Supervised Linear Dimension Reduction

Wei Bian

Faculty of Engineering and Information Technology

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2012

Certificate of Authorship/Originality

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature

Acknowledgements

On having completed this thesis, I am especially thankful to my advisor Prof. Dacheng Tao, who had led me to an at one time unfamiliar area of academic research, and trusted me and gave me as much as possible freedom to purse my own research interests. Dacheng has taught me how to think and study independently and how to solve a difficult scientific problem in flexible but rigorous ways. He has sacrificed much of his precious time for developing my academic research skills. He has also given me great help and support in life.

I am thankful to the group members I met in the Hong Kong Polytechnic University, Nanyang Technological University, and University of Technology, Sydney, including Tianyi Zhou, Bo Geng, Chao Zhang, Bo Xie, Yang Mu, and many others. I learned a lot from these smart people, and I was always inspired by the interesting and in-depth discussions with them. I enjoyed the wonderful atmosphere,being with them, of both academic research and daily life.

I am incredibly grateful to my mother for her generosity and encouragement. This thesis is definitely impossible to be completed without her constant support and understanding. I am also thankful to my friends who have companied me, though not always at my side, through the arduous journey of four and a half years.

Abstract

Supervised linear dimension reduction (SLDR) is one of the most effective methods for complexity reduction, which has been widely applied in pattern recognition, computer vision, information retrieval, and multimedia data processing. This thesis explores SLDR by enriching the theory of existing methods and by proposing new methods.

In the first part of this thesis, we present theoretical analysis of Fisher's linear discriminant analysis (LDA), one of the most representative methods for SLDR. 1) Classical asymptotic analysis of LDA is based on a fixed dimensionality, and thus does not apply in the case where the dimensionality and the training sample number are proportionally large. Besides, the classical result does not provide quantitative information on the performance of LDA. To address these limitations, we present an asymptotic generalization analysis of LDA, allowing both the dimensionality and the training sample number to be proportionally large, from which we principally obtain an asymptotic generalization bound that quantitatively describes the performance of LDA in terms of the dimensionality and the training sample number. 2) We study a new regularization method for LDA, termed the block-diagonal regularization. By partitioning variables into small groups and treating them independently, block-diagonal regularization effectively reduces the dimensionality to training sample number ratio and thus improves the generalization ability of LDA. We present a theoretical justification of the block-diagonally regularized LDA by investigating its approximation and sample errors. We show that the block-diagonally regularized LDA performs competitively compared to other types of regularized LDA, e.g., with the Tikhonov regularization and the banded regularization.

In the second part of this thesis, we propose two new methods for SLDR. 1) The first method is for parametric SLDR, termed max-min distance analysis (MMDA). MMDA optimizes the projection matrix by maximizing the minimum pairwise distance of all class pairs in the dimension reduced space. Thus, it duly considers the separation of all classes and overcomes the "class separation" problem of existing parametric SLDR methods that close class pairs tend to merge in the dimension reduced space. 2) The second method is for nonparametric SLDR, which uses minimizing the asymptotic nearest neighbor classification error (MNNE) as the criterion for optimizing the projection matrix. Theoretically, we compare MNNE with other criteria, e.g., maximizing mutual information (MMI) and minimizing Bhattacharyya bound. We show that MMNE is superior to these two criteria in terms of the closeness to the Bayes optimal criterion. Empirical studies show that the proposed methods, MMDA and MNNE, achieve state-of-the-art performance for parametric and nonparametric SLDR, respectively.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Dimension Reduction: an Overview

Dimension reduction is an important data processing technique in pattern recognition and machine learning Duda et al. [2001] Devroye et al. [1996] Bishop [2006]. By transforming data from the original space to low-dimensional representations, the subsequent data analysis benefits from many aspects; for example, by providing a two or three-dimensional graphical visualization of data and reducing the computational cost. Most importantly, dimension reduction helps to reduce the complexity of the problem to be studied, and thus improves the performance of data analysis. In density estimation, the phenomenon of *the curse of dimensionality* says that: "... to get good error rates, the number of training samples should be exponentially large in the number of dimensionality" Devroye et al. [1996]. Therefore, dimension reduction is generally a necessary step even when the original dimensionality is moderate. In classification, dimension reduction is useful for improving the generalization ability of a classifier. Taking linear classification, for example: reducing data dimensionality reduces the VC (Vapnik-Chervonenkis) dimension of a classification algorithm, enabling better generalization to be obtained with the same number of training samples Vapnik [1998].

The study on dimension reduction has a long history in pattern recognition and statistics, and has received increasing attention over the past decade due to the demand of high-dimensional data analysis. According to the information used or the characteristics of the data transformation, dimension reduction can be categorized in the following two ways:

- *Supervised v.s. Unsupervised.* The difference between supervised and unsupervised dimension reduction lies in whether the label information of the data is utilized or accessible. Generally, the purpose of supervised dimension reduction is to improve classification performance. Thus, it aims to find a low-dimensional data representation by which different classes of data can be well separated. Popular supervised dimension reduction methods include Fisher's linear discriminant analysis (LDA) Fisher [1936] Rao [1948], kernel discrimination analysis (KDA) Mika et al. [1999] and nonparametric discriminant analysis (NDA) Fukunaga and Mantock [1983]. Unsupervised dimension reduction does not utilize label information, and is usually performed for de-noising or obtaining a compact/semantic data representation. In contrast to supervised dimension reduction, which focuses mostly on the discrimination between classes, unsupervised dimension reduction is more diverse due to the different aspects of the data characteristics concerned. For instance, when aiming for minimum reconstruction error, principal component analysis (PCA) Jolliffe [2002] should be applied, while if intending to discover the latent uncorrelated factors behind observations, factor analysis (FA)Gorsuch [1983] will be preferable. Semantic data representation has recently become very popular in the machine learning field, and many algorithms have been developed, e.g., probabilistic latent semantic analysis (pLSA) Hofmann [1999] and latent Dirichlet allocation (LDA) Blei et al. [2003]. Although these algorithms are often categorized as generative probabilistic models, they do perform unsupervised dimension reduction functionally.

- *Linear v.s. Nonlinear.* Linear dimension reduction exploits linear projection as data transformation. It learns a projection matrix $\mathbf{W} \in \mathbb{R}^{D\times d}$, such that $\mathbf{z} = \mathbf{W}^T\mathbf{x}$ is the dimension reduced representation of the original data $\mathbf{x} \in \mathbb{R}^D$. Just as linear models are generally the simplest way to conduct statistical inference, so linear dimension reduction is superior to nonlinear dimension reduction in terms of model complexity. The most representative linear dimension reduction methods include Fisher's LDA and PCA, both of which have been applied in a wide range of areas, from computer vision and information retrieval to multimedia applications. Nonlinear dimension reduction uses nonlinear functions $f : \mathbb{R}^D \mapsto \mathbb{R}^d$ as data transformation, i.e., $\mathbf{z} = f(\mathbf{x})$. It is more satisfactory when the nonlinear structure of data distribution becomes important or the trade-off between the training sample number and model complexity is not difficult. Two major subcategories of nonlinear dimension reduction are kernel dimension reduction and manifold learning. The former can be understood as performing linear dimension reduction in a very high (possibly infinite) dimensional space induced by a kernel function, and the most well-known methods are kernel principal component analysis (KPCA) Schölkopf et al. [1997] and KDA Mika et al. [1999]. Manifold learning assumes that data are distributed on a low-dimensional manifold, which is embedded in the original high-dimensional ambient space. Accordingly, in obtaining the low-dimensional representation, a major aim of manifold learning is to preserve the local geometric structure of data distribution. Popular manifold learning methods include locally linear embedding (LLE) Roweis and Saul [2000], ISOMAP Tenenbaum et al. [2000] , Laplacian eigenmaps (LE) Belkin and Niyogi [2003], Hessian eigenmaps (HLLE) Donoho and Grimes [2003], Generative Topographic Mapping (GTM) Bishop et al. [1998], and local tangent space alignment (LTSA) Zhang and Zha [2004].

Since dimension reduction has a considerably large literature, it is impossible to include all the references here. We point readers to several useful surveys on dimension reduction techniques, including Fodor [2002], Cayton [2005], Saul et al., and Pless and Souvenir [2009].

1.2 Supervised Linear Dimension Reduction

This thesis is devoted to a subcategory of dimension reduction, i.e., supervised linear dimension reduction (SLDR). For a multi-class problem, with a joint probability distribution $p(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^D$ and $y \in \{1, 2, ..., c\}$, SLDR can be described informally as

> *Given the conditional probability densities $p(\mathbf{x}|y = i)$, $i = 1, 2, ..., c$, SLDR aims to find a linear transformation $\mathbf{z} = \mathbf{W}^T\mathbf{x}$, wherein $\mathbf{W} \in \mathbb{R}^{D \times d}$, such that the conditional probability densities after dimension reduction, $p(\mathbf{z}|y = i)$, $i = 1, 2, ..., c$, can be well separated from one another.*

Two key issues in SLDR are 1) estimating conditional probability densities, $p(\mathbf{x}|y = i)$ and/or $p(\mathbf{z}|y = i)$, and 2) defining the separability measurement among multiple $p(\mathbf{z}|y = i)$, which we refer to as the discrimination power. In addressing these two problems, different types of SLDR methods have been proposed. For instance, there are parametric and nonparametric SLDR methods regarding conditional density estimation. The former assumes that data are generated from certain probability distribution families, e.g., the homoscedastic or heteroscedastic Gaussian distributions, while the latter is generally distribution-free and utilizes nonparametric methods to perform density estimation. As for the definition of discrimination power, Bayes error should be the optimal choice whenever classification is the data analysis task. We refer to minimizing Bayes error as the Bayes optimal criterion for SLDR. However, due to the difficulty in calculating Bayes error, the Bayes optimal criterion is only tractable for quite limited data distributions, e.g., the homoscedastic Gaussian distributions. For general cases, a practical strategy is to use a proxy to approximate the Bayes optimal criterion.

1.2.1 Parametric Methods

The most remarkable method for parametric SLDR is Fisher's linear discriminant analysis (LDA), first proposed by Fisher [1936] for binary classification and then extended by Rao [1948] to the multi-class scenario. It assumes that the conditional densities, $p(\mathbf{x}|y = j)$, are homoscedastic Gaussian distributions, i.e., $p(\mathbf{x}|y = j) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_j$ is the j-th class mean and $\boldsymbol{\Sigma}$ is the common covariance matrix. In LDA, the discrimination power is defined by the ratio of the between-class scatter to the within-class scatter, and the optimal projection matrix $\mathbf{W}^*$ that gives the maximized discrimination power can be obtained by generalized eigendecomposition. One important property of LDA is that for a $c+1$-class problem, it obtains a c dimensional subspace which is asymptotically Bayes optimal, i.e., asymptotically, the subspace has the minimum Bayes error among all c dimensional subspaces and also the same Bayes error as in the original space. This is often referred to as the asymptotical Bayes optimality of LDA.

Another approach to parameterizing SLDA, with a less restrictive assumption than LDA, is to allow different classes to have distinct covariance structures, i.e., exploiting heteroscedastic Gaussian distributions for conditional density modeling. To define the discrimination power between heteroscedastic Gaussian distributions, Loog and Duin [2004] proposed to utilize the Chernoff bound, and the resulting Chernoff Criterion (CC) can be regarded as a proxy of the Bayes optimal criterion under the heteroscedastic Gaussian assumption. Information quantities, such as Kullback-Leibler divergence, have also been used for the same purpose of discrimination power measurement Decell and Mayekar [1977] Tao et al. [2009]. Since the heteroscedastic Gaussian assumption takes into account the discrimination information provided by covariance matrices, the corresponding SLDA methods generally outperform LDA given a sufficient number of training samples. However, when training samples are limited, these methods can show inferior performance compared to LDA due to the inaccuracy of the estimation of multiple covariance matrices.

1.2.2 Nonparametric Methods

In order to extend SLDR to the scenario of general data distributions, i.e., developing a distribution-free SLDR method, nonparametric approaches are usually applied. Fukunaga and Mantock [1983] proposed the first nonparametric SLDR method, termed nonparametric discriminant analysis (NDA). NDA follows the same idea as LDA by defining the discrimination power as the ratio of the between-class scatter to the within-class scatter. However, its between-class scatter is defined in a nonparametric way, distinct from that of LDA which uses the means of Gaussian distributions. Further, NDA puts more weights on the samples near the classification boundary when calculating the between-class scatter and is able to deal with nonlinearly separable distributions.

Kernel density estimation has also been introduced to nonparametric SLDR. For example, Lee and Landgrebe [1993] proposed to use the kernel method for conditional density estimation and to define the discrimination power, similar to NDA, by considering the decision boundary of a nonparametric classifier. Torkkola [2003] also utilized the kernel method to estimate the conditional densities of different classes, and then introduced quadratic mutual information based on Renyi's entropy to measure the discrimination power among classes.

1.2.3 Proxies of the Bayes Optimal Criterion

Although the Bayes optimal criterion is theoretically the best choice for SLDR, the calculation of Bayes error is generally intractable, since it requires a complex integral over the entire data space. Therefore, a tractable strategy for SLDR is to use a proxy criterion to approximate the Bayes optimal criterion. Even in the case of LDA, Fisher's criterion is a proxy criterion, which can be proved to be equivalent to the Bayes optimal criterion under the homoscedastic Gaussian assumption. For general data distributions, minimizing an upper bound of Bayes error is a common choice as a proxy of the Bayes optimal criterion. For example, Loog and Duin [2004] used the Chernoff bound and proposed the Chernoff Criterion (CC) for SLDR, while Saon and Padmanabhan [2001] proposed a proxy criterion which minimizes a generalized Bhattacharyya Bound. Both criteria show promising performance under the heteroscedastic Gaussian assumption of data distributions. Another proxy criterion is the maximizing mutual information (MMI), proposed by Torkkola [2003], which is based on the fact that the conditional entropy provides an upper bound of Bayes error and minimizing conditional entropy is equivalent to maximizing mutual information. MMI further integrates the kernel method for conditional density estimation and shows state-of-the-art performance for nonparametric SLDR.

1.3 Contributions of This Thesis

This thesis explores SLDR from two aspects: theoretical analyses and algorithmic extensions. In the first part of this thesis, we present theoretical analyses of LDA, one of the most representative methods for SLDR. Specifically, we prove an asymptotic generalization bound of LDA, and based on this result we propose a new regularization method for LDA. In the second part of this thesis, we present two new methods for parametric and nonparametric SLDR, respectively. The parametric method is based on the criterion of maximizing the minimum pairwise distance in the dimension reduced space, which we refer to as max-min distance analysis (MMDA). MMDA solves the "class separation" problem suffered by existing parametric SLDR methods, including LDA and its many extensions. The nonparametric method utilizes minimizing the nearest neighbor classification error (MNNE) as the criterion for dimension reduction. By examining the closeness to the Bayes optimal criterion, we show that MNNE is superior to other criteria for nonparametric SLDR, e.g., maximizing mutual information (MMI) and minimizing the Bhattacharyya bound.

1.3.1 Theoretical Analyses

LDA is one of the most representative methods for SLDR, and also a fundamental model in pattern recognition and statistics. However, given its popularity and wide applications, there are few results on the theoretical analysis of LDA; for example, an exact generalization analysis of LDA has not been found in the literature. To enrich the theory of LDA, we present an asymptotic generalization analysis of LDA, from which we principally obtain an asymptotic generalization bound that quantitatively describes the performance of LDA in terms of the dimensionality and the training sample number. Motivated by the bound, we propose a block-diagonal regularization LDA, which shows favorable performance in dealing with problems with insufficient training samples.

1.3.1.1 Chapter 2

This chapter is devoted to an asymptotic generalization analysis of LDA, which aims to provide a quantitative description of the performance of LDA in terms of the dimensionality D and the training sample number N. Classical asymptotic analysis shows that for a fixed D the generalization discrimination power of LDA approaches the population discrimination power as N goes to infinity. However, this theory is inferior in two aspects. First, it assumes a fixed D and a sufficiently large N, which makes the theory inapplicable for practical problems where D and N are proportionally large. Second, it does not provide quantitative information on the performance of LDA, and thus we are still unaware of how large N needs to be with respect to D for LDA to obtain an acceptable generalization discrimination power. To address these limitations, we present an asymptotic generalization analysis of LDA. Unlike classical results based on multivariate statistics, our analysis is carried out by using powerful tools from random matrix theory. The asymptotic generalization bound obtained allows D and N to be proportionally large and quantitatively describes how the dimensionality to training sample number ratio D/N affects the performance of LDA.

1.3.1.2 Chapter 3

This chapter studies regularized LDA. It is common in practical applications, e.g., face recognition, that the training sample number N is insufficient with respect to the dimensionality D. In such a case, the direct application of LDA will readily fail due to the highly inaccurate parameter estimation. To address this problem, regularization has been introduced to LDA. Motivated by the asymptotic generalization bound obtained in Chapter 2, we propose a new regularization method for LDA, termed the block-diagonal regularization. By partitioning variables into k groups and treating them independently, block-diagonal regularization effectively reduces the dimensionality to training sample number ratio and thus improves the generalization ability of LDA. We present a theoretical justification of the block-diagonally regularized LDA by investigating its approximation and sample errors. Empirically, we evaluate the block-diagonally regularized LDA by face recognition experiments, and compare it with other types of regularized LDA, e.g., with the Tikhonov regularization and the banded regularization.

1.3.2 Algorithmic Extensions

In algorithmic extensions, we treat parametric and nonparametric SLDR separately, because the two subcategories have different application scenarios. Generally, parametric SLDR has a lower model complexity and requires relatively fewer training samples, and thus is favorable for high-dimensional applications, e.g., face recognition. In contrast, nonparametric SLDR has a higher model complexity and requires more training samples, which is more favorable for moderate dimensional problems. We propose two new methods for parametric and nonparametric SLDR, respectively. We show that they perform competitively compared to the state-of-the-art methods.

1.3.2.1 Chapter 4

This chapter proposes a new method for parametric SLDR, termed max-min distance analysis (MMDA). A major problem with existing parametric SLDR methods, including LDA and its many extensions, is that when the dimensionality of the learned subspace is low close class pairs tend to merge. This is referred to as the "class separation" problem in the literature, and it has received considerab attention in recent years. MMDA is proposed to solve the class separation problem. It optimizes the projection matrix by maximizing the minimum pairwise distance among all class pairs in the dimension reduced space. Thus, it duly considers the separation of all classes. Unfortunately, MMDA is hard optimize directly due to the non-smoothness of the objective function and the orthonormal constraints. Therefore, we derive an approximate algorithm for MMDA by using the sequential convex relaxation technique. Empirical evaluations, on both synthetic data experiments and face recognition, show the competitive performance of MMDA compared to the state-of-the-art parametric SLDR methods.

1.3.2.2 Chapter 5

This chapter proposes a new method for nonparametric SLDR, which optimizes the projection matrix by minimizing the asymptotic nearest neighbor classification error (MNNE). Previous study shows that asymptotic nearest neighbor classification error upper bounds Bayes error by a factor of at most 2. Therefore, MNNE can be regarded as a proxy of the Bayes optimal criterion for SLDR. In the literature, maximizing mutual information (MMI) and minimizing the Bhattacharyya bound have also been utilized as proxy criteria for SLDR. One of our contributions is that we prove MNNE is superior to these two criteria in terms of the closeness to the Bayes optimal criterion. We derive an algorithm for MNNE, based on kernel density estimation and a gradient descent method on the Grassmann manifold. Empirical evaluations on real datasets show the promising performance of MNNE compared to the state-of-the-art nonparametric SLDR methods.

