very different covariances, suggesting that a nonlinear boundary is needed for classification.

We performed a kernel discriminant analysis on the training data using the inhomogeneous polynomial kernels of degree 1 to 6, and obtained the corresponding polynomial discriminants. For computational efficiency, we estimated the moment difference $\tilde{\Delta}$ and covariance matrix $\tilde{\mathbf{W}}$ directly using the training data and solved a sample version of (15) instead of (2). Figure 12 shows the estimated coefficients for the discriminants that are normalized to unit length using a color map. High order terms, especially beyond the cubic terms, have negligible coefficients. We need to decide on a threshold for discriminant scores to make a decision for spam filtering. We chose the threshold value by minimizing the training error. Figure 13 displays the decision boundaries of the final discriminant functions using the chosen threshold. All nonlinear

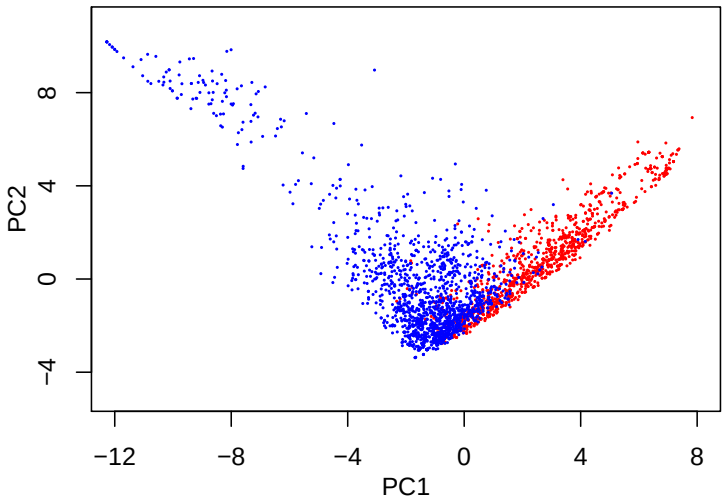


Figure 11: A scatterplot of the first two principal components scores on the email messages in the training data (blue: regular and red: spam).

polynomial discriminants in the figure seem to have similar boundaries at least in the region where data density is high. Table 3 presents their test error rates for comparison along with the rates for misclassifying spam as regular and vice versa. The fifth and sixth order polynomial discriminants have the lowest error rate in this case. However, reduction in the test error rate is marginal after the third order, which we may expect from the result in Figure 12 and diminishing returns in the ratio from degree as shown in Table 3. We may well consider the cubic discriminant sufficient for this application. It provides a good compromise between the two kinds of errors while maintaining simplicity.

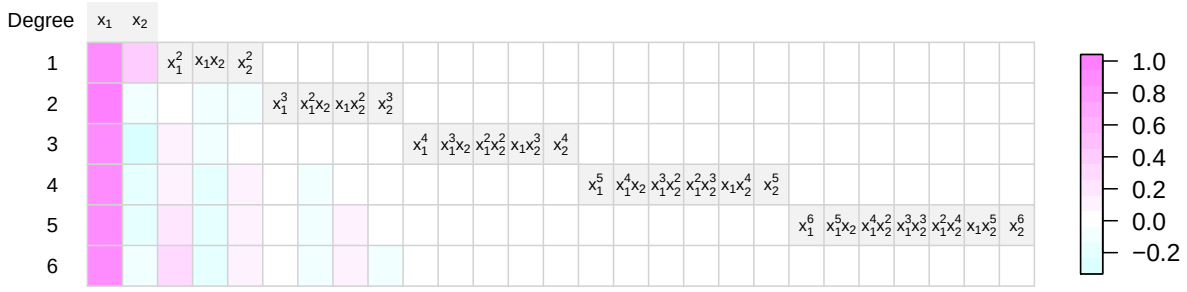


Figure 12: A color map of the estimated coefficients for the polynomial discriminants of degree 1 to 6 using two principal components from the spam email data displayed in the lower triangular array. The column label in the gray band (e.g., $x_1 = PC_1$ and $x_2 = PC_2$) indicates the term corresponding to each coefficient.