1.4 Notations

Throughout this thesis, we will use the following notations. Lower case letter $\mathbf{a}$ denotes a vector. Bold upper case letter $\mathbf{A}$ denotes a matrix. $\mathbb{R}^D$ denotes a D-dimensional vector space. $\mathbb{R}^{D_1 \times D_2}$ denotes the set of all D_1 by D_2 matrices. $\mathbf{A}_{ii}$ or $\{\mathbf{A}\}_{ii}$ denotes the i-th diagonal entry of a symmetric matrix $\mathbf{A}$. $\mathbf{A}_i$ denotes the i-th column of $\mathbf{A}$. $\mathbf{A}_{1:c}$ denotes the matrix composed by the first c columns of $\mathbf{A}$. $\mathbb{S}^{D-1}$ denotes the D-dimensional unit sphere located on the original point. $\mathbb{S}_{+}^{D\times D}$ and $\mathbb{S}_{++}^{D\times D}$ denotes the set of all D by D positive semidefinite and positive definite matrices, respectively. $\mathbf{A} \preceq \mathbf{B}$ denotes $\mathbf{B} - \mathbf{A}$ is positive semidefinite. $\mathrm{diag}(\mathbf{A}, \mathbf{B})$ denotes a block-diagonal matrix composed by $\mathbf{A}$ and $\mathbf{B}$. $\|\mathbf{a}\|$ denotes the ℓ_2 norm of $\mathbf{a}$. $\|\mathbf{A}\|$ denotes the operator norm, i.e., the largest singular value, of $\mathbf{A}$. $\lambda_i(\mathbf{A})$ denotes the i-th eigenvalue of $\mathbf{A}$, sorted in a descent order. $\Lambda(\mathbf{A})$ denotes the diagonal matrix composed of the eigenvalues of $\mathbf{A}$, with the eigenvalues sorted in a descent order. $\det(\mathbf{A})$ denotes the determinant of $\mathbf{A}$. $\mathcal{R}(\mathbf{A})$ denotes the range or the column space of $\mathbf{A}$.

Part I

Theoretical Analyses

Chapter 2

Asymptotic Generalization Analysis of Linear Discriminant Analysis

2.1 Introduction

Fisher's linear discriminant analysis (LDA) Fisher [1936] Rao [1948] is among the most representative SLDR methods. Given $c+1$ classes in space $\mathbb{R}^D$, represented by homoscedastic Gaussian distributions, LDA selects a c-dimensional subspace by simultaneously minimizing the within-class scatter and maximizing the between-class scatter. Since the within- and between-class scatters are measured by the sample covariance $\widehat{\boldsymbol{\Sigma}}$ and sample means $\widehat{\boldsymbol{\mu}}_i$, it can be shown by multivariate statistics Anderson [1984] that the c-dimensional subspace selected by LDA is asymptotically Bayes optimal, under conditions that the dimensionality D is fixed and the training sample number N goes to infinity.

Because of its asymptotic Bayes optimality, we can trust the discriminative subspace selected by LDA provided N is sufficiently large compared to D. However, the requirement of "sufficiently large" N is still unclear, and asymptotic Bayes optimality does not provide quantitative justification for the performance of LDA. Besides, practical problems often encounter the situation that D and N are proportionally large, or have the same order of magnitudes, e.g., face recognition and object categorization. In this case, the asymptotic results from classical multivariate statistics Anderson [1984] become invalid. For example, the sample covariance may significantly deviate from its population counterpart when both D and N are large Yin et al. [1988] EL Karoui [2008]. As a result, the asymptotic Bayes optimality of LDA, which is built upon classical multivariate statistics, is no longer applicable.

Given aforementioned limitations of the asymptotic Bayes optimality, but the practical importance of LDA, there is a need of establishing new theoretical results to justify the performance of this SLDR method. We fulfill this by proving an asymptotic generalization bound of LDA, in which we allow both D and N increase and the ratio $D/N \longrightarrow \gamma \in (0,1)$. First, the new result, i.e., the asymptotic generalization bound, is applicable to the situations where D and N are proportionally large. Second, it provides quantitative justification of the performance of LDA, by showing how the dimensionality to training sample number ratio D/N affects the generalization discrimination power preserved by empirical learning over training samples. Informally speaking, given the population discrimination power $\boldsymbol{\lambda}$, the generalization discrimination power of LDA should be larger than

$$\cos^2(\arccos(\sqrt{\boldsymbol{\lambda}/(\boldsymbol{\lambda}+\gamma)}) + \arccos(\sqrt{1-\gamma}))\boldsymbol{\lambda}$$

under mild conditions. Compared with the asymptotic Bayes optimality, such result is considerably informative; for example, if $\gamma = 0.2$ and $\boldsymbol{\lambda} \geq 10$, we know that LDA would preserve about 70% of the discrimination power.

The technical tools used in developing the new theory are from Random matrix theory (RMT) Wigner [1955] Wigner [1958] Marčenko and Pastur [1967] Bai and Silverstein [1998] Edelman and Rao [2005]. The main goal of RMT is to provide understanding of the diverse properties, most notably, statistics of eigenvalues, of matrices with entries drawn randomly from various probability distributions. RMT was originally motivated by applications in nuclear physics in 1950's, and after that it was intensively studied in mathematics and statistics. It also found successful applications in engineering fields, e.g., wireless communications Tulino and Verdú [2004], recently. In this chapter, we make use of two important results from RMT. The first result is the Marčenko-Pastur Law Marčenko and Pastur

[1967], which states that the empirical spectral distribution of the eigenvalues of the sample covariance converges almost surely to a deterministic distribution $F_\gamma(\lambda)$ as $D/N \longrightarrow \gamma \in [0, \infty)$. The second result is on the almost sure convergence of the extreme singular values of a large Gaussian random matrix. Both results play fundamental roles in our proof of the asymptotic generalization bound of LDA.

The rest of this chapter is organized as follows. Section 2.2 introduces LDA and briefly reviews its asymptotic Bayes optimality under the condition of fixed dimensionality. Section 2.3 presents our main result, an asymptotic generalization bound of LDA. Section 2.4 shows empirical evaluations of the asymptotic generalization bound on both synthetic and real datasets. Technical proofs are arranged as appendixes in Section 2.6.

2.2 LDA and its Asymptotic Optimality with Fixed Dimensionality

The motivation of LDA is as follows. Given $c+1$ classes in a high-dimensional space $\mathbb{R}^D$, it seeks a linear projection $\mathbf{z} = \mathbf{W}^T\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{W} \in \mathbb{R}^{D\times d}$, such that the discrimination power among the classes is maximally preserved after the projection. Suppose the $c+1$ classes are represented by homoscedastic Gaussian distributions, $\mathcal{N}_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2, ..., c+1$, with the class means $\boldsymbol{\mu}_i \in \mathbb{R}^D$ and the common covariance matrix $\boldsymbol{\Sigma} \in \mathbb{S}_{++}^{D\times D}$, and they have equal prior probabilities[1] $\frac{1}{c+1}$. According to Fisher's criterion, the discrimination power in the dimension reduced space is given by

$$\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\mathbf{W}) = \mathrm{Tr}\left((\mathbf{W}^T\boldsymbol{\Sigma}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S}\mathbf{W}\right), \tag{2.1}$$

where the matrix

$$\mathbf{S} = \frac{1}{c+1}\sum_{i=1}^{c+1}(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T, \text{ with } \boldsymbol{\mu} = \frac{1}{c+1}\sum_{i=1}^{c+1}\boldsymbol{\mu}_i, \tag{2.2}$$

[1] This does not substantially affect the results obtained later and is only assumed for the convenience of expression.

is called the between-class scatter matrix, which measures the separation between classes' centers. Therefore, the optimal projection matrix $\mathbf{W}^*$ of LDA can be obtained by the maximization problem,

$$\mathbf{W}^* = \arg \max_{\mathbf{W} \in \mathbb{R}^{D \times d}} \Delta(\mathbf{\Sigma}, \mathbf{S} | \mathbf{W}). \tag{2.3}$$

Observing that $\mathbf{S}$ defined in (2.2) has only rank c, thus it does not affect the attaining of the optimal value of the maximization problem (2.3) to restrict $\mathbf{W} \in \mathbb{R}^{D \times c}$. In other words, a c-dimensional subspace is sufficient to preserve all the discrimination power as defined in (2.1). Moreover, by basic property of trace operator, (2.1) is invariant to the transformation $\mathbf{W} \leftarrow \mathbf{WA}$, with $\mathbf{A} \in \mathbb{R}^{c \times c}$ being any nonsingular matrix. Thus, we can further require $\mathbf{W}^T \mathbf{\Sigma} \mathbf{W} = \mathbf{I}_c$, which also does not affect (2.3). As a result, we can rewrite (2.3) as below

$$\mathbf{W}^* = \arg \max_{\mathbf{W}^T \mathbf{\Sigma} \mathbf{W} = \mathbf{I}_c} \Delta(\mathbf{\Sigma}, \mathbf{S} | \mathbf{W}). \tag{2.4}$$

Usually, (2.4) is solved by the generalized eigendecomposition

$$\mathbf{S} \boldsymbol{\zeta}_i = \boldsymbol{\lambda}_i \mathbf{\Sigma} \boldsymbol{\zeta}_i, \tag{2.5}$$

and $\mathbf{W}^*$ is composed of the first c eigenvectors, $\boldsymbol{\zeta}_i$, $i = 1, 2, ..., c$. However, for the convenience of the theoretical analysis, we utilize the simultaneous diagonalization Fukunaga [1990], an equivalent characterization of the generalized eigendecomposition, to describe $\mathbf{W}^*$. This is given in the proposition below.

Proposition 2.1. *There exists a nonsingular matrix* $\mathbf{X}^* = [\mathbf{W}^* \ \mathbf{V}^*]$, *with* $\mathbf{W}^* \in \mathbb{R}^{D \times c}$ *and* $\mathbf{V}^* \in \mathbb{R}^{D \times (D-c)}$, *that simultaneously diagonalizes* $\mathbf{\Sigma}$ *and* $\mathbf{S}$, *i.e.,*

$$\mathbf{X}^{*T} \mathbf{\Sigma} \mathbf{X}^* = \mathbf{I} \text{ and } \mathbf{X}^{*T} \mathbf{S} \mathbf{X}^* = \mathbf{\Lambda}, \tag{2.6}$$

where $\mathbf{\Lambda}$ *is a diagonal matrix, with only the first* c *diagonal entries being nonzero. Further,* $\mathbf{X}^*$ *can be explicitly expressed as*

$$\mathbf{X}^* = \mathbf{\Sigma}^{-\frac{1}{2}} \mathbf{U}^*, \tag{2.7}$$

where $\mathbf{U}^*$ *is from the eigendecomposition* $\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{U}^*\mathbf{\Lambda}\mathbf{U}^{*T}$*; and,*

$$\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*) = \sum_{i=1}^{c} \boldsymbol{\lambda}_i, \tag{2.8}$$

where $\boldsymbol{\lambda}_i$*,* $i = 1, 2, ..., c$*, is the first* c *diagonal entries of* $\mathbf{\Lambda}$*.*

In fact, the nonzero diagonal entries $\boldsymbol{\lambda}_i$, $i = 1, 2, ..., c$, of $\mathbf{\Lambda}$ are just the nonzero eigenvalues of (2.5), and $\mathbf{W}^*$ and $\mathbf{V}^*$ are the two invariant subspaces associated to the nonzero and zero eigenvalues, respectively. In addition, the $\boldsymbol{\lambda}_i$'s measure the discrimination power in each of the c directions $\mathbf{W}^*_i$, $i = 1, 2, ..., c$. Since the rest diagonal entries of $\mathbf{\Lambda}$ are all zero, (2.8) explains why a c-dimensional subspace is sufficient to preserve all discrimination power among the $c+1$ classes.

All above discussions are based on known population parameters, i.e., $\mathbf{\Sigma}$ and $\mathbf{S}$. In practice, we usually do not have access to these parameters but have a set of training samples. Suppose there are n samples $\mathbf{x}^i_j$ for each class, $i = 1, 2, ..., c+1$, $j = 1, 2, ..., n$, and in total $N = (c+1)n$ training samples for all classes. We have the following empirical estimates for $\mathbf{\Sigma}$ and $\mathbf{S}$,

$$\widehat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{i=1}^{c+1}\sum_{j=1}^{n}(\mathbf{x}^i_j - \widehat{\boldsymbol{\mu}}_i)(\mathbf{x}^i_j - \widehat{\boldsymbol{\mu}}_i)^T, \tag{2.9}$$

$$\widehat{\mathbf{S}} = \frac{1}{c+1}\sum_{i=1}^{c+1}(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}})(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}})^T, \tag{2.10}$$

where

$$\widehat{\boldsymbol{\mu}}_i = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}^i_j \text{ and } \widehat{\boldsymbol{\mu}} = \frac{1}{c+1}\sum_{i=1}^{c+1}\widehat{\boldsymbol{\mu}}_i. \tag{2.11}$$

Therefore, the empirical estimate of $\mathbf{W}^*$ can be obtained by

$$\widehat{\mathbf{W}}^* = \arg \max_{\mathbf{W}^T\widehat{\mathbf{\Sigma}}\mathbf{W}=\mathbf{I}_c} \Delta(\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{S}}|\mathbf{W}). \tag{2.12}$$

The key question regarding the performance of the empirical learning of LDA is how much of the population discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*)$ is preserved in the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$? By using results from classical multivariate statistics Anderson [1984], we can answer this question by the following proposition.

Proposition 2.2. *Fixing dimensionality D, as the training sample number $N \longrightarrow \infty$, it holds*

$$\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) - \Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*) \xrightarrow{a.s.} 0. \tag{2.13}$$

Indeed, it can be shown that Proposition 2.2 is a corollary of the convergence of empirical estimates $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{S}}$. As $N \to \infty$, by the strong law of large numbers (SLLN), we have $\widehat{\mathbf{\Sigma}} \xrightarrow{a.s.} \mathbf{\Sigma}$ and $\widehat{\mathbf{S}} \xrightarrow{a.s.} \mathbf{S}$. Thus, it holds $\widehat{\mathbf{W}}^* \xrightarrow{a.s.} \mathbf{W}^*$, which consequently gives rise to (2.13).

The almost sure convergence in Proposition 2.2 provides a theoretical guarantee, at least in the case of a large training sample number, to the performance of LDA. However, such classical result is limited by remarkable weaknesses:

1. *The almost sure convergence is obtained based on the condition of a fixed dimensionality D, and it is not applicable for practical problems where the dimensionality D is proportionally large to the training sample number N.*

2. *It does not provide quantitative results on the performance of LDA. Especially, given the dimensionality D and the training sample number N, it is still unknown how the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$ is compared to the population discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*)$.*

2.3 Asymptotic Generalization Analysis

In this section, we propose to give an asymptotic generalization analysis of LDA in the setting where the dimensionality D increases proportionably with the training sample size N and the ratio D/N $(D < N)$ has a positive limit $\gamma \in (0, 1)$. It will be clear at the end of this section our new result overcomes all the aforementioned weaknesses of the classical result on LDA.

The analysis is completed by three steps. First, we express the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$ in a new formula with two auxiliary estimates $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{S}}_0$ which are independent of population parameters, e.g., the covariance matrix, but only depend on D and N. Next, we present asymptotic results on the eigensystems (eigenvalues and eigenvectors) of $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{S}}_0$. With above results, we finally obtain an asymptotic lower bound of the generalization discrimination power of LDA.

2.3.1 Generalization Discrimination Power

Recall the generalization discrimination power

$$\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) = \mathrm{Tr}((\widehat{\mathbf{W}}^{*T}\mathbf{\Sigma}\widehat{\mathbf{W}}^*)^{-1}\widehat{\mathbf{W}}^{*T}\mathbf{S}\widehat{\mathbf{W}}^*), \tag{2.14}$$

where $\widehat{\mathbf{W}}^*$ is given by

$$\widehat{\mathbf{W}}^* = \arg \max_{\mathbf{W}^T\widehat{\mathbf{\Sigma}}\mathbf{W}=\mathbf{I}_c} \Delta(\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{S}}|\mathbf{W}). \tag{2.15}$$

At first sight, it may be the case that $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$ is affected by population parameters $\mathbf{\Sigma}$ and $\mathbf{S}$. For example, it is reasonable to think that a problem with a simple covariance matrix, say $\mathbf{I}$, is easy for empirical learning by LDA, and thus could lead to better generalization ability, i.e., larger $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$. However, in the following, we show that $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$ is independent of the covariance structure and only depends on dimensionality D and training sample size N, given fixed population discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*)$.

First, we introduce two auxiliary estimates

$$\widehat{\mathbf{\Sigma}}_0 = \mathbf{X}^{*T}\widehat{\mathbf{\Sigma}}\mathbf{X}^* \text{ and } \widehat{\mathbf{S}}_0 = \mathbf{X}^{*T}\widehat{\mathbf{S}}\mathbf{X}^*. \tag{2.16}$$

where $\mathbf{X}^*$ is from Proposition 2.1, i.e., it simultaneously diagonalizes $\mathbf{\Sigma}$ and $\mathbf{S}$,

$$\mathbf{X}^{*T}\mathbf{\Sigma}\mathbf{X}^* = \mathbf{I} \text{ and } \mathbf{X}^{*T}\mathbf{S}\mathbf{X}^* = \mathbf{\Lambda}. \tag{2.17}$$

From (2.16) and (2.17), we know that $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{S}}_0$ are the empirical estimates of $\mathbf{I}$ and $\mathbf{\Lambda}$, respectively. This is summarized in Proposition 2.3

Proposition 2.3. *Suppose the $c+1$ classes have the common covariance matrix $\mathbf{I}$, and the class means $\boldsymbol{\mu}_i$ are in particular locations in $\mathbb{R}^D$ such that the between-class scatter matrix via (2.2) is given by $\mathbf{\Lambda}$. Then, $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{S}}_0$ are the corresponding estimates of $\mathbf{I}$ and $\mathbf{\Lambda}$, respectively.*

Then, the following Lemma 2.1 shows that $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$ can be expressed by using the eigenvalues and eigenvectors of $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{S}}_0$.

Lemma 2.1. *Suppose the eigendecompositions* $\widehat{\mathbf{\Sigma}}_0 = \mathbf{U}\Lambda(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T$ *and* $\widehat{\mathbf{S}}_0 = \mathbf{V}\Lambda(\widehat{\mathbf{S}}_0)\mathbf{V}^T$*, then*

$$\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_i, \tag{2.18}$$

where

$$\boldsymbol{\delta}_i = \|\mathcal{R}^T(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})\mathbf{U}^T\mathbf{e}_i\|^2. \tag{2.19}$$

Although Lemma 2.1 does not show $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$ explicitly, it provides several insight to the generalization ability of LDA:

1. *Given the population discrimination power* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*)$*, i.e.,* $\boldsymbol{\lambda}_i$*'s, the generalization discrimination power* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$ *is independent of the covariance* $\mathbf{\Sigma}$*. (3.34) and (3.35) show that* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$ *is exactly determined by* $\widehat{\mathbf{\Sigma}}_0$ *and* $\widehat{\mathbf{S}}_0$*, which are the empirical estimates of* $\mathbf{I}$ *and* $\mathbf{\Lambda}$*, respectively, and are independent of* $\mathbf{\Sigma}$*. This observation is important since it helps us get rid of the covariance structures, especially the conditional number* $\lambda_{max}(\mathbf{\Sigma})/\lambda_{min}(\mathbf{\Sigma})$*, which is an important regularity condition for learning the mixture of Gaussians Dasgupta [1999].*

2. *The eigenvalues and eigenvectors of* $\widehat{\mathbf{\Sigma}}_0$ *and* $\widehat{\mathbf{S}}_0$ *play the key roles in evaluating the generalization discrimination power* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$*. In particular, the eigenvalues of* $\widehat{\mathbf{\Sigma}}_0$ *is especially important due to the inversion operation* $\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)$ *in (3.35). The eigenvectors of* $\widehat{\mathbf{\Sigma}}_0$ *are rather uninformative, as we will see later* $\mathbf{U}$ *is a uniformly distributed random variable on the set of all orthonormal matrices. Moreover, regarding* $\widehat{\mathbf{S}}_0$*, the first* c *eigenvectors* $\mathbf{V}_{1:c}$ *also affect the generalization discrimination power* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$*.*

3. *Like the the population discrimination power* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*) = \sum_{i=1}^{c} \boldsymbol{\lambda}_i$*, the generalization discrimination power* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$ *is also expressed as a sum of* c *components, each corresponding to its counterpart in* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*)$ *but multiplied by a factor* δ_i*.*

2.3.2 Properties of the Auxiliary Estimates

We have known from Section 2.3.1 that the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$ is exactly determined by eigenvalues and eigenvectors of the auxiliary estimates $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{S}}_0$. In this section, we present some useful lemmas on properties of these eigenvalues and eigenvectors for the asymptotic generalization bound to be proved later.

2.3.2.1 Asymptotic Properties of $\widehat{\mathbf{\Sigma}}_0$

First, we have the following lemma on the eigenvalues and eigenvectors of $\widehat{\mathbf{\Sigma}}_0$.

Lemma 2.2. *Given the eigendecomposition $\widehat{\mathbf{\Sigma}}_0 = \mathbf{U}\Lambda(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T$, it holds*

1. *$\mathbf{U}$ and $\Lambda(\widehat{\mathbf{\Sigma}}_0)$ are independent random variables;*

2. *$\mathbf{U}$ follows the Haar distribution, i.e., it is uniformly distributed on the set of all orthonormal matrices in $\mathbb{R}^{D\times D}$;*

3. *denoting by $F_N(\lambda)$ the empirical spectral distribution of the eigenvalues of $\widehat{\mathbf{\Sigma}}_0$, i.e.,*

$$F_N(\lambda) = \frac{1}{D}\sum_{i=1}^{D} 1\{\lambda_i(\widehat{\mathbf{\Sigma}}_0) \le \lambda\},\ \lambda \ge 0, \tag{2.20}$$

then, as $D/N \longrightarrow \gamma \in (0,1)$,

$$F_N(\lambda) \xrightarrow{a.s.} F_\gamma(\lambda), \tag{2.21}$$

where the limit distribution $F_\gamma(\lambda)$ has the density

$$dF_\gamma(\lambda) = \frac{1}{2\pi\gamma}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}, \tag{2.22}$$

with

$$\lambda_+ = (1+\sqrt{\gamma})^2 \text{ and } \lambda_- = (1-\sqrt{\gamma})^2. \tag{2.23}$$

The first and second statements in Lemma 2.2 can be understood by the fact that $\widehat{\mathbf{\Sigma}}_0$ is the empirical estimate of $\mathbf{I}$ and its probability density function is invariant to any orthogonal transformation, while the last statement is a corollary of the well-known Marčenko-Pastur law, which says that the empirical spectral distribution of the matrix $\mathbf{A} = \frac{1}{N}\mathbf{G}\mathbf{G}^T$, wherein $\mathbf{G} \in \mathbb{R}^{D\times N}$ has i.i.d entries sampled from $\mathcal{N}(0,1)$, converges almost surely to the deterministic distribution $F_\gamma(\lambda)$ as $D/N \longrightarrow \gamma \in (0,1)$.

Further, we have the following two lemmas on $\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)$ and $\Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)$.

Lemma 2.3. *Suppose ξ is a unit-length random vector uniformly distributed on the unit sphere $\mathbb{S}^{D-1}$ and it is independent of $\widehat{\mathbf{\Sigma}}_0$, then, as $D/N \longrightarrow \gamma \in (0,1)$, it holds*

$$\xi^T \Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi \xrightarrow{a.s.} \int \lambda^{-1} dF_\gamma(\lambda) = \frac{1}{1-\gamma}. \tag{2.24}$$

Lemma 2.4. *Suppose ξ is a unit-length random vector uniformly distributed on the unit sphere $\mathbb{S}^{D-1}$ and it is independent of $\widehat{\mathbf{\Sigma}}_0$, then, as $D/N \longrightarrow \gamma \in (0,1)$, it holds*

$$\xi^T \Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\xi \xrightarrow{a.s.} \int \lambda^{-2} dF_\gamma(\lambda) = \frac{1}{(1-\gamma)^3}. \tag{2.25}$$

These two lemmas say that, in the limit, the projection of $\Lambda^{-}(\widehat{\mathbf{\Sigma}}_0)$ and $\Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)$ onto a random direction is almost surely deterministic.

2.3.2.2 Asymptotic Properties of $\widehat{\mathbf{S}}_0$

We have the following lemma on the first c eigenvectors of $\widehat{\mathbf{S}}_0$.

Lemma 2.5. *Given $\mathbf{\Lambda}$ and the eigendecomposition $\widehat{\mathbf{S}}_0 = \mathbf{V}\Lambda(\widehat{\mathbf{S}}_0)\mathbf{V}^T$, then, as $D/N \longrightarrow \gamma \in (0,1)$, it holds*

$$\lim_{D/N\longrightarrow\gamma} \|\mathbf{V}_{1:c}^T \mathbf{e}_i\|^2 \geq \frac{\boldsymbol{\lambda}_i}{\boldsymbol{\lambda}_i + \gamma}, \ \ a.s., \ \ i = 1, 2, ..., c, \tag{2.26}$$

where $\boldsymbol{\lambda}_i$ is the i-th diagonal entry of $\mathbf{\Lambda}$.

Note that the first c eigenvectors of $\mathbf{\Lambda}$ are $\mathbf{I}_{1:c} = [\mathbf{e}_1, ..., \mathbf{e}_c]$. Thus, from the relationship between $\widehat{\mathbf{S}}_0$ and $\mathbf{\Lambda}$, $\mathbf{V}_{1:c}$ is actually an estimate of $\mathbf{I}_{1:c}$. Lemma 2.5 describes the performance of this estimation in terms of $\boldsymbol{\lambda}_i$ and γ. Specifically, if $\frac{\boldsymbol{\lambda}_i}{\boldsymbol{\lambda}_i+\gamma}$ is close to 1, then $\mathbf{e}_i$ is mostly included in $\mathbf{V}_{1:c}$.

If we treat $\widehat{\mathbf{S}}_0$ as obtained by a perturbation on $\mathbf{\Lambda}$, matrix perturbation theory Stewart and Sun [1990] can be directly applied to examine the performance of $\mathbf{V}_{1:c}$ as an estimate of $\mathbf{I}_{1:c}$. However, we found that the the corresponding lower bound would be $\frac{\boldsymbol{\lambda}_i - 1}{\boldsymbol{\lambda}_i}$, which is much looser than (2.26). The superiority of our bound comes from random matrix theory, in particular, the result on the largest singular value of a random Gaussian matrix Edelman and Rao [2005].

2.3.3 Asymptotic Generalization Bound

In this section, we prove our main result, which is an asymptotic lower bound of the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$. Recall the result in Lemma 2.1, i.e., $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_i$. We first present a lower bound of $\boldsymbol{\delta}_i$.

Lemma 2.6. *Given the eigenvalues $\Lambda(\widehat{\mathbf{\Sigma}}_0)$ of $\widehat{\mathbf{\Sigma}}_0$ and the first c eigenvectors $\mathbf{V}_{1:c}$ of $\widehat{\mathbf{S}}_0$, it holds*

$$\boldsymbol{\delta}_i \geq \max{}^2\{\cos(\theta), 0\}, \tag{2.27}$$

where

$$\theta = \arccos(\|\mathbf{V}_{1:c}^T \mathbf{e}_i\|) + \arccos\left(\xi^T \Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi \Big/ \sqrt{\xi^T \Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\xi}\right), \tag{2.28}$$

with ξ a unit-length random vector uniformly distributed on the unit sphere $\mathbb{S}^{D-1}$.

Then by Lemma 2.3, Lemma 2.4, Lemma 2.5, and Lemma 2.6, we have the following theorem on the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$.

Theorem 2.1. *Suppose the population discrimination power is given by $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*) = \sum_{i=1}^{c} \boldsymbol{\lambda}_i$, and $\widehat{\mathbf{W}}^*$ is the empirical optimal projection matrix obtained by $\max \Delta(\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{S}}|\mathbf{W})$. For the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_i$, as both the dimensionality D and the training sample number N increase $(N > D)$ and $D/N \longrightarrow \gamma \in (0, 1)$, it holds almost surely*

$$\boldsymbol{\delta}_i \geq \boldsymbol{\eta}_i = \max{}^2\Big\{\cos(\arccos(\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i + \gamma)}) + \arccos(\sqrt{1-\gamma})), 0\Big\}. \tag{2.29}$$

Proof. By Lemma 2.3 and Lemma 2.4, we have

$$\lim_{D/N \longrightarrow \gamma} \frac{\xi^T \Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi}{\sqrt{\xi^T \Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\xi}} = \frac{\frac{1}{1-\gamma}}{\frac{1}{(1-\gamma)^{1.5}}} = \sqrt{1-\gamma}, \text{ a.s.}, \tag{2.30}$$

and by Lemma 2.5, we have

$$\lim_{D/N \longrightarrow \gamma} \|\mathbf{V}_{1:c}^T \mathbf{e}_i\| = \sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i + \gamma)}, \text{ a.s..} \tag{2.31}$$

Then the proof is completed by substituting (2.30) and (2.31) into Lemma 2.6 and recalling the fact $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) = \sum_{i=1}^c \boldsymbol{\delta}_i \boldsymbol{\lambda}_i$. □

From Theorem 2.1, we have the following observations:

1. *Given the population discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*)$, i.e., $\boldsymbol{\lambda}_i$'s, the lower bound of the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*)$, i.e., $\boldsymbol{\eta}_i$'s, is only determined by the dimensionality to training sample number ratio $\gamma = D/N$, which is quantitatively described by (3.23). Figure 2.1 shows the lower bound $\boldsymbol{\eta}_i$ as a function of γ and $\boldsymbol{\lambda}_i$.*

2. *The affection of $\gamma = D/N$ to the generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}^*})$ comes from two aspects, each though the term $\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i + \gamma)}$ and the term $\sqrt{1-\gamma}$. Note that the first term $\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i + \gamma)}$ allows a tradeoff between $\boldsymbol{\lambda}_i$ and γ, i.e., the affection caused by a large γ can be relatively reduced by a large $\boldsymbol{\lambda}_i$. This is consistent with the intuition that a problem with a larger population discrimination power should be easier to be dealt with by LDA. In contrast, the second term $\sqrt{1-\gamma}$ is only related to γ. According to (2.30), $\sqrt{1-\gamma}$ is due to $\Lambda(\widehat{\mathbf{\Sigma}}_0)$, i.e., the eigenvalues of the sample covariance. Actually, by assuming a sufficient large $\boldsymbol{\lambda}_i$ so that $\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i + \gamma)} \approx 1$, we have*

$$\boldsymbol{\eta}_i \approx 1 - \gamma, \tag{2.32}$$

which shows that given the dimensionality to training sample number ratio $\gamma = D/N$, the loss of discrimination power due to the imperfection of sample covariance is approximately γ. To the best of our knowledge, this is the

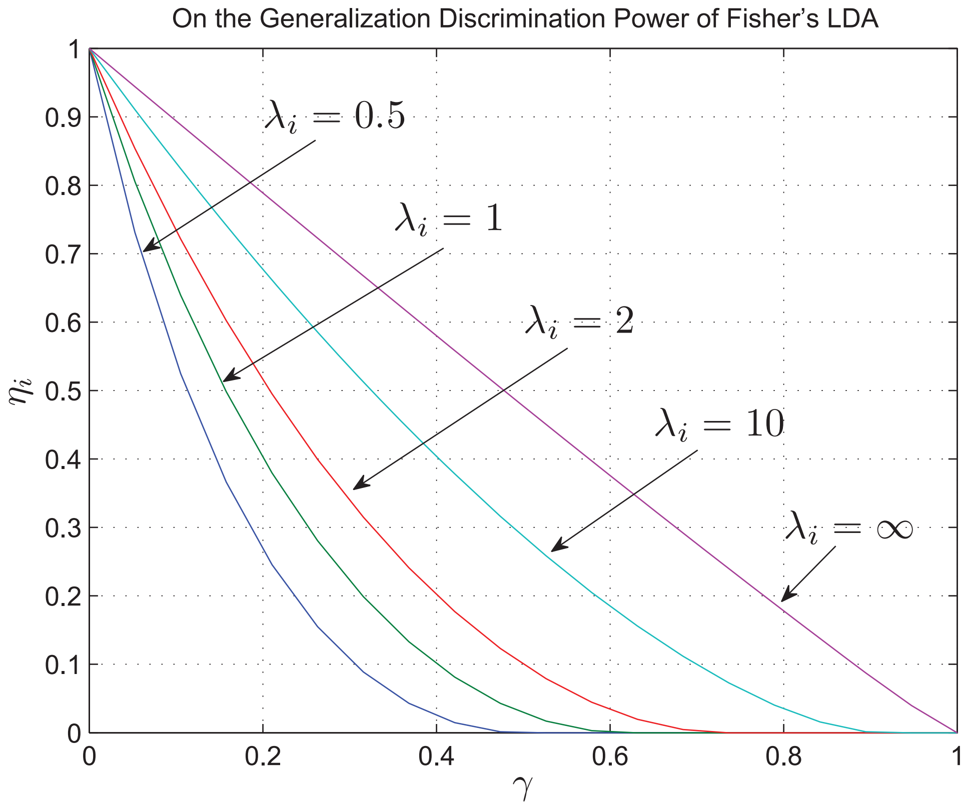


Figure 2.1: Asymptotic generalization bound of LDA.

first quantitative result on the influence of covariance estimation to LDA, although it has been commonly noticed in the literature.

3. *In the lower bound of the generalization discrimination power, each $\boldsymbol{\eta}_i$ is individually determined by the corresponding $\boldsymbol{\lambda}_i$. Since $\boldsymbol{\lambda}_i$ is sorted in a decreasing order, $\boldsymbol{\eta}_i$ is also in a decreasing order according to (3.23). This implies that, in the c components of the generalization discrimination power $\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_i$, the last few ones may not be as useful as the first few ones. We will come back to this point in Chapter 4 by considering how to further reduce dimensionality from c to lower cases.*

2.4 Empirical Evaluations

In this section, we present empirical experiments on both synthetic and real datasets to evaluate the validity of the asymptotic generalization bound obtained before. We first learn $\widehat{\mathbf{W}}^*$ by performing LDA and then compare the lower bound of the generalization discrimination power, i.e., $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) \geq \sum_{i=1}^{c} \boldsymbol{\eta}_i \boldsymbol{\lambda}_i$ from Theorem 2.1, with the true generalization discrimination power, i.e., $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_i$ from Lemma 2.1. Specifically, one can see that the comparison is actually between $\boldsymbol{\eta}_i$ and $\boldsymbol{\delta}_i$, $i = 1, 2, ..., c$.

Recall that

$$\boldsymbol{\delta}_i = \|\mathcal{R}^T(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})\mathbf{U}^T\mathbf{e}_i\|^2, \tag{2.33}$$

wherein $\Lambda(\widehat{\mathbf{\Sigma}}_0)$ and $\mathbf{U}$ are the eigenvalues and the eigenvectors of $\widehat{\mathbf{\Sigma}}_0$ while $\mathbf{V}_{1:c}$ contains the first c eigenvectors of $\widehat{\mathbf{S}}_0$. As $\widehat{\mathbf{\Sigma}}_0$ and $\widehat{\mathbf{S}}_0$ are obtained by normalizing the sample estimators $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{S}}$, respectively, with $\mathbf{X}^*$ which simultaneously diagonalizes population parameters $\mathbf{\Sigma}$ and $\mathbf{S}$, it is necessary to know $\mathbf{\Sigma}$ and $\mathbf{S}$ before hand. For the synthetic data case, we can specify these parameters. But for the real data case, they are unknown. To this end, we choose real datasets with sufficiently number of samples compared to the dimensionality, i.e., $N \gg D$, and treat the estimates with entire dataset as the "population" parameters. In addition, note that $\boldsymbol{\delta}_i = \|\mathcal{R}^T(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})\mathbf{U}^T\mathbf{e}_i\|^2$ is a random variable due to the empirical learning by LDA. Thus, we do Monte Carlo experiments to obtain the realizations of $\boldsymbol{\delta}_i$.

As for

$$\boldsymbol{\eta}_i = \max{}^2\big\{\cos(\arccos(\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i+\gamma)}) + \arccos(\sqrt{1-\gamma})), 0\big\}, \tag{2.34}$$

it is a deterministic variable related to $\boldsymbol{\lambda}_i$ and γ. We vary $\boldsymbol{\lambda}_i$ and γ so as to evaluate the asymptotic generalization bound in different situations. Besides, in order to evaluate this asymptotic result, we are supposed to vary the dimensionality D to a sufficient large case. However, by experiments, we found that $D \geq 100$ is almost sufficient for the asymptotic generalization bound to be valid.

2.4.1 On Synthetic Datasets

The evaluation in this subsection is based on synthetic datasets. From previous discussion, we know the covariance structure does not affect the generalization ability of LDA, and thus we properly choose the covariance matrix to be $\mathbf{\Sigma} = \mathbf{I}$. The between-class matrix is specifies as $\mathbf{S} = \mathrm{diag}(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_c, 0, ..., 0)$, where c is class number. Note that we can always choose the class means $\boldsymbol{\mu}_i$ such that via (2.2) they give the specified $\mathbf{S}$. Below, we design three examples with different setting of c and $\boldsymbol{\lambda}_i$, $i = 1, ..., c$. In each example, we vary D and γ to examine how these two parameters affect the generalization ability of LDA and the validity of the obtained asymptotic generalization bound.

The experiment is conducted in following steps: 1) according to the given c and $\boldsymbol{\lambda}_i$, $i = 1, ..., c$, we choose class means $\boldsymbol{\mu}_i$, $i = 1, ..., c+1$; 2) generate in total $N = D/\gamma$ samples from the $c+1$ Gaussian distributions, $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I})$, $i = 1, ..., c+1$, each with $N/(c+1)$ samples; 3) repeat step 2) 10,000 times, and for each time calculate $\boldsymbol{\delta}_i$, $i = 1, ..., c$, by (2.33); 4) calculate $\boldsymbol{\eta}_i$, $i = 1, ..., c$, by (2.34); 5) compare $\boldsymbol{\delta}_i$ and $\boldsymbol{\eta}_i$ by plotting them on one figure.

Example 1: $c = 2$, $\boldsymbol{\lambda}_1 = 1$, $D = \{10, 50, 100, 200\}$.