5 Discussion

We have examined the population version of kernel discriminant analysis and the generalized eigenvalue problem with between-class and within-class kernel covariance operators to shed light on the relation between the data distribution and resulting kernel discriminant. Our analysis shows that polynomial discrimi-

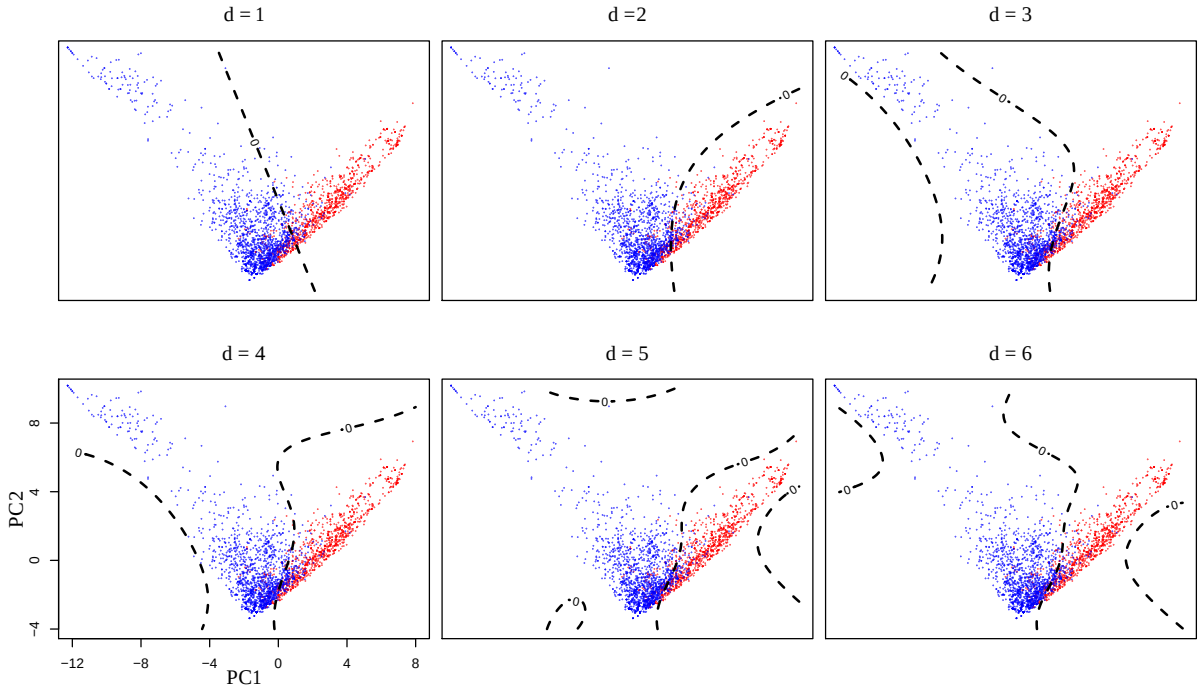


Figure 13: Decision boundaries of the polynomial discriminants with the inhomogeneous polynomial kernels of degree 1 to 6 obtained from the spam email data. The black dashed lines are the boundaries with minimum training error for each kernel.

nants capture the difference between two distributions through their moments of a certain order specified by the polynomial kernel. Depending on the representation of the Gaussian kernel, on the other hand, Gaussian discriminants encode the class difference using all polynomial features or Fourier features of random projections.

Whenever we have some discriminative predictors in the data by design as is typically the case, kernels of a simple form aligned with those predictors will work well. For instance, if we use polynomial kernels in such a setting, we expect the Rayleigh quotient as a measure of class separation to become saturated quickly with degree and low-order polynomial features to prevail. The geometric perspective of kernel

Table 3: Test error rates of kernel discriminant analysis on the spam email data set with the inhomogeneous polynomial kernels of varying degrees. The training error rates and between-class to within-class variation ratio are provided for comparison.