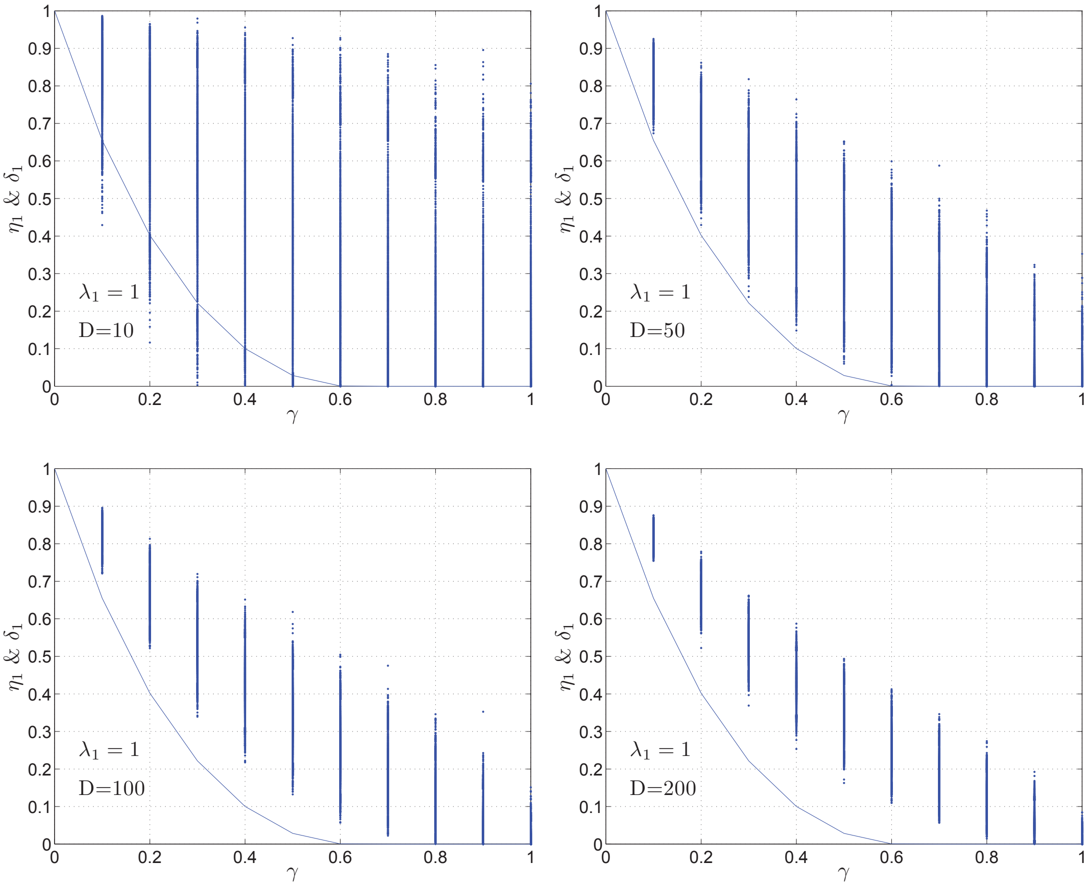


Figure 2.2: Evaluation of the asymptotic generalization bound on Example 1.

Example 2: $c = 2$, $\boldsymbol{\lambda}_1 = 10$, $D = \{10, 50, 100, 200\}$.

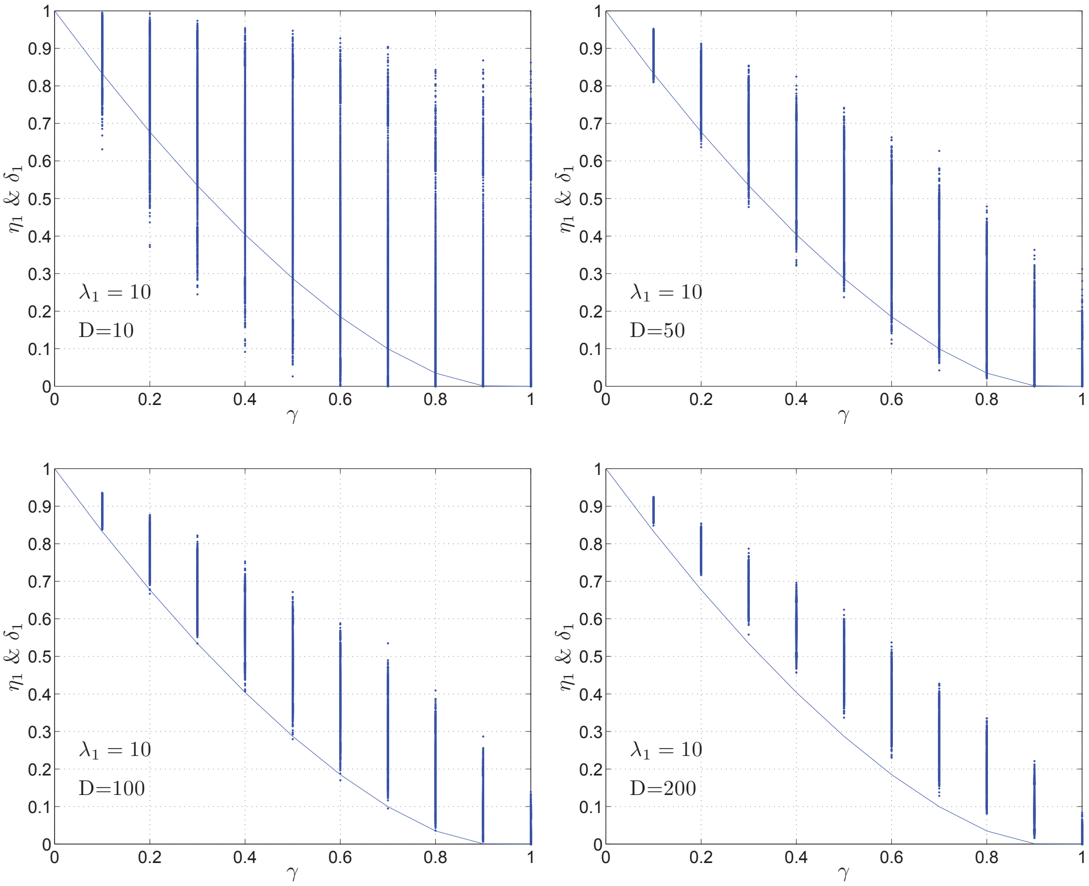


Figure 2.3: Evaluation of the asymptotic generalization bound on Example 2.

Example 3: $c = 5$, $\boldsymbol{\lambda}_1 = 10$, $\boldsymbol{\lambda}_2 = 2$, $\boldsymbol{\lambda}_3 = 1$, $\boldsymbol{\lambda}_4 = 0.5$, $D = 100$.

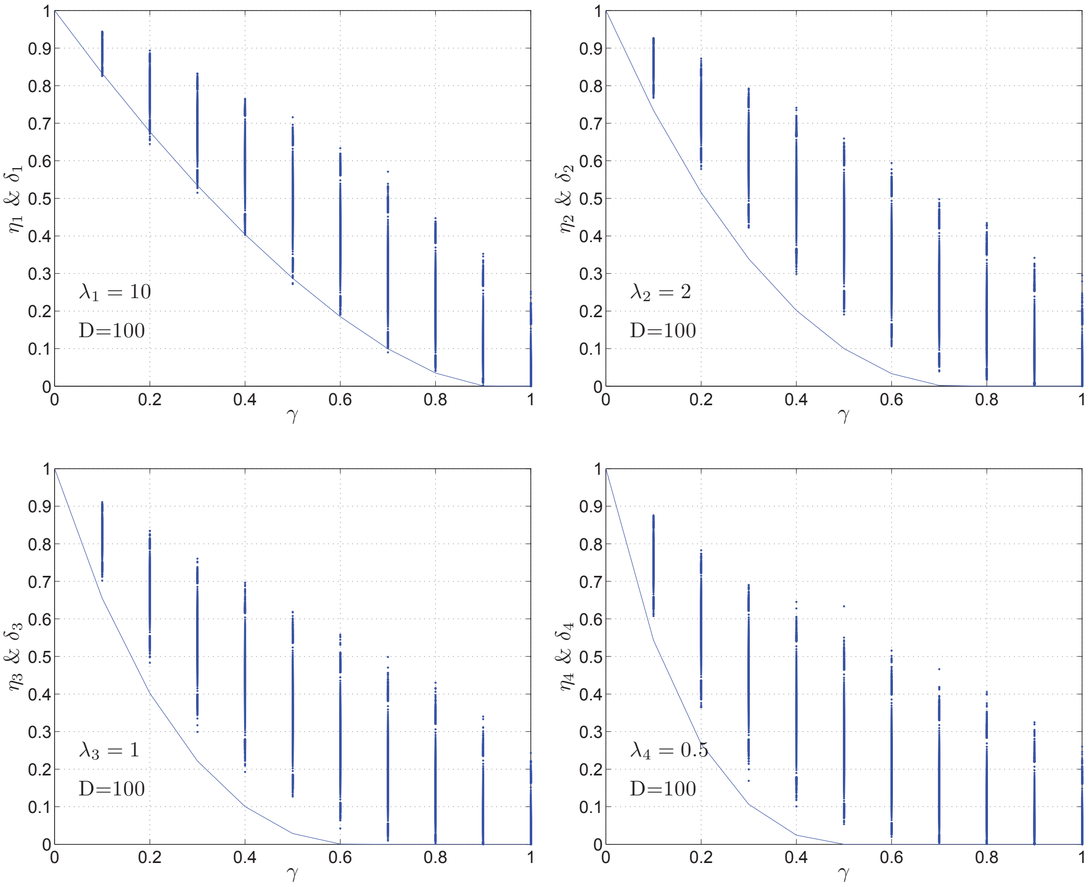


Figure 2.4: Evaluation of the asymptotic generalization bound on Example 3.

We have the following remarks on the evaluation results shown by Figures 2.2 to 2.4:

1. *Although the asymptotic generalization bound holds theoretically in the limit case, i.e., $D = \infty$, in all the three examples above $D \geq 100$ is enough for it to be valid. Recall that our asymptotic result is obtained based on asymptotic results from random matrix theory, e.g., the Marčenko-Pastur Law and the convergence of extreme singular values of a random matrix, which themselves hold satisfyingly for a moderate dimensionality.*

2. *The plots show that the bound is considerably tight: the shape of the $\boldsymbol{\eta}_i$ curve fits the scatters of $\boldsymbol{\delta}_i$ well. This is due to the deterministic character of the bound, i.e., when D is sufficient large the bound holds almost surely rather than in a probabilistic sense.*

2.4.2 On Real Datasets

This subsection presents empirical evaluations of the asymptotic generalization bound on three real datasets from the UCI machine learning repository Blake and Merz [1998]. 1) The image segmentation (ImageSeg) dataset, which contains samples randomly drawn from a database of seven outdoor images. Nineteen continuous valued features are extracted for each 3×3 region, and thus each sample is a vector from $\mathbb{R}^{19}$. The class label is obtained by manual segmentation, including brick-face, sky, foliage, cement, window, path, and grass. There are 2,310 samples and 7 classes in total. 2) The Landsat dataset, which has been used in the Statlog project, consists of the multi-spectral values of pixels in 3×3 neighborhoods in a satellite image. It constants in total 6,435 samples from 6 classes, including red soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble, and very damp grey soil, and each sample is a vector in $\mathbb{R}^{36}$. 3) The optical recognition of handwritten digits (Optdigits) dataset, which contains 5,620 samples from $\mathbb{R}^{60}$ for optical recognition of 10 handwritten digits from 0 to 9.

On each dataset, we model the classes by homoscedastic Gaussian distributions, where the classes means $\boldsymbol{\mu}_i$, between-class scatter matrix $\mathbf{S}$, and the common covariance matrix $\boldsymbol{\Sigma}$ are estimated by using the entire dataset. Note that for all the three datasets, it holds $N \gg D$, and thus we can suppose these estimations to be reliable. The procedure of experiments for the evaluation of the asymptotic generalization bound is similar to the synthetic data case: 1) we randomly select a certain number of training samples according to a given γ; 2)performance LDA and calculate $\boldsymbol{\delta}_i$ by (2.33); 3) repeat step 1) and 2) 10,000 times; 4) calculate $\boldsymbol{\eta}_i$ by (2.34); 5) compare $\boldsymbol{\delta}_i$ and $\boldsymbol{\eta}_i$ by plotting them on one figure. The results of evaluation on the three datasets are shown in Figure 2.5, 2.6 and 2.7, respectively.

On all the three datasets, our asymptotic generalization bound are valid, i.e., the scatters of $\boldsymbol{\delta}_i$ is lower bounded by the curve of $\boldsymbol{\eta}_i$. However, the tightness of the bound is not as good as in the synthetic data case. This is because the real data are located on a finite support of the data space, which cannot provide the worst-case simulation results as in the synthetic data case.

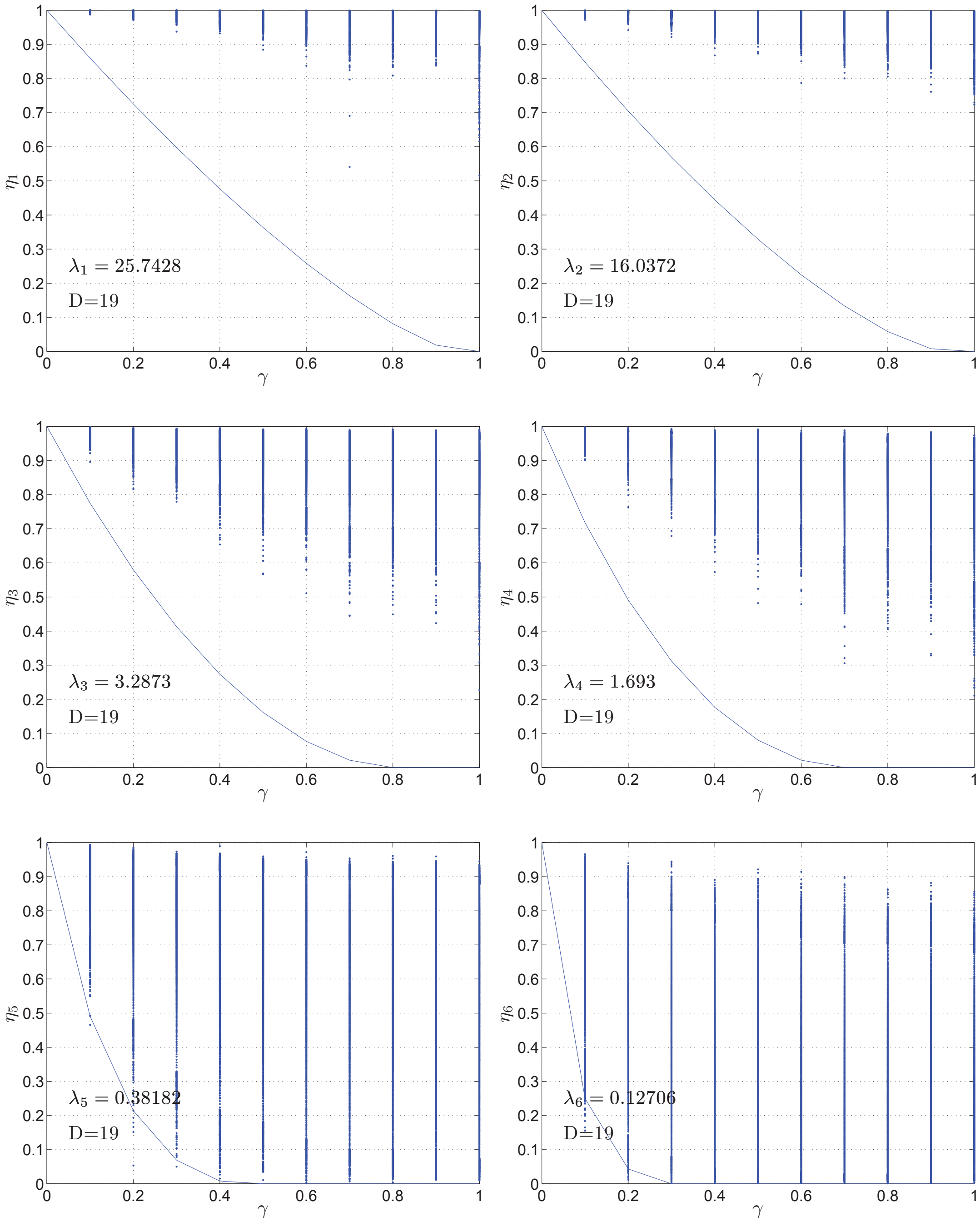


Figure 2.5: Evaluation of the asymptotic generalization bound on the ImageSeg dataset

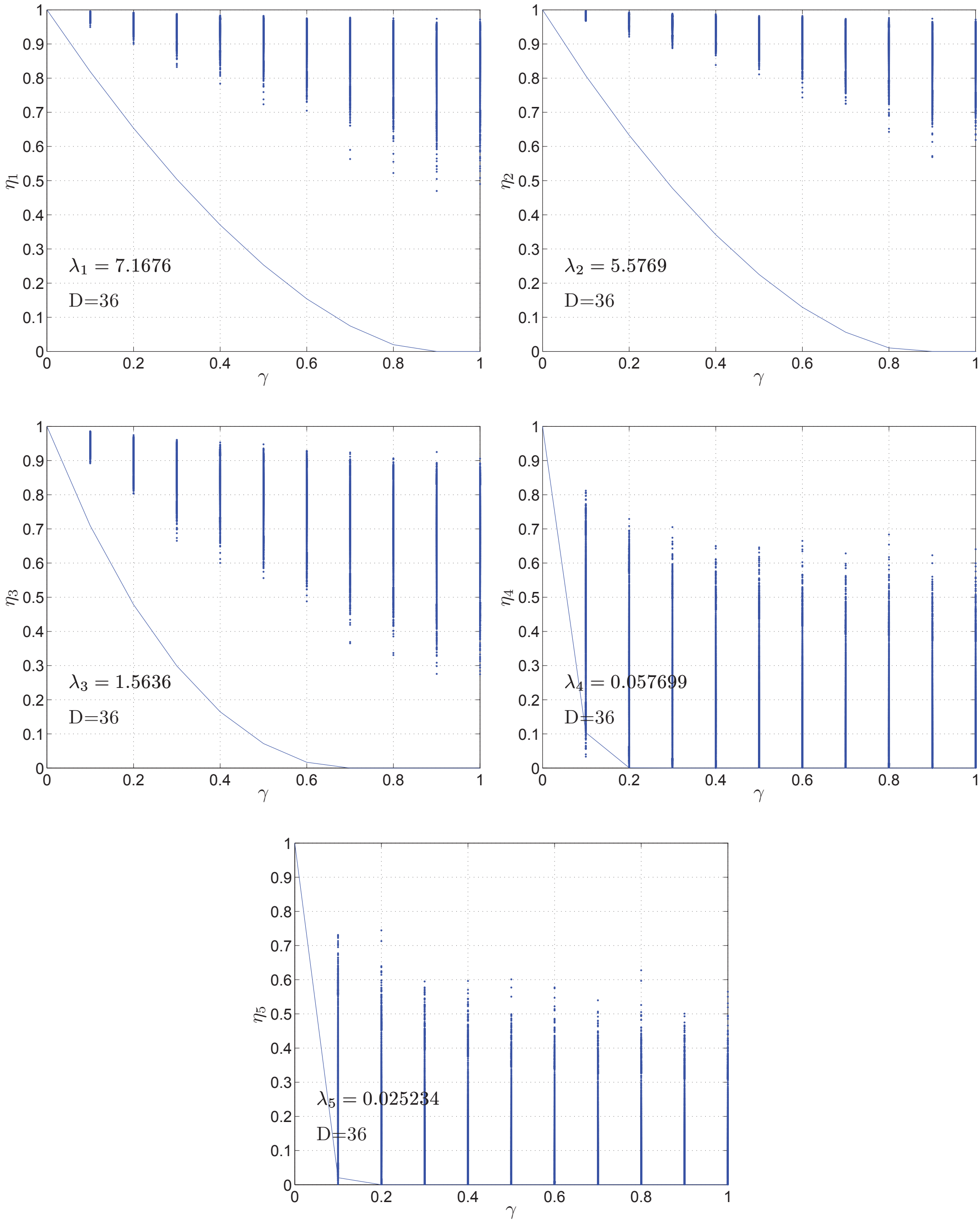


Figure 2.6: Evaluation of the asymptotic generalization bound on the ImageSeg dataset

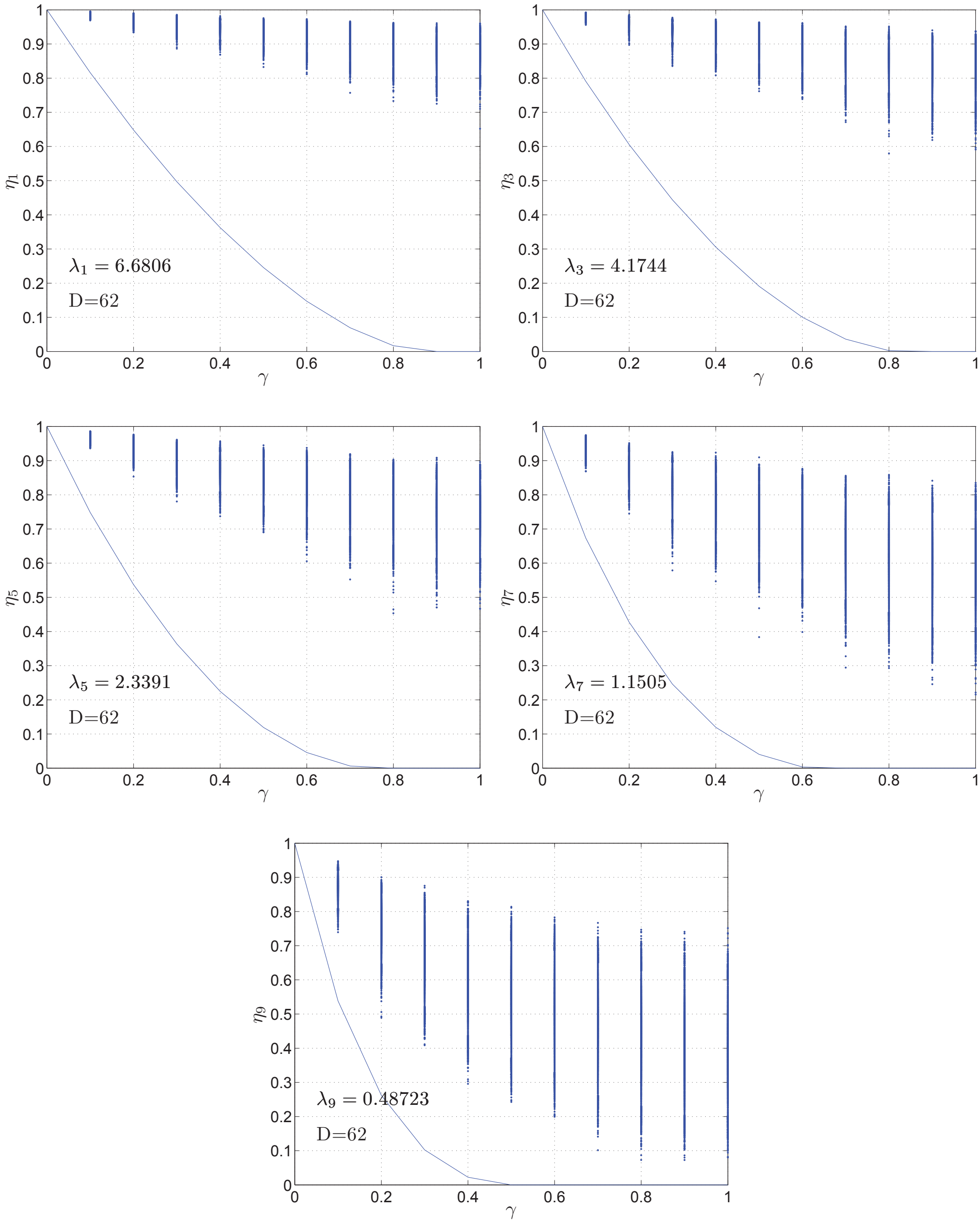


Figure 2.7: Evaluation of the asymptotic generalization bound on the OptDigits dataset

2.5 Discussions

We have proved the asymptotic generalization bound of Fisher's LDA, in the setting where both dimensionality D and training sample size N increase and $D/N \longrightarrow \gamma \in (0,1)$. In this section, we discuss possible extensions of this result in different settings.

In practice, it is possible that the class number $c+1$ increases along with D and N. The following corollary shows that the same generalization ability holds for LDA as long as the increasing speed of $c+1$ is lower than those of D and N.

Corollary 2.1. *Suppose the population discrimination power is given by* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*) = \sum_{i=1}^{c} \boldsymbol{\lambda}_i$, *and* $\widehat{\mathbf{W}}^*$ *is the empirical optimal projection matrix obtained by* $\max \Delta(\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{S}}|\mathbf{W})$. *For the generalization discrimination power* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_i$, *as both the dimensionality* D *and the training sample number* N *increase* ($N > D$) *and* $D/N \longrightarrow \gamma \in (0,1)$, *and the class number* $c+1$ *also increases but satisfying* $(c+1)/D \longrightarrow 0$, *it holds almost surely*

$$\boldsymbol{\delta}_i \geq \boldsymbol{\eta}_i = \max{}^2\Big\{\cos(\arccos(\sqrt{\boldsymbol{\lambda}_i/(\boldsymbol{\lambda}_i+\gamma)}) + \arccos(\sqrt{1-\gamma})), 0\Big\}. \tag{2.35}$$

It can be seen that Corollary 2.1 extends the result in Theorem 2.1, where the class number $c+1$ is assumed to be a fixed constant. Note that the condition $(c+1)/D \longrightarrow 0$ is essential for the validity of (2.35). It means that though the class number $c+1$ can be considerably large but it should not be comparable to the dimensionality D. This is realistic for practical problems, where data dimensionality can be tens of thousands while the class number has a lower order of magnitude.

Besides, the condition $D/N \longrightarrow \gamma \in (0,1)$ in Theorem 2.1 requires that the growing speeds of D and N are linearly comparable, i.e., $D = \gamma N + o(N)$ or $D = O(N)$. It will be interesting to consider the case where D and N may not increase in sync. The following two corollaries show the results for two settings where D grows essentially slower and faster than N, respectively.

Corollary 2.2. *Suppose the population discrimination power is given by* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*) = \sum_{i=1}^{c} \boldsymbol{\lambda}_i$, *and* $\widehat{\mathbf{W}}^*$ *is the empirical optimal projection matrix obtained by* $\max \Delta(\widehat{\mathbf{\Sigma}}, \widehat{\mathbf{S}}|\mathbf{W})$. *For the generalization discrimination power* $\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_i$, *as both* *the dimensionality D and the training sample number N increase and satisfying $D = o(N)$, it holds almost surely*

$$\boldsymbol{\delta}_i \longrightarrow 1. \tag{2.36}$$

Corollary 2.3. *Suppose the population discrimination power is given by $\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\mathbf{W}^*) = \sum_{i=1}^{c} \boldsymbol{\lambda}_i$, and $\widehat{\mathbf{W}}^*$ is the empirical optimal projection matrix obtained by $\max \Delta(\widehat{\boldsymbol{\Sigma}}, \widehat{\mathbf{S}}|\mathbf{W})$. For the generalization discrimination power $\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_i$, as both the dimensionality D and the training sample number N increase and satisfying $N = o(D)$, it only leads to trivial lower bound*

$$\boldsymbol{\delta}_i \geq 0. \tag{2.37}$$

We have the following observations from Corollaries 2.2 and 2.3. First, when D grows slower than N, i.e., $D = o(N)$, we have that $\delta_i \longrightarrow 1$. In this case, there is no loss of discrimination power asymptotically, i.e., LDA is asymptotically Bayes optimal as long as $D = o(N)$. This substantially generalizes the classical result on LDA's asymptotic Bayes optimality, which requires D being a fixed constant. Second, when D grows faster than N, i.e., $N = o(D)$, we only obtain a trivial lower bound $\delta_i \geq 0$. Such result is quite informative and one may wonder if it is possible to give more informative results on LDA' generalization ability when $N = o(D)$. Actually, Bickel and Levina [2004] gives a negative answer to this question. It has been proved that for a binary-class classification problem with equal prior probabilities, the 1-dimensional subsapce learned by LDA has classification error rate 0.5, i.e., like random guessing, if $N = o(D)$. Thus, LDA has 0 generalization ability in this situation, which implies the lower bound (2.3), though trivial, is sharp in the sense that it can be attained.

2.6 Appendixes

2.6.1 Proof of Lemma 2.1

Proof. The proof is divided into two steps.
i) Since $\mathbf{X}^*$ is nonsingular in Proposition 2.1, we can express $\widehat{\mathbf{W}}^*$ as

$$\widehat{\mathbf{W}}^* = \mathbf{X}^*\mathbf{Q}, \tag{2.38}$$

for some $\mathbf{Q} \in \mathbb{R}^{D\times c}$. Then,

$$\begin{aligned}\Delta(\boldsymbol{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}^*) &= \mathrm{Tr}((\widehat{\mathbf{W}}^{*T}\boldsymbol{\Sigma}\widehat{\mathbf{W}}^*)^{-1}\widehat{\mathbf{W}}^{*T}\mathbf{S}\widehat{\mathbf{W}}^*)\\ &= \mathrm{Tr}((\mathbf{Q}^T\mathbf{X}^{*T}\boldsymbol{\Sigma}\mathbf{X}^*\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{X}^{*T}\mathbf{S}\mathbf{X}^*\mathbf{Q})\\ &= \mathrm{Tr}((\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{X}^{*T}\boldsymbol{\Lambda}\mathbf{Q})\\ &= \mathrm{Tr}((\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}_1^T\boldsymbol{\Lambda}_1\mathbf{Q}_1)\\ &= \mathrm{Tr}(\mathbf{Q}_1(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}_1^T\boldsymbol{\Lambda}_1)\\ &= \sum_{i=1}^{c}\boldsymbol{\delta}_i\boldsymbol{\lambda}_i,\end{aligned} \tag{2.39}$$

where $\mathbf{Q}_1$ contains the first c rows of $\mathbf{Q}$ and

$$\boldsymbol{\delta}_i = \{\mathbf{Q}_1(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}_1^T\}_{ii}. \tag{2.40}$$

ii) Similar to Proposition 2.1, we can augment $\widehat{\mathbf{W}}^*$ with some $\hat{\mathbf{V}}^* \in \mathbb{R}^{D\times c}$ to simultaneously diagonalize $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{S}}$, and thus have

$$\widehat{\mathbf{W}}^{*T}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{W}}^* = \mathbf{I}_c \text{ and } \widehat{\mathbf{W}}^{*T}\widehat{\mathbf{S}}\widehat{\mathbf{W}}^* = \widehat{\boldsymbol{\Lambda}}_1, \tag{2.41}$$

where $\widehat{\boldsymbol{\Lambda}}_1$ is some $c\times c$ diagonal matrix. Then, substituting (2.38) into (2.41) and recalling $\widehat{\boldsymbol{\Sigma}}_0 = \mathbf{X}^{*T}\widehat{\boldsymbol{\Sigma}}\mathbf{X}^*$ and $\widehat{\mathbf{S}}_0 = \mathbf{X}^{*T}\widehat{\mathbf{S}}\mathbf{X}^*$, we get

$$\mathbf{Q}^T\widehat{\boldsymbol{\Sigma}}_0\mathbf{Q} = \mathbf{I}_c \text{ and } \mathbf{Q}^T\widehat{\mathbf{S}}_0\mathbf{Q} = \widehat{\boldsymbol{\Lambda}}_1. \tag{2.42}$$

Given the eigendecomposition $\widehat{\boldsymbol{\Sigma}}_0 = \mathbf{U}\Lambda(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T$, we have from the first equation in (2.42) that there must exist some orthogonal matrix $\mathbf{O} \in \mathbb{R}^{D\times c}$, $\mathbf{O}^T\mathbf{O} = \mathbf{I}_c$, such that

$$\mathbf{Q} = \mathbf{U}\Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{O}. \tag{2.43}$$

Further, given the eigendecomposition $\widehat{\mathbf{S}}_0 = \mathbf{V}^T\Lambda(\widehat{\mathbf{S}}_0)\mathbf{V}$, we get from the second equation in (2.42) that

$$\mathbf{O}^T\Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}\Lambda(\widehat{\mathbf{S}}_0)\mathbf{V}^T\mathbf{U}\Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{O} = \widehat{\boldsymbol{\Lambda}}_1. \tag{2.44}$$

In addition, since $\widehat{\mathbf{S}}_0$ has rank c, we can rewrite (2.44) as

$$\mathbf{O}^T\Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\Lambda_1^{\frac{1}{2}}(\widehat{\mathbf{S}}_0)\Lambda_1^{\frac{1}{2}}(\widehat{\mathbf{S}}_0)\mathbf{V}_{1:c}^T\mathbf{U}\Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{O} = \widehat{\boldsymbol{\Lambda}}_1, \tag{2.45}$$

where $\Lambda_1(\widehat{\boldsymbol{\Sigma}}_0)$ is the first $c \times c$ diagonal block of $\Lambda(\widehat{\boldsymbol{\Sigma}}_0)$. (2.45) implies the columns of $\mathbf{O}$ must be the left singular vectors of $\Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\Lambda_1^{\frac{1}{2}}(\widehat{\mathbf{S}}_0)$. Thus, $\mathbf{O}$ spans the range space of $\Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\Lambda_1^{\frac{1}{2}}(\widehat{\mathbf{S}}_0)$ and also the range space of $\Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}$[1]. Then, there must exist some matrix $\mathbf{A} \in \mathbb{R}^{c\times c}$ such that $\Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c} = \mathbf{OA}$, and thus

$$\mathbf{O} = \Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{A}^{-1}, \tag{2.46}$$

where the nonsingularity of $\mathbf{A}$ is implied by the nonsingularity of $\Lambda^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T$.

By (2.43) and (2.46), we have

$$\mathbf{Q} = \mathbf{U}\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{A}, \tag{2.47}$$

and

$$\mathbf{Q}_1 = \mathbf{I}_{1:c}^T\mathbf{U}\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{A}. \tag{2.48}$$

Therefore,

$$\begin{aligned}&\{\mathbf{Q}_1(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}_1\}_{ii} = \\ &\quad \mathbf{e}_i^T\mathbf{U}\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}(\mathbf{V}_{1:c}^T\mathbf{U}\Lambda^{-2}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})^{-1}\mathbf{V}_{1:c}^T\mathbf{U}\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{e}_i\end{aligned} \tag{2.49}$$

[1] It actually requires $\Lambda_1(\widehat{\mathbf{S}}_0)$ to be invertible, i.e., the first c eigenvalues of $\widehat{\mathbf{S}}_0$ are all nonzero. Note that with probability one $\widehat{\mathbf{S}}_0$ and $\mathbf{S}$ has the same rank, and thus this requirement is always satisfied as long as $\mathbf{S}$ has rank c.

Letting $\mathbf{R}$ span the range space $\mathbf{R} = \mathcal{R}(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})$, then

$$\mathbf{R}\mathbf{R}^T = \Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}(\mathbf{V}_{1:c}^T\mathbf{U}\Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})^{-1}\mathbf{V}_{1:c}^T\mathbf{U}\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0), \tag{2.50}$$

which together with (2.49) gives

$$\{\mathbf{Q}_1(\mathbf{Q}_\ell^T\mathbf{Q}_\ell)^{-1}\mathbf{Q}_1\}_{ii} = \mathbf{e}_i^T\mathbf{U}\mathbf{R}\mathbf{R}^T\mathbf{U}^T\mathbf{e}_i = \|\mathbf{R}^T\mathbf{U}^T\mathbf{e}_i\|^2. \tag{2.51}$$

This completes the proof. □

2.6.2 Proof of Lemma 2.2

We need the following propositions, which are referred to as the Marčenko-Pastur Law Marčenko and Pastur [1967] in the literature of random matrix theory. Note that Proposition 2.5 is a corollary of Proposition 2.4.

Proposition 2.4. *Marčenko and Pastur [1967] Given $\mathbf{H} \in \mathbb{R}^{D\times N}$, whose entries are independent zero-mean real (or complex) random variables with variance $1/N$ and fourth moments of order $\mathcal{O}(1/N^2)$, then as both D and $N \longrightarrow \infty$, and $D/N \longrightarrow \gamma$, the empirical distribution of the eigenvalues of $\mathbf{H}\mathbf{H}^T$ converges almost surely to a deterministic limiting distribution with density*

$$f_\gamma(\lambda) = \max(1-1/\gamma, 0)\mathbf{1}(\lambda = 0) + \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\gamma\lambda} \tag{2.52}$$

where

$$\lambda_+ = (1+\sqrt{\gamma})^2 \text{ and } \lambda_- = (1-\sqrt{\gamma})^2 \tag{2.53}$$

Proposition 2.5. *Letting $\widehat{\Sigma}$ be the sample covariance, obtained by N i.i.d. samples of the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ in $\mathbb{R}^D$, then as both D and $N \longrightarrow \infty$, and $D/N \longrightarrow \gamma \in (0,1)$, the empirical distribution of the eigenvalues of $\widehat{\Sigma}$ converges almost surely to a deterministic limiting distribution with density*

$$f_\gamma(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\gamma\lambda}, \tag{2.54}$$

where

$$\lambda_+ = (1+\sqrt{\gamma})^2 \text{ and } \lambda_- = (1-\sqrt{\gamma})^2. \tag{2.55}$$

The proof of Lemma 2.2 is provided below.

Proof. By Proposition 2.3, we have

$$\widehat{\mathbf{\Sigma}}_0 = \frac{1}{N}\sum_{i=1}^{c+1}\sum_{j=1}^{n}(\mathbf{x}_j^i - \bar{\mathbf{x}}_i)(\mathbf{x}_j^i - \bar{\mathbf{x}}_i)^T, \tag{2.56}$$

where $\mathbf{x}_j^i$ is sampled from $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I})$ and $\bar{\mathbf{x}}_i$ is the sample mean. Letting $\mathbf{z}_j^i = \mathbf{x}_j^i - \boldsymbol{\mu}_i$, which means $\mathbf{z}_j^i$ is sampled from the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$, then $\widehat{\mathbf{\Sigma}}_0$ can be rewritten as

$$\widehat{\mathbf{\Sigma}}_0 = \frac{1}{N}\sum_{i=1}^{c+1}\sum_{j=1}^{n}(\mathbf{z}_j^i - \bar{\mathbf{z}}^i)(\mathbf{z}_j^i - \bar{\mathbf{z}}^i)^T, \tag{2.57}$$

with $\bar{\mathbf{z}}^i \sim \mathcal{N}(0, \frac{1}{n}\mathbf{I})$. One property of $\widehat{\mathbf{\Sigma}}_0$ in (2.57) is that, as a random variable, its distribution is invariant to orthogonal similarity transformation, i.e., $\widehat{\mathbf{\Sigma}}_0$ and $\mathbf{U}\widehat{\mathbf{\Sigma}}_0\mathbf{U}^T$, where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ have the same distribution. This is a result of the fact that $\mathbf{O}^T\widehat{\mathbf{\Sigma}}_0\mathbf{O}$ corresponds to (2.57) in the case of replacing $\mathbf{z}_j^i$ by $\mathbf{O}\mathbf{z}_j^i$ and $\mathbf{U}\mathbf{z}_j^i$ has the same distribution with $\mathbf{z}_j^i$, i.e., the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. Then, according to Theorem 3.2 in Edelman [1989], due to the invariant property to orthogonal similarity transformation, the distribution of $\widehat{\mathbf{\Sigma}}_0$ is independent of its eigenvectors $\mathbf{U}$ but only depends on its eigenvalues $\Lambda(\widehat{\mathbf{\Sigma}}_0)$, and thus $\mathbf{U}$ should be a random variable uniformly distributed on the set of all possible orthonormal matrices. This completes the statements 1) and 2) in Lemma 2.2.