Degree	Ratio	Training error	Test error		
			Misclassified spam	Misclassified regular	Overall
1	4.3154	0.1381	0.2218	0.0744	0.1325
2	6.4226	0.1163	0.1887	0.0645	0.1135
3	7.2928	0.1116	0.1736	0.0645	0.1075
4	7.7043	0.1095	0.1377	0.0959	0.1124
5	8.0941	0.1058	0.1612	0.0672	0.1042
6	8.3192	0.1033	0.1543	0.0717	0.1042

discriminant analysis presented in this paper suggests that the ideal kernel for discrimination retains only those features necessary for describing the difference in two distributions. This promotes a compositional view of kernels (e.g., $\tilde{K}_d(\mathbf{x}, \mathbf{u}) = \sum_{m=0}^{d} \binom{d}{m} K_m(\mathbf{x}, \mathbf{u})$) and further points to the potential benefits of selecting kernel components relevant to discrimination similar to the way feature selection is incorporated into linear discriminant analysis using sparsity inducing penalties (Cai and Liu 2011, Clemmensen et al. 2011). For instance, Kim et al. (2006) formulated a convex optimization problem for kernel selection in KDA. It is also of interest to compare this kernel selection approach with other approaches for numerical approximation of kernel matrices themselves through Nyström approximation (Drineas and Mahoney 2005, Williams and Seeger 2001) or random projections (Ye et al. 2017).

As a related issue, it has not been formally examined how the Rayleigh quotient maximized in kernel discriminant analysis is related to the error rate of the induced classifier except for some special cases only. It is of particular interest how the relation changes with the form of a kernel and associated features given the difference between two distributions.

While our analysis has focused on the case of two classes, we can generalize it to the case of multiple classes where more than one kernel discriminants need to be considered and properly combined to make a decision. We leave this extension as future research.

Acknowledgements

This research was supported in part by the National Science Foundation under grant DMS-15-13566. We thank Professor Mikyoung Lim at KAIST for helpful conversations on linear operators.

References

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

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

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis, *Journal of the American Statistical Association* **106**(496): 1566–1577.

Clemmensen, L., Hastie, T., Witten, D. and Ersboll, B. (2011). Sparse discriminant analysis, *Technometrics* **53**(4): 406–413.

Drineas, P. and Mahoney, M. W. (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning, *J. Mach. Learn. Res.* **6**: 2153–2175.

Dua, D. and Graff, C. (2017). UCI machine learning repository.
URL: *http://archive.ics.uci.edu/ml*

Gu, C. (2002). *Smoothing Spline ANOVA Models*, New York: Springer.

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

Holmquist, B. (1996). The d-variate vector Hermite polynomial of order k, *Linear algebra and its applications* **237**: 155–190.

Joachims, T. (2002). Optimizing search engines using clickthrough data, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, Association for Computing Machinery, New York, NY, USA, p. 133–142.

Kim, S., Magnani, A. and Boyd, S. (2006). Optimal kernel selection in kernel Fisher discriminant analysis, *Proceedings of the 23rd international conference on Machine learning*, ACM, pp. 465–472.

Liang, Z. and Lee, Y. (2013). Eigen-analysis of nonlinear PCA with polynomial kernels, *Statistical Analysis and Data Mining* **6**: 529–544.

Mika, S., Ratsch, G., Weston, J., Schölkopf, B. and Müllers, K. (1999). Fisher discriminant analysis with kernels, *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop*, IEEE, pp. 41–48.

Rahimi, A. and Recht, B. (2008a). Random features for large-scale kernel machines, *Advances in neural information processing systems*, pp. 1177–1184.

Rahimi, A. and Recht, B. (2008b). Uniform approximation of functions with random bases, *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 555–561.

Rudin, W. (2017). *Fourier analysis on groups*, Courier Dover Publications.

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

Schölkopf, B., Smola, A. J. and Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation* **10**: 1299–1319.

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

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*, Cambridge University Press, USA.

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

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

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

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

Williams, C. K. I. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines, *in* T. K. Leen, T. G. Dietterich and V. Tresp (eds), *Advances in Neural Information Processing Systems 13*, MIT Press, pp. 682–688.

Ye, H., Li, Y., Chen, C. and Zhang, Z. (2017). Fast Fisher discriminant analysis with randomized algorithms, *Pattern Recognition* **72**: 82–92.

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