In addition, (2.57) can be rewritten as

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_0 &= \frac{1}{N}\sum_{i=1}^{c+1}\sum_{j=1}^{n}\mathbf{z}_j^i\mathbf{z}_j^{iT} - \frac{1}{c+1}\sum_{i=1}^{c+1}\bar{\mathbf{z}}^i\bar{\mathbf{z}}^{iT} \\ &= \frac{1}{N}\sum_{i=1}^{c+1}\sum_{j=1}^{n}\mathbf{z}_j^i\mathbf{z}_j^{iT} - \frac{1}{(c+1)n}\sum_{i=1}^{c+1}\sqrt{n}\bar{\mathbf{z}}^i\sqrt{n}\bar{\mathbf{z}}^{iT} \\ &= \frac{1}{N}\mathbf{G}_1\mathbf{G}_1^T - \frac{1}{N}\mathbf{G}_2\mathbf{G}_2^T \\ &= T_1 + T_2.\end{aligned} \tag{2.58}$$

where $\mathbf{G}_1 \in \mathbb{R}^{D\times N}$, $\mathbf{G}_2 \in \mathbb{R}^{D\times(c+1)}$, and both have entries i.i.d. from $\mathcal{N}(0,1)$. For the first term $T_1 = \frac{1}{N}\mathbf{G}_1\mathbf{G}_1^T$, by Proposition 2.4, we know that the empirical distribution of its eigenvalues converges almost surely to $F_\gamma(\gamma)$ with density,

$$f_\gamma(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\gamma\lambda}, \tag{2.59}$$

where $\gamma = D/N$ and

$$\lambda_+ = (1+\sqrt{\gamma})^2 \text{ and } \lambda_- = (1-\sqrt{\gamma})^2. \tag{2.60}$$

For the second term $T_2 = \frac{1}{N}\mathbf{G}_2\mathbf{G}_2^T$, clearly it has finite rank $c+1$. According to Tao [2012], a finite rank perturbation does not effect the convergence of the empirical spectral distribution, i.e., $\lim F_N(\lambda(T_1 + T2)) = \lim F_N(\lambda(T_1)) = F_\gamma(\lambda)$. This completes the proof. □

2.6.3 Proof of Lemma 2.3

Proof. The condition that ξ is a unit-length random vector uniformly distributed on the unit sphere $\mathbb{S}^{D-1}$ can be replaced by $\xi \in \mathbb{R}^D$ and its entries are i.i.d. samples from $\mathcal{N}(0, 1/D)$. This is because, in the later case, $\xi/\|\xi\|$ is uniformly distributed on $\mathbb{S}^{D-1}$, and in the limit $\|\xi\|^2 \xrightarrow{a.s.} 1$ due to the strong law of large numbers. Then, we divide the proof into two steps. First, we show that $\xi^T\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi \xrightarrow{a.s.} \int \lambda^{-1} dF_\gamma(\lambda)$ and then we calculate the integral.

i) Recall $\lambda_- = (1-\sqrt{\gamma})^2$, and let $\overline{\Lambda}^{-1}(\widehat{\mathbf{\Sigma}}_0) = \mathrm{diag}(\min\{\lambda_-, \lambda_i^{-1}(\widehat{\mathbf{\Sigma}}_0)\})$, i.e., a truncated version of $\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)$ by clamping $\lambda_i^{-1}(\widehat{\boldsymbol{\Sigma}}_0)$ to be λ_-^{-1} if $\lambda_i^{-1}(\widehat{\boldsymbol{\Sigma}}_0) \geq \lambda_-^{-1}$. Then, we divide the lefthand side of (2.24) into three terms

$$\xi^T \Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi - \xi^T \overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi, \tag{2.61}$$

$$\xi^T \overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi - \frac{1}{D}\mathrm{Tr}(\overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0)), \tag{2.62}$$

and

$$\frac{1}{D}\mathrm{Tr}(\overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0)) - \int \lambda^{-1} dF_\gamma(\lambda). \tag{2.63}$$

Then, we show that all the three terms almost surely converge to zero.

For the first term, we have

$$\begin{aligned} 0 \leq & \xi^T(\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0) - \overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0))\xi \\ \leq & \|\xi\|^2 \max\{0, \lambda_{\min}^{-1}(\widehat{\boldsymbol{\Sigma}}_0) - \lambda_-^{-1}\}. \end{aligned} \tag{2.64}$$

Since $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_0) \xrightarrow{a.s.} \lambda_-$ Edelman and Rao [2005], we have

$$\max\{0, \lambda_{\min}^{-1}(\widehat{\boldsymbol{\Sigma}}_0) - \lambda_-^{-1}\} \xrightarrow{a.s.} 0. \tag{2.65}$$

Then, by $\|\xi\|^2 \xrightarrow{a.s.} 1$, (2.64) and (2.65), we have

$$\xi^T \Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi - \xi^T \overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi \xrightarrow{a.s.} 0. \tag{2.66}$$

For the second term, since $\|\overline{\Lambda}^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\| \leq \lambda_-$ for all D, i.e., it is uniformly bounded, then we apply Theorem 3.4 in Tulino and Verdú [2004] and get

$$\xi^T \overline{\Lambda}_\alpha^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi - \frac{1}{D}\mathrm{Tr}(\overline{\Lambda}_\alpha^{-1}(\widehat{\boldsymbol{\Sigma}}_0)) \xrightarrow{a.s.} 0. \tag{2.67}$$

For the third term, since $dF_\gamma(\lambda)$ is nonzero on the $[\lambda_-, \lambda_+]$, it is sufficient to examine

$$\begin{aligned}
&\frac{1}{D}\mathrm{Tr}(\overline{\Lambda}^{-1}(\widehat{\Sigma}_0)) - \int \lambda^{-1} dF_\gamma(\lambda) \\
&= \int_0^\infty \min(\lambda_-, \lambda^{-1}) dF_N(\lambda) - \int_{\lambda_-}^{\lambda_+} \lambda^{-1} dF_\gamma(\lambda) \\
&= \int_{\lambda_-}^{\lambda_+} \lambda^{-1} d(F_N(\lambda) - F_\gamma(\lambda)) + \lambda_-^{-1} \int_0^{\lambda_-} dF_N(\lambda) + \int_{\lambda_+}^\infty \lambda^{-1} dF_N(\lambda).
\end{aligned} \tag{2.68}$$

Sine $F_N(\lambda) \xrightarrow{a.s.} F_\gamma(\lambda)$ and λ^{-1} is bounded on $[\lambda_-, \lambda_+]$, it holds Billingsley [1999]

$$\int_{\lambda_-}^{\lambda_+} \lambda^{-1} d(F_N(\lambda) - F_\gamma(\lambda)) =\!\!\xrightarrow{a.s.} 0. \tag{2.69}$$

Further, sine $F_\gamma(\lambda_-) = 0$ and $F_\gamma(\lambda_+) = 1$, it holds

$$\int_0^{\lambda_-} dF_N(\lambda) = F_N(\lambda_-) \xrightarrow{a.s.} F_\gamma(\lambda_-) = 0, \tag{2.70}$$

and

$$0 \leq \int_{\lambda_+}^\infty \lambda^{-1} dF_N(\lambda) \leq \lambda_+^{-1}(1 - F_N(\lambda_+)) \xrightarrow{a.s.} \lambda_+^{-1}(1 - F_\gamma(\lambda_+)) = 0. \tag{2.71}$$

Thus,

$$\frac{1}{D}\mathrm{Tr}(\overline{\Lambda}_\alpha^{-1}(\widehat{\Sigma}_0)) - \int \lambda^{-1} dF_\gamma(\lambda) \xrightarrow{a.s.} 0. \tag{2.72}$$

ii) We now calculate the integral

$$I = \int \lambda^{-1} dF_\gamma(\lambda) = \int_{\lambda_-}^{\lambda_+} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\gamma\lambda^2} d\lambda \tag{2.73}$$

where $\lambda_+ = (1 + \sqrt{\gamma})^2$ and $\lambda_- = (1 - \sqrt{\gamma})^2$.

Letting $\lambda = 1 + \gamma - 2\sqrt{\gamma}\cos x$, $x \in [0, \pi]$ and substituting it into (2.73), we have

$$I = \frac{2}{\pi} \int_0^\pi \frac{\sin^2 x}{(1 + \gamma - 2\sqrt{\gamma}\cos x)^2} dx. \tag{2.74}$$

Further, letting $t = \tan\frac{x}{2}$, we have

$$\begin{aligned} I &= \frac{2}{\pi}\int_0^\infty \frac{\left(\frac{2t}{1+t^2}\right)^2}{\left(1+\gamma-2\sqrt{\gamma}\frac{1-t^2}{1+t^2}\right)^2}\frac{2}{1+t^2}dt \\ &= \frac{16}{\pi}\int_0^\infty \frac{t^2}{\left((1+\gamma)(t^2+1)-2\sqrt{\gamma}(1-t^2)\right)^2}\frac{1}{1+t^2}dt \\ &= \frac{16}{\pi}\int_0^\infty \frac{t^2}{\left((1+\sqrt{\gamma})^2t^2+(1-\sqrt{\gamma})^2\right)^2}\frac{1}{1+t^2}dt \\ &= \frac{16}{\pi(1+\sqrt{\gamma})^4}\int_0^\infty \frac{t^2}{\left(t^2+\left(\frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}\right)^2\right)^2}\frac{1}{1+t^2}dt. \end{aligned} \tag{2.75}$$

Letting $\alpha = \frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}$ and by partial fraction, we have

$$\begin{aligned} \int_0^\infty \frac{t^2}{(t^2+\alpha^2)^2}\frac{1}{1+t^2}dt = &\int_0^\infty \frac{-\frac{1}{(1-\alpha^2)^2}}{t^2+1}dt \\ &+\int_0^\infty \frac{\frac{1}{(1-\alpha^2)^2}}{t^2+\alpha^2}dt + \int_0^\infty \frac{-\frac{\alpha^2}{(1-\alpha^2)}}{(t^2+\alpha^2)^2}dt. \end{aligned} \tag{2.76}$$

Denoting by I_1, I_2 and I_3 the terms in the righthand side of (2.76), we have

$$I_1 = \int_0^\infty \frac{-\frac{1}{(1-\alpha^2)^2}}{t^2+1}dt = \frac{-1}{(1-\alpha^2)^2}\int_0^\infty d\arctan t = \frac{-\pi}{2(1-\alpha^2)^2}, \tag{2.77}$$

$$I_2 = \int_0^\infty \frac{\frac{1}{(1-\alpha^2)^2}}{t^2+\alpha^2}dt = \frac{1}{\alpha(1-\alpha^2)^2}\int_0^\infty d\arctan\frac{t}{\alpha} = \frac{\pi}{2\alpha(1-\alpha^2)^2}, \tag{2.78}$$

$$\begin{aligned} I_3 &= \int_0^\infty \frac{-\frac{\alpha^2}{(1-\alpha^2)}}{(t^2+\alpha^2)^2}dt \\ &= \frac{-1}{2(1-\alpha^2)}\int_0^\infty d\frac{t}{t^2+\alpha^2} + \frac{-1}{2(1-\alpha^2)}\int_0^\infty \frac{1}{t^2+\alpha^2}dt \\ &= 0 + \frac{-\pi}{4\alpha(1-\alpha^2)} = \frac{-\pi}{4\alpha(1-\alpha^2)}. \end{aligned} \tag{2.79}$$

Combining (2.75) to (2.79) and noticing $\alpha = \frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}$, we get

$$\begin{aligned} I &= \frac{16}{\pi(1+\sqrt{\gamma})^4}\left(\frac{-\pi}{2(1-\alpha^2)^2} + \frac{\pi}{2\alpha(1-\alpha^2)^2} + \frac{-\pi}{4\alpha(1-\alpha^2)}\right) \\ &= \frac{16}{\pi(1+\sqrt{\gamma})^4}\frac{\pi}{4\alpha(1+\alpha)^2} \\ &= \frac{1}{1-\gamma}. \end{aligned} \tag{2.80}$$

This completes the proof. □

2.6.4 Proof of Lemma 2.4

Proof. By the same strategy as used in the proof of Lemma 2.3, we have $\xi^T\Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\xi \xrightarrow{a.s.} \int \lambda^{-2}dF_\gamma(\lambda)$. Below, we calculate the integral.

$$I = \int \lambda^{-2}dF_\gamma(\lambda) = \int_{\lambda_-}^{\lambda_+}\frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{2\pi\gamma\lambda^3}d\lambda, \tag{2.81}$$

where $\lambda_+ = (1+\sqrt{\gamma})^2$ and $\lambda_- = (1-\sqrt{\gamma})^2$. Letting $\lambda = 1+\gamma-2\sqrt{\gamma}\cos x$, $x \in [0,\pi]$ and substituting it into (2.73), we have

$$I = \frac{2}{\pi}\int_0^\pi \frac{\sin^2 x}{(1+\gamma-2\sqrt{\gamma}\cos x)^3}dx. \tag{2.82}$$

Further, letting $t = \tan\frac{x}{2}$, we have

$$\begin{aligned} I &= \frac{2}{\pi}\int_0^\infty \frac{\left(\frac{2t}{1+t^2}\right)^2}{\left(1+\gamma-2\sqrt{\gamma}\frac{1-t^2}{1+t^2}\right)^3}\frac{2}{1+t^2}dt \\ &= \frac{16}{\pi}\int_0^\infty \frac{t^2}{\left((1+\gamma)(t^2+1)-2\sqrt{\gamma}(1-t^2)\right)^3}dt \\ &= \frac{16}{\pi}\int_0^\infty \frac{t^2}{\left((1+\sqrt{\gamma})^2t^2+(1-\sqrt{\gamma})^2\right)^3}dt \\ &= \frac{16}{\pi(1+\sqrt{\gamma})^6}\int_0^\infty \frac{t^2}{\left(t^2+\left(\frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}\right)^2\right)^3}dt. \end{aligned} \tag{2.83}$$

Letting $\alpha = \frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}$, we have

$$\begin{aligned}\int_0^\infty \frac{t^2}{(t^2+\alpha^2)^3}dt =& -\frac{1}{4}\int_0^\infty d\frac{t}{(t^2+\alpha^2)^2} + \frac{1}{4}\int_0^\infty \frac{1}{(t^2+\alpha^2)^2}dt \\ =& \frac{\pi}{16\alpha^3}.\end{aligned} \tag{2.84}$$

Thus, by $\alpha = \frac{1-\sqrt{\gamma}}{1+\sqrt{\gamma}}$, we get

$$I = \frac{16}{\pi(1+\sqrt{\gamma})^6}\frac{\pi}{16\alpha^3} = \frac{1}{(1-\gamma)^3}. \tag{2.85}$$

This completes the proof. □

2.6.5 Proof of Lemma 2.5

We need the following Proposition, which describes the operator norm of a large Gaussian random matrix Edelman and Rao [2005].

Proposition 2.6. *Letting* $\mathbf{G} \in \mathbb{R}^{D\times m}$ *with i.i.d. entries sampled from* $\mathcal{N}(0,1)$*, then as* $m/D \longrightarrow \gamma \in [0,1]$*,*

$$\frac{1}{\sqrt{D}}\|\mathbf{G}\| \xrightarrow{a.s.} 1+\sqrt{\gamma}. \tag{2.86}$$

The proof of Lemma 2.5 is provided below.

Proof. Since $\mathbf{\Lambda}$ is a between-class scatter matrix, it can be expressed as $\mathbf{\Lambda} = \frac{1}{c+1}\sum_{i=1}^{c+1}(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$, where $\boldsymbol{\mu} = \frac{1}{c+1}\sum_{i=1}^{c+1}\boldsymbol{\mu}_i$. Letting $\mathbf{M} = [\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_{c+1}]$ and $\mathbf{E} \in \mathbb{R}^{(c+1)\times(c+1)}$ with all entries equal to $\frac{1}{c+1}$, we have $\mathbf{\Lambda} = \frac{1}{c+1}\mathbf{M}(\mathbf{I}-\mathbf{E})(\mathbf{I}-\mathbf{E})^T\mathbf{M}^T$. Similarly, we have $\widehat{\mathbf{S}}_0 = \frac{1}{c+1}\widehat{\mathbf{M}}(\mathbf{I}-\mathbf{E})(\mathbf{I}-\mathbf{E})^T\widehat{\mathbf{M}}^T$, with $\widehat{\mathbf{M}} = [\widehat{\boldsymbol{\mu}}_1, ..., \widehat{\boldsymbol{\mu}}_{c+1}]$. According to Proposition 2.3, the population covariance matrix is $\mathbf{I}$, and thus we have $\widehat{\mathbf{M}} = \mathbf{M} + \mathbf{X}$, wherein the entries of $\mathbf{X} \in \mathbb{R}^{D\times(c+1)}$ are i.i.d. samples from $\mathcal{N}(0, 1/n)$, with n being the training sample number for each class.

Note that the nonzero diagonal entries of $\mathbf{\Lambda}$ are $\boldsymbol{\lambda}_i$, $i = 1, 2, ..., c$, and its eigenvectors are $\mathbf{e}_i$, $i = 1, 2, ..., c$. Then, $\mathbf{\Lambda} = \frac{1}{c+1}\mathbf{M}(\mathbf{I}-\mathbf{E})(\mathbf{I}-\mathbf{E})^T\mathbf{M}^T$ implies that $\mathbf{M}(\mathbf{I}-\mathbf{E})$ has singular values $\sqrt{(c+1)\boldsymbol{\lambda}_i}$, $i = 1, 2, ..., c$ and $\mathbf{I}_{1:c} = [\mathbf{e}_1, ..., \mathbf{e}_c]$

are the corresponding left singular vectors. Thus, denoting by $\mathbf{Q} \in \mathbb{R}^{(c+1)\times c}$ the right singular vectors, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_c$, we have

$$\mathbf{M}(\mathbf{I}-\mathbf{E})\mathbf{Q} = \left[\sqrt{(c+1)\boldsymbol{\lambda}_1}\mathbf{e}_1, ..., \sqrt{(c+1)\boldsymbol{\lambda}_c}\mathbf{e}_c\right]. \tag{2.87}$$

Consequently, by $\widehat{\mathbf{M}} = \mathbf{M} + \mathbf{X}$, we have

$$\begin{aligned}&\widehat{\mathbf{M}}(\mathbf{I}-\mathbf{E})\mathbf{Q}\\ &= \left[\sqrt{(c+1)\boldsymbol{\lambda}_1}\mathbf{e}_1, ..., \sqrt{(c+1)\boldsymbol{\lambda}_c}\mathbf{e}_c\right] + \mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}\\ &= [\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_c],\end{aligned} \tag{2.88}$$

where $\boldsymbol{\xi}_i = \sqrt{(c+1)\boldsymbol{\lambda}_i}\mathbf{e}_i + \mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i$, $i = 1, 2, ..., c$. Then, by $\widehat{\mathbf{S}}_0 = \frac{1}{c+1}\widehat{\mathbf{M}}(\mathbf{I}-\mathbf{E})(\mathbf{I}-\mathbf{E})^T\widehat{\mathbf{M}}^T$, we have for the first c eigenvectors $\mathbf{V}_{1:c}$ of $\widehat{\mathbf{S}}_0$ that

$$\begin{aligned}\mathbf{V}_{1:c} &= \mathcal{R}(\widehat{\mathbf{M}}(\mathbf{I}-\mathbf{E}))\\ &= \mathcal{R}(\widehat{\mathbf{M}}(\mathbf{I}-\mathbf{E})\mathbf{Q})\\ &= \mathcal{R}([\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_c]).\end{aligned} \tag{2.89}$$

Thus,

$$\begin{aligned}\|\mathbf{V}_{1:c}^T\mathbf{e}_i\| &= \|\mathcal{R}^T([\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_c])\mathbf{e}_i\|\\ &\geq \|\mathcal{R}^T(\boldsymbol{\xi}_i)\mathbf{e}_i\|\\ &= \frac{1}{\|\boldsymbol{\xi}_i\|}|\boldsymbol{\xi}_i^T\mathbf{e}_i|\\ &= \frac{|\mathbf{e}_i^T\sqrt{(c+1)\boldsymbol{\lambda}_i}\mathbf{e}_i + \mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i|}{\|\sqrt{(c+1)\boldsymbol{\lambda}_i}\mathbf{e}_i + \mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\|}\\ &\geq \frac{\sqrt{(c+1)\boldsymbol{\lambda}_i} - |\mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i|}{\sqrt{(c+1)\boldsymbol{\lambda}_i} + \|\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\|}.\end{aligned} \tag{2.90}$$

It can be verified that as $N = (c+1)n \longrightarrow \infty$

$$|\mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i| \leq \|\mathbf{e}_i^T\mathbf{X}\| = \sqrt{\sum_{j=1}^{c+1}\mathbf{X}_{ij}^2} \xrightarrow{a.s.} 0, \tag{2.91}$$

where the inequality is due to $\|(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\| \leq \|(\mathbf{I}-\mathbf{E})\|\|\mathbf{Q}_i\| \leq 1$ and the limit is because $\mathbf{X}_{ij}$ follows the distribution $\mathcal{N}(0, \frac{1}{n})$.

In addition, by Proposition 2.6 and letting $\mathbf{G} = \sqrt{n}\mathbf{X}$, we have

$$\|\mathbf{X}\| = \frac{1}{\sqrt{n}}\|\mathbf{G}\| \xrightarrow{a.s.} \sqrt{\frac{D}{n}} = \sqrt{\frac{(c+1)D}{N}} \longrightarrow \sqrt{(c+1)\gamma}. \tag{2.92}$$

Thus,

$$\|\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\| \le \|\mathbf{X}\| \xrightarrow{a.s.} \sqrt{(c+1)\gamma}. \tag{2.93}$$

Combining (2.107), (2.91) and (2.109), we obtain

$$\lim_{D/N\longrightarrow\gamma} \|\mathbf{V}_{1:c}^T\mathbf{e}_i\|^2 \ge \frac{\boldsymbol{\lambda}_i}{\boldsymbol{\lambda}_i+\gamma},\ a.s. \tag{2.94}$$

This completes the proof. □

2.6.6 Proof of Lemma 2.6

Proof. Recall Lemma 2.1 that $\boldsymbol{\delta}_i = \|\mathcal{R}^T(\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})\mathbf{U}^T\mathbf{e}_i\|^2$. Denote by $\measuredangle(\mathbf{U}^T\mathbf{e}_i, \mathcal{R}(\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}))$ the angle between vector $\mathbf{U}^T\mathbf{e}_i$ and subspace $\mathcal{R}^T(\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})$, we have

$$\boldsymbol{\delta}_i = \cos^2(\measuredangle(\mathbf{U}^T\mathbf{e}_i, \mathcal{R}(\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}))). \tag{2.95}$$

Two basic facts that hold for arbitrary vector $\mathbf{x}_1$, $\mathbf{x}_2$ and subspace $\mathbf{X}$ are

$$\measuredangle(\mathbf{x}_1, \mathbf{X}) \le \measuredangle(\mathbf{x}_1, \mathbf{x}_2) + \measuredangle(\mathbf{x}_2, \mathbf{X}). \tag{2.96}$$

and

$$\measuredangle(\mathbf{x}_1, \mathbf{X}) \le \measuredangle(\mathbf{x}_1, \mathbf{x}),\ \text{if } \mathbf{x} \in \mathbf{X}. \tag{2.97}$$

Then, by using (2.96) and (2.97), we get

$$\begin{aligned}
&\measuredangle(\mathbf{U}^T\mathbf{e}_i, \mathcal{R}(\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_i))\\
\le&\measuredangle(\mathbf{U}^T\mathbf{e}_i, \mathbf{U}^T\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T\mathbf{e}_i) + \measuredangle(\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T\mathbf{e}_i, \mathcal{R}(\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}))\\
\le&\measuredangle(\mathbf{U}^T\mathbf{e}_i, \mathbf{U}^T\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T\mathbf{e}_i) + \measuredangle(\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T\mathbf{e}_i, \Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T\mathbf{e}_i)\\
=&\theta_1 + \theta_2.
\end{aligned} \tag{2.98}$$

Denoting $\theta = \theta_1 + \theta_2$, since $\cos(x)$ is positive and decreasing on $[0, \pi/2]$, x^2 is increasing on $[0, 1]$, and $\boldsymbol{\delta}_i$ is nonnegative, we have

$$\begin{aligned} \boldsymbol{\delta}_i &\geq \begin{cases} \cos^2(\theta), & \theta \leq \frac{\pi}{2} \\ 0, & \text{else} \end{cases} \\ &= \max{}^2\{\cos(\theta), 0\}. \end{aligned} \tag{2.99}$$

It remains to calculate θ_1 and θ_2. For θ_1, We have

$$\cos^2(\theta_1) = \frac{|\mathbf{e}_i \mathbf{V}_{1:c}^T \mathbf{U}\mathbf{U}^T \mathbf{V}_{1:c}\mathbf{e}_i|^2}{\|\mathbf{U}^T \mathbf{V}_{1:c}\mathbf{V}_{1:c}^T \mathbf{e}_i\|^2} = \frac{|\mathbf{e}_i^T \mathbf{V}_{1:c}\mathbf{V}_{1:c}^T \mathbf{e}_i|^2}{\mathbf{e}_i^T \mathbf{V}_{1:c}\mathbf{V}_{1:c}^T \mathbf{e}_i} = \|\mathbf{V}_{1:c}^T \mathbf{e}_i\|^2, \tag{2.100}$$

which gives

$$\theta_1 = \arccos(\|\mathbf{V}_{1:c}^T \mathbf{e}_i\|). \tag{2.101}$$

For θ_2, as rescaling does not change the direction of a vector, we can rewrite θ_2 as

$$\theta_2 = \measuredangle(\mathbf{U}^T \xi, \Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T \xi), \tag{2.102}$$

where

$$\zeta = \frac{\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T \mathbf{e}_i}{\|\mathbf{V}_{1:c}\mathbf{V}_{1:c}^T \mathbf{e}_i\|}. \tag{2.103}$$

Note that ζ is a unit-length random vector and is independent of $\mathbf{U}$ due to the independency between $\mathbf{V}_{1:c}$ and $\mathbf{U}$. Then, we have

$$\cos^2(\theta_2) = \frac{|\zeta^T \mathbf{U}\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T \zeta|^2}{\|\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T \zeta\|^2} = \frac{(\zeta^T \mathbf{U}\Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T \zeta)^2}{\zeta^T \mathbf{U}\Lambda^{-2}(\widehat{\boldsymbol{\Sigma}}_0)\mathbf{U}^T \zeta}. \tag{2.104}$$

We have known, from Lemma 2.2, $\mathbf{U}$ is uniformly distributed on the set of all orthonormal matrices in $\mathbb{R}^{D\times D}$, and ζ is a unit-length random vector independent of $\mathbf{U}$. Thus, $\xi = \mathbf{U}^T \zeta$ must be a unit-length random vector uniformly distributed on the unit sphere $\mathbb{S}^{D-1}$. Finally, (2.104) gives

$$\theta_2 = \arccos\left(\xi^T \Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi \Big/ \sqrt{\xi^T \Lambda^{-2}(\widehat{\boldsymbol{\Sigma}}_0)\xi}\right). \tag{2.105}$$

This completes the proof. □

2.6.7 Proof of Corollary 2.1

Proof. It is sufficient to show that the results of Lemma 2.2 and Lemma 2.5 still hold under the conditional $(c+1)/D \longrightarrow 0$.

First, in the proof of Lemma 2.2, we have shown in (2.58) that

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_0 &= \frac{1}{N}\mathbf{G}_1\mathbf{G}_1^T - \frac{1}{N}\mathbf{G}_2\mathbf{G}_2^T \\ &= T_1 + T_2.\end{aligned} \tag{2.106}$$

where $\mathbf{G}_1 \in \mathbb{R}^{D\times N}$, $\mathbf{G}_2 \in \mathbb{R}^{D\times(c+1)}$, and both have entries i.i.d. from $\mathcal{N}(0,1)$. Under the condition $(c+1)/D \longrightarrow 0$, we have $\frac{1}{D}\text{rank}(T_2) \longrightarrow 0$, which is a sufficient condition Tao [2012] for that, as long as $F_N(\lambda(T_1))$ converges almost surely to a deterministic distribution $F_\gamma(\lambda)$, $F_N(\lambda(T_1+T_2))$ will also converge almost surely to the same distribution. Thus, the result of Lemma 2.2 still holds.

Second, in the proof of Lemma 2.5, we have shown in (2.107) that

$$\begin{aligned}\|\mathbf{V}_{1:c}^T\mathbf{e}_i\| &\geq \frac{\sqrt{(c+1)\boldsymbol{\lambda}_i} - |\mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i|}{\sqrt{(c+1)\boldsymbol{\lambda}_i} + \|\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\|} \\ &= \frac{\boldsymbol{\lambda}_i - \frac{|\mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i|}{\sqrt{c+1}}}{\boldsymbol{\lambda}_i + \frac{\|\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\|}{\sqrt{c+1}}}.\end{aligned} \tag{2.107}$$

Similar to (2.91), we have

$$\frac{|\mathbf{e}_i^T\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i|}{\sqrt{c+1}} \leq \frac{\|\mathbf{e}_i^T\mathbf{X}\|}{\sqrt{c+1}} = \sqrt{\frac{\sum_{j=1}^{c+1}\mathbf{X}_{ij}^2}{c+1}} \xrightarrow{a.s.} 0, \tag{2.108}$$

since $\mathbf{X}_{ij} \sim \mathcal{N}(0, 1/n)$. Further, by (2.109), we have

$$\frac{\|\mathbf{X}(\mathbf{I}-\mathbf{E})\mathbf{Q}_i\|}{\sqrt{c+1}} \leq \frac{\|\mathbf{X}\|}{\sqrt{c+1}} \xrightarrow{a.s.} \gamma. \tag{2.109}$$

Therefore,

$$\lim_{D/N\longrightarrow\gamma} \|\mathbf{V}_{1:c}^T\mathbf{e}_i\|^2 \geq \frac{\boldsymbol{\lambda}_i}{\boldsymbol{\lambda}_i+\gamma},\ a.s. \tag{2.110}$$

i.e., the result of Lemma 2.2 also holds. □

2.6.8 Proof of Corollary 2.2

Proof. When $D = o(N)$, we have $D/N \longrightarrow 0$, i.e, $\gamma = 0$. By calculation, we have $\boldsymbol{\eta}_i = 1$ in Theorem 2.1, and thus $\boldsymbol{\delta}_i \geq 1$, a.s.. Besides, since it holds always $0 \leq \boldsymbol{\delta}_i \leq 1$, we $\boldsymbol{\delta}_i \longrightarrow 1$, a.s.. □

2.6.9 Proof of Corollary 2.3

Proof. The proof is trivial since it holds always $0 \leq \boldsymbol{\delta}_i \leq 1$. □

Chapter 3

Block-Diagonal Regularization for Linear Discriminant Analysis

3.1 Introduction

We have known, from the asymptotic generalization analysis in Chapter 2, that in order to obtain an acceptable generalization discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{W}\widehat{\mathbf{W}}^*)$, LDA needs a rationally small dimensionality to training sample number ratio $\gamma = D/N$. Practically, however, such requirement is not always satisfied. For example, in face recognition Belhumeur et al. [1997], data usually have a high dimensionality D (e.g., thousands), but the training sample number N can be relatively small (e.g., hundreds). In such case, the sample covariance $\widehat{\mathbf{\Sigma}}$ significantly deviates from its population counterpart $\mathbf{\Sigma}$ and becomes singular if $D > N$, and consequently the direct using of LDA is readily to fail.

Regularization provides a principal method for parameter estimation from insufficient training samples. In the case of LDA, it has been shown that replacing the sample covariance with regularized estimators can significantly improve the performance Ye et al. [2006] Guo et al. [2007] Lu et al. [2005]. The most commonly used regularization method in LDA is the Tikhonov regularization, which replaces the sample covariance $\widehat{\mathbf{\Sigma}}$ in Fisher's criterion by $\widehat{\mathbf{\Sigma}} + \rho\mathbf{I}$, where $\rho > 0$ serves as a tuning parameter. Analogous to the Tikhonov regularized LDA, one can introduce other types of regularized covariance estimators to LDA.

In the literature of covariance estimation, a number of regularized estimation methods have been proposed, e.g., the factor model based estimation Lv [2007], the banded estimation Bickel and Levina [2008b], the thresholding estimation Bickel and Levina [2008a], and the sparse inverse covariance matrix estimation (also known as covariance selection) Dempster [1972] Friedman et al. [2008], to name a few. However, derived from a covariance estimation point of view, these regularization methods are designed on purpose for LDA, and to the best of our knowledge, they have not been applied to LDA by any studies in the literature.

Motivated by the results in Chapter 2, we propose a new regularization method for LDA, which is referred to as the block-diagonal regularization. According to the asymptotic generalization bound in Theorem 2.1, one efficient way to improve the generalization ability of LDA is to reduce the dimension to training sample number ratio D/N. Variable partitioning provides a simple method to reduce the ratio D/N, e.g., if we partition the variables into k groups with equal size, then the ratio will be reduced by k times into $D/(kN)$. In the block-diagonal regularization, we first partition the variables into k groups and treat the groups independently in covariance estimation. Therefore, instead of using $\widehat{\mathbf{\Sigma}}$, the block-diagonally regularized LDA uses $\widehat{\mathbf{\Sigma}}_r = \text{diag}(\widehat{\mathbf{\Sigma}}_1, ..., \widehat{\mathbf{\Sigma}}_k)$, where $\widehat{\mathbf{\Sigma}}_j$ is the sample covariance of the j-th group.

We present theoretical justification of the proposed block-diagonally regularized LDA, by examining its approximation and sample errors. Specifically, we show that the approximation error is bounded by using the closeness between the population covariance matrix and its block-diagonal approximation, and the sample error is bounded in terms of D, N, and k. We further propose two intuitive methods for variable partitioning, where the first method is based on Laplacian eigenmaps embedding and the second method uses an entirely random strategy. Empirical evaluations show that the proposed block-diagonally regularized LDA performs competitively compared with other types of regularized LDA, e.g., with the Tikhonov regularization and the banded regularization.

In addition to regularization, there have been other methods in the literature to extend LDA to deal with the insufficient training sample problem. The first class of extensions is usually referred to as two-stage LDA, which first applies an intermediate method, e.g., PCA (on the sample covariance) Belhumeur et al.

[1997] or QR decomposition (on the centralized class mean matrix) Ye and Li [2005], to reduce the dimensionality to a proper case and then perform LDA afterwards. Another class of extensions utilizes the pseudoinverse method to deal with the singularity problem occurred when the training sample number is less than the dimensionality Ye et al. [2004] Ye [2005]. Besides, null space LDA Chen et al. [2000] and dual space LDA Wang and Tang [2004] were also proposed to deal with the singularity problem, where the former only utilizes the discriminative information of the null space of the sample covariance but the latter utilizes the discriminative information of both the principal and null spaces of the sample covariance. The rest of this chapter is organized as follows. Section 3.2 presents the block-diagonally regularized LDA, including the analysis on the approximation and sample errors, as well as two intuitive methods for variable partitioning. Section 3.3 reports empirical evaluations and the comparison with other regularization methods. Technical proofs are arranged as appendixes in Section 3.4.

3.2 Block-Diagonally Regularized LDA

Suppose the D variables are divided into k groups, each with a covariance matrix $\mathbf{\Sigma}_j$, $j = 1, 2, ..., k$. Then, we refer to

$$\mathbf{\Sigma}_r = \text{diag}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, ..., \mathbf{\Sigma}_k). \tag{3.1}$$

as a block-diagonal approximation of the original covariance $\mathbf{\Sigma}$. We introduce the following measurement

$$\varrho = \|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma} - \mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\|. \tag{3.2}$$

to measure the closeness between $\mathbf{\Sigma}_r$ and $\mathbf{\Sigma}$. Since $\mathbf{\Sigma}$ is nonsingular, $\varrho = 0$ leads to $\mathbf{\Sigma} = \mathbf{\Sigma}_r$. Through this section, we assume $\varrho < 1$, which serves as a regularity condition for theoretical analysis.

Empirically, an estimate of $\mathbf{\Sigma}_r$ is given by

$$\widehat{\mathbf{\Sigma}}_r = \mathrm{diag}(\widehat{\mathbf{\Sigma}}_1, \widehat{\mathbf{\Sigma}}_2, ..., \widehat{\mathbf{\Sigma}}_k). \tag{3.3}$$

where each $\widehat{\mathbf{\Sigma}}_j$, $j = 1, 2, ..., k$, is the corresponding sample covariance of the j-th block $\mathbf{\Sigma}_j$. By submitting $\widehat{\mathbf{\Sigma}}_r$ into Fisher's criterion, we get the block-diagonally regularized LDA,

$$\widehat{\mathbf{W}}_r^* = \arg\max_{\mathbf{W}} \Delta(\widehat{\mathbf{\Sigma}}_r, \widehat{\mathbf{S}}|\mathbf{W}), \tag{3.4}$$

where $\widehat{\mathbf{S}}$ is the unchanged sample between-class scatter matrix.

The performance of the block-diagonally regularized LDA (3.4) can be justified by using the following error

$$\mathcal{E} = \Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*) - \Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}_r^*), \tag{3.5}$$

which measures the loss of discrimination power with respect to the population optimal projection matrix $\mathbf{W}^*$. In order to understand how the block-diagonal regularization works in improving the performance of LDA, we further express the error $\mathcal{E}$ as a summation of the approximation error $\mathcal{E}_a$, which is independent of the training samples, and the sample error $\mathcal{E}_s$, which depends on the training samples. To this end, we need the following projection matrix,

$$\mathbf{W}_r^* = \arg\max_{\mathbf{W}} \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}). \tag{3.6}$$

Note that $\mathbf{W}_r^*$ is also independent of training samples. Then, by direct calculation, we have

$$\begin{aligned}
\mathcal{E} =& \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}_r^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\widehat{\mathbf{W}}_r^*) \\
& + \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\widehat{\mathbf{W}}_r^*) - \Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}_r^*) + \Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}_r^*) \\
\leq& \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}_r^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\widehat{\mathbf{W}}_r^*) \\
& + \Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}_r^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\widehat{\mathbf{W}}_r^*) + \Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}^*) \qquad (3.7) \\
\leq& \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}_r^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\widehat{\mathbf{W}}_r^*) \\
& + \left|\Delta(\mathbf{\Sigma}, \mathbf{S}|\widehat{\mathbf{W}}_r^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\widehat{\mathbf{W}}_r^*)\right| + \left|\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}^*)\right| \\
\leq& \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}_r^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\widehat{\mathbf{W}}_r^*) + 2\max_{\mathbf{W}} \left|\Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W})\right|.
\end{aligned}$$

Then, we define the $\mathcal{E}_a$ and $\mathcal{E}_s$ as below,

$$\mathcal{E}_a = 2\max_{\mathbf{W}} \left|\Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}) - \Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W})\right|, \qquad (3.8)$$

$$\mathcal{E}_s = \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}_r^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\widehat{\mathbf{W}}_r^*). \qquad (3.9)$$

In the following two subsections, we will derive upper bounds of the approximation error $\mathcal{E}_a$ and the sample error $\mathcal{E}_s$, respectively.

3.2.1 On the Approximation Error

According to (3.8), the approximation error $\mathcal{E}_a$ comes from the difference between $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_r$. Specifically, we have the following Theorem on $\mathcal{E}_a$.

Theorem 3.1. *Given the measurement $\varrho = \|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma} - \mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\| < 1$, the approximation error $\mathcal{E}_a$ of the block-diagonally regularized LDA is bounded by*

$$\mathcal{E}_a \leq \frac{2\varrho}{1-\varrho}\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*). \qquad (3.10)$$

Theorem 3.1 indicates the approximation error $\mathcal{E}_a$ is bounded by a factor $\frac{2\varrho}{1-\varrho}$ with respect to the population discrimination power $\Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*)$. Thus, whenever $\mathbf{\Sigma}_r$ gives a good approximation of $\mathbf{\Sigma}$, we can expect a reasonable small approximation error $\mathcal{E}_a$.

3.2.2 On the Sample Error

In this subsection, we examine the sample error $\mathcal{E}_s$ of the block-diagonally regularized LDA. For the convenience of analysis, we assume the sizes of the k groups of variables are equal, i.e. each has a size $D' = D/k$. Recall that

$$\mathcal{E}_s = \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}_r^*) - \Delta(\mathbf{\Sigma}_r, \mathbf{S}|\widehat{\mathbf{W}}_r^*). \tag{3.11}$$

We intend to derive an asymptotic low bound of $\mathcal{E}_s$, with a similar strategy as used in the asymptotic generalization analysis in Chapter 2. Again, we begin with the simultaneous diagonalization proposition.

Proposition 3.1. *There exists a nonsingular matrix $\mathbf{X}_r^* = [\mathbf{W}_r^* \ \mathbf{V}_r^*]$, with $\mathbf{W}_r^* \in \mathbb{R}^{D\times c}$ and $\mathbf{V}_r^* \in \mathbb{R}^{D\times(D-c)}$, that simultaneously diagonalizes $\mathbf{\Sigma}_r$ and $\mathbf{S}$, i.e.,*

$$\mathbf{X}_r^{*T}\mathbf{\Sigma}_r\mathbf{X}_r^* = \mathbf{I} \textit{ and } \mathbf{X}_r^{*T}\mathbf{S}\mathbf{X}_r^* = \mathbf{\Lambda}_r, \tag{3.12}$$

where $\mathbf{\Lambda}_r$ is a diagonal matrix, with only the first c diagonal entries being nonzero. Further, $\mathbf{X}_r$ can be explicitly expressed as

$$\mathbf{X}_r^* = \mathbf{\Sigma}_r^{-\frac{1}{2}}\mathbf{U}_r^*, \tag{3.13}$$

where $\mathbf{U}_r^*$ is from the eigendecomposition $\mathbf{\Sigma}_r^{-\frac{1}{2}}\mathbf{S}\mathbf{\Sigma}_r^{-\frac{1}{2}} = \mathbf{U}_r^*\mathbf{\Lambda}_r\mathbf{U}_r^{*T}$; and,

$$\Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}_r^*) = \sum_{i=1}^{c} \boldsymbol{\lambda}_{ri}, \tag{3.14}$$

where $\boldsymbol{\lambda}_{ri}$, $i = 1, 2, ..., c$, is the first c diagonal entries of $\mathbf{\Lambda}_r$.

Then, we introduce the auxiliary estimates,

$$\widehat{\mathbf{\Sigma}}_0 = \mathbf{X}_r^{*T}\widehat{\mathbf{\Sigma}}_r\mathbf{X}_r^* \text{ and } \widehat{\mathbf{S}}_0 = \mathbf{X}_r^{*T}\widehat{\mathbf{S}}\mathbf{X}_r^*. \tag{3.15}$$

Note that $\widehat{\mathbf{\Sigma}}_0$ is more structured here than in Chapter 2. Specifically, by substituting (3.13) into (3.15), we have

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_0 &= \mathbf{U}_r^{*T}\mathbf{\Sigma}_r^{-\frac{1}{2}}\widehat{\mathbf{\Sigma}}_r\mathbf{\Sigma}_r^{-\frac{1}{2}}\mathbf{U}_r^* \\ &= \mathbf{U}_r^{*T}\mathrm{diag}(\mathbf{\Sigma}_{r1}^{-\frac{1}{2}}\widehat{\mathbf{\Sigma}}_{r1}\mathbf{\Sigma}_{r1}^{-\frac{1}{2}}, \mathbf{\Sigma}_{r2}^{-\frac{1}{2}}\widehat{\mathbf{\Sigma}}_{r2}\mathbf{\Sigma}_{r2}^{-\frac{1}{2}}, ..., \mathbf{\Sigma}_{rk}^{-\frac{1}{2}}\widehat{\mathbf{\Sigma}}_{rk}\mathbf{\Sigma}_{rk}^{-\frac{1}{2}})\mathbf{U}_r^* \\ &= \mathbf{U}_r^{*T}\mathrm{diag}(\widehat{\mathbf{\Sigma}}_{01}, \widehat{\mathbf{\Sigma}}_{02}, ..., \widehat{\mathbf{\Sigma}}_{0k},)\mathbf{U}_r^*.\end{aligned} \tag{3.16}$$

where $\widehat{\mathbf{\Sigma}}_{0j}$, $j = 1, 2, ..., k$, is the empirical estimate of the identity covariance matrix $\mathbf{I}_{D'}$.

By applying Lemma 2.2 to each individual $\widehat{\mathbf{\Sigma}}_{0j}$, we have the following lemma on the eigenvalues and the eigenvectors of $\widehat{\mathbf{\Sigma}}_0$.

Lemma 3.1. *Given the eigendecomposition $\widehat{\mathbf{\Sigma}}_0 = \mathbf{U}\Lambda(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T$, then*

$$\mathbf{U} = \mathbf{U}_r^{*T}\mathrm{diag}(\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_k), \tag{3.17}$$

$$\Lambda(\widehat{\mathbf{\Sigma}}_0) = \mathrm{diag}(\Lambda(\widehat{\mathbf{\Sigma}}_{01}), \Lambda(\widehat{\mathbf{\Sigma}}_{02})..., \Lambda(\widehat{\mathbf{\Sigma}}_{0k})), \tag{3.18}$$

where

1. *$\mathbf{U}_j$ and $\Lambda(\widehat{\mathbf{\Sigma}}_{0j})$ are independent, $j = 1, 2, ..., k$;*

2. *$\mathbf{U}_j$, $j = 1, 2, ..., k$, is uniformly distributed on the set of all orthonormal matrices in $\mathbf{R}^{D'\times D'}$, i.e., follows the Haar distribution;*

3. *denoting the empirical spectral distribution of the eigenvalues of $\widehat{\mathbf{\Sigma}}_{0j}$, $j = 1, 2, ..., k$, by $F_N(\lambda)$, then as $D'/N \longrightarrow \gamma' \in (0, 1)$*

$$F_N(\lambda) \xrightarrow{a.s.} F_{\gamma'}(\lambda), \tag{3.19}$$

where

$$dF_{\gamma'}(\lambda) = \frac{1}{2\pi\gamma'}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}, \tag{3.20}$$

with

$$\lambda_+ = (1 + \sqrt{\gamma'})^2 \text{ and } \lambda_- = (1 - \sqrt{\gamma'})^2. \tag{3.21}$$

From Lemma 3.1, one can see that asymptotically the eigenvalues of $\widehat{\mathbf{\Sigma}}_0$ are spread on $[(1 - \sqrt{\gamma'})^2, (1 + \sqrt{\gamma'})^2]$. Since $\gamma' = \gamma/k$, this is more concentrated around 1 than the original case without block-diagonal regularization, where the eigenvalue are spread on $[(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]$.

Analogous to Theorem 2.1, we have the following theorem on the sample error of the block-diagonally regularized LDA.

Theorem 3.2. *Given $\mathbf{\Sigma}_r = diag(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, ..., \mathbf{\Sigma}_k)$, with an equal block size $D' = D/k$, then as both the dimensionality D and the training sample size N increase, such that $D/N \longrightarrow \gamma \in (0, \infty)$ and $\gamma/k \in (0,1)$, the sample error $\mathcal{E}_s$ of the block-regularized LDA satisfies almost surely*

$$\mathcal{E}_s \leq \sum_{i=1}^{c} (1-\boldsymbol{\eta}_i)\boldsymbol{\lambda}_{ri}, \tag{3.22}$$

where

$$\boldsymbol{\eta}_i = \max{}^2\big\{\cos(\arccos(\sqrt{\boldsymbol{\lambda}_{ri}/(\boldsymbol{\lambda}_{ri}+\gamma)}) + \arccos(\sqrt{1-\gamma/k})), 0\big\}. \tag{3.23}$$

Comparing Theorem 2.1 and Theorem 3.2, one can see that the term $\sqrt{1-\gamma}$ is replaced by $\sqrt{1-\gamma/k}$ due to the block-diagonal regularization. This helps reduce the sample error $\mathcal{E}_s$. Especially, in the extreme case where $\boldsymbol{\lambda}_{ri}$ is sufficient large and $\sqrt{\boldsymbol{\lambda}_{ri}/(\boldsymbol{\lambda}_{ri}+\gamma)} \approx 1$, then we have $1 - \boldsymbol{\eta}_i = \gamma/k$. This implies that smaller $\mathcal{E}_s$ can be obtained by increasing the group number k. An illustration of Theorem 3.2 is given in Figure 3.1, where we set $\boldsymbol{\lambda}_i = 50$ and vary k from 1 to 20. The plots indicate that, when k increases a larger γ can be tolerated. Therefore, given a sufficient large k, even when the dimensionality D is several times of the training sample number N, we can still maintain an acceptable small sample error.

3.2.3 Intuitive Variable Partitioning

To perform the block-diagonally regularized LDA, we need to partition the variables into k groups. At first sight, it is a variable clustering problem, i.e., according to the correlation between variables we cluster them into k approximately independent groups. However, this is unrealistic, because in the situation of insufficient training samples we cannot obtain sufficiently accurate variable cor-

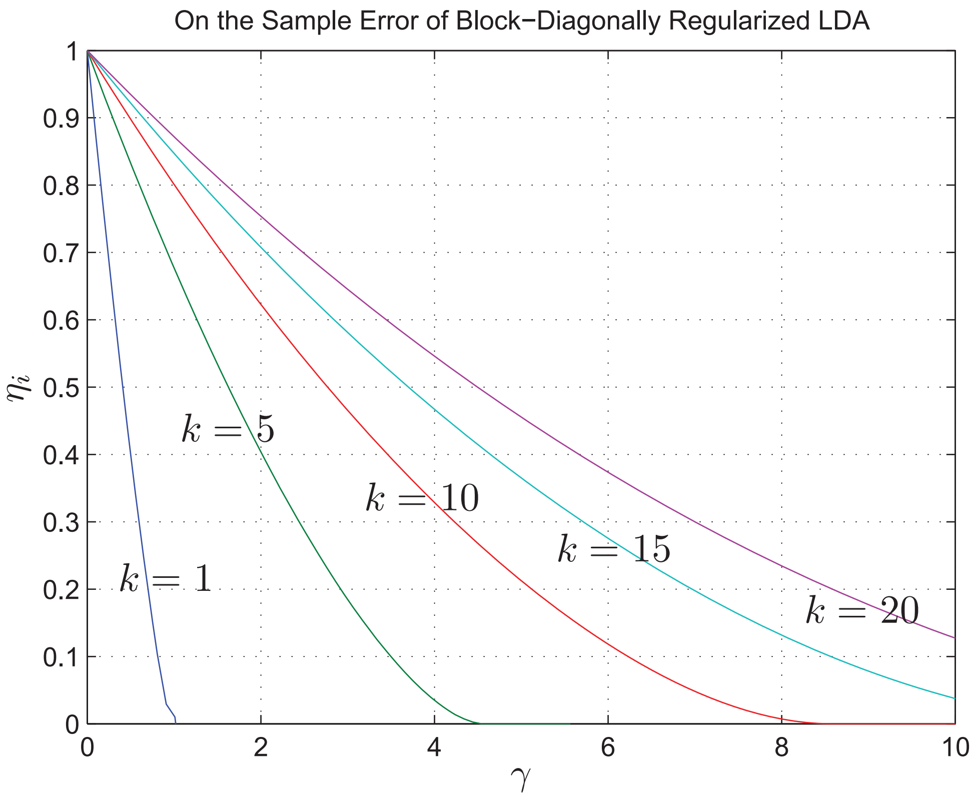


Figure 3.1: The sample error of the block-diagonally regularized LDA.

relations to perform variable clustering. Instead, we propose two intuitive methods for variable partitioning, both of which are considerably simple but perform favorably as shown by empirical evaluations.

The first method uses a deterministic partitioning method. Suppose the correlation coefficients matrix is $\mathbf{C}$. We treat the D variables as vertexes of a graph, and define the adjacent matrix $\mathbf{A}$ as below

$$\mathbf{A}_{ij} = |\mathbf{C}_{ij}|^{\alpha}, \tag{3.24}$$

where $\alpha > 1$ is a penalization parameter, i.e., A large α penalizes the small (absolute) correlation coefficients. Then, we utilize Laplacian eigenmaps Belkin and Niyogi [2003] to get a 1-dimensional embedding ξ of the D variables, i.e., ξ is a vector in $\mathbb{R}^D$. Specifically, the embedding ξ is an eigenvector of $\mathbf{D}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{A})\mathbf{D}^{-\frac{1}{2}}$ associated with the smallest nonzero eigenvalues, wherein $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$. We sort the variables according to their corresponding values in ξ in an ascending order. Finally, we partition the sorted variables evenly into k groups. One may consider using spectral clustering (also based on $\mathbf{A}$) to partition the variables into groups. However, empirical studies show that it is inferior to the above described method, mainly because spectral clustering does not guarantees the groups to be evenly small, and the large groups may lead to an large sample error.

The second method uses a random partitioning method, i.e., we randomly partition the D variables into k groups with a equal size. Clearly, the second method is considerably simple and does not utilize any information of the covariance or the correlation coefficients obtained from the training samples. However, as shown later by empirical evaluations, the random partitioning method works comparably with the deterministic method. This is consistent with our intuition that when the training sample number is small the structural information from covariance or correlation estimation can be highly inaccurate. Therefore, the entirely randomized method can have comparable performance with the estimation based method.

3.3 Empirical Evaluations

3.3.1 On Variable Partitioning

We evaluate the proposed block-diagonally regularized LDA, associated with the two methods for variable partitioning, on face recognition experiments. Four benchmark face image databases, “Feret”, “Orl”, “Pie”, and “Yale”, are used in our experiments. The Feret database contains 13,539 face images from 1,565 individuals, with varying pose, facial expression and age. In our experiments, we used a subset of it, containing 50 individuals with 7 images for each. The Orl dataset Orl [1994] contains 400 images of 40 individuals.The images are taken at different times, varying the lighting, facial expressions and facial details. The Pie database from CMU contains 68 individuals and 41,368 face images in total with varying pose, illumination condition and expression. We used a subset of PIE according to He et al. [2005], which contains all 68 subjects with 170 images for each. The YALE database contains face images collected from 15 individuals, each of which has 11 images with varying lighting condition and facial expression. The data dimensionality for Feret, Orl, and Yale, are 1,600, i.e., each face image has a size of 40×40, while the data dimensionality for Pie is 1,024, i.e., each face image has a size of 32×32. On Feret, Orl and Yale, we randomly select 80% data for training and use the rest 20% for test, while on Pie we randomly select 10% data for training and use the rest 90% for test.

The classification performance on the four datasets over 20 random training/test splits are shown by Figure 3.4 to Figure 3.5. From these results, we have the following observations:

1. *As the block number k increases until reaching an optimal k^*, the classification error rate decreases, and afterwards the classification error rate increases again. This reflects the tradeoff between the approximation and the sample errors. First, the sample error is dominated. Increasing k reduces the sampler error and thus leads to a lower classification error rate. When the sample error is small enough, the approximation error becomes dominated. Further increasing k increases the approximation and thus leads to a higher classification error rate.*

2. *The random variable partitioning method works comparably well with the deterministic partitioning method. Besides, the optimal block numbers k^*, corresponding to the best classification error rate, are very close for these two methods. As the random variable partitioning method does not utilize any structural information of covariance or correlations, these results confirm the importance of reducing the dimensionality to training sample number ratio itself in improving the performance of LDA.*

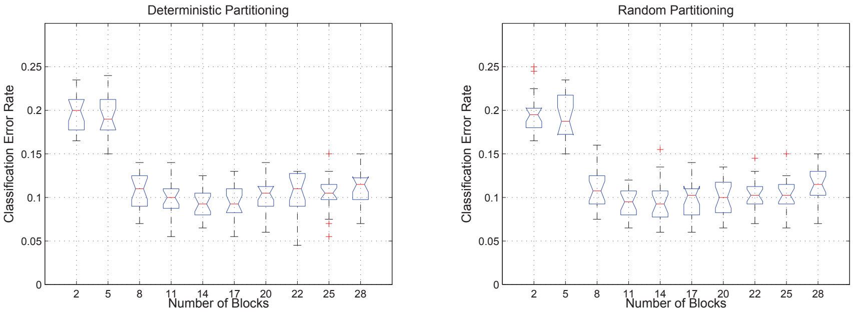


Figure 3.2: Evaluation of the block-diagonally regularized LDA on the Feret dataset

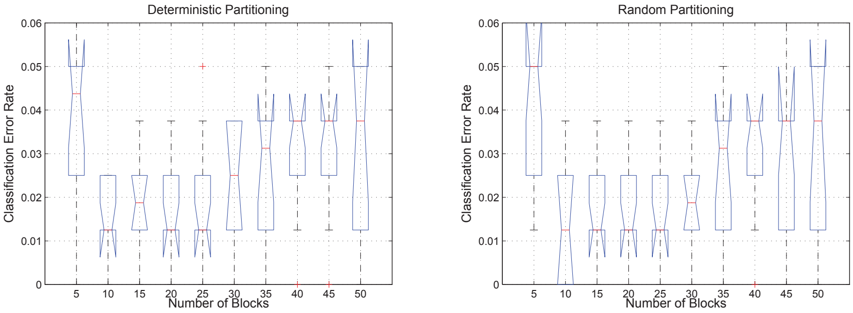


Figure 3.3: Evaluation of the block-diagonally regularized LDA on the Orl dataset

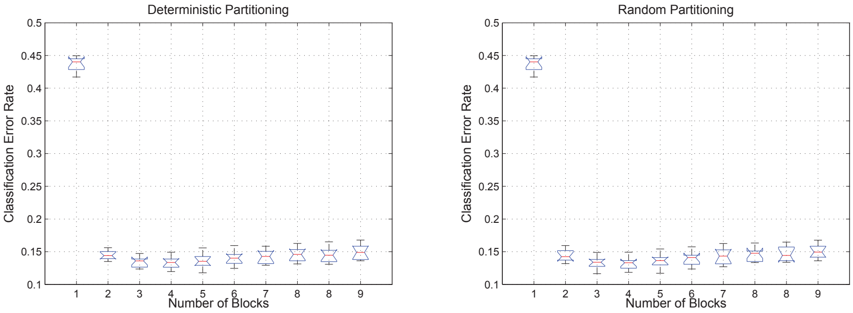


Figure 3.4: Evaluation of the block-diagonally regularized LDA on the Pie dataset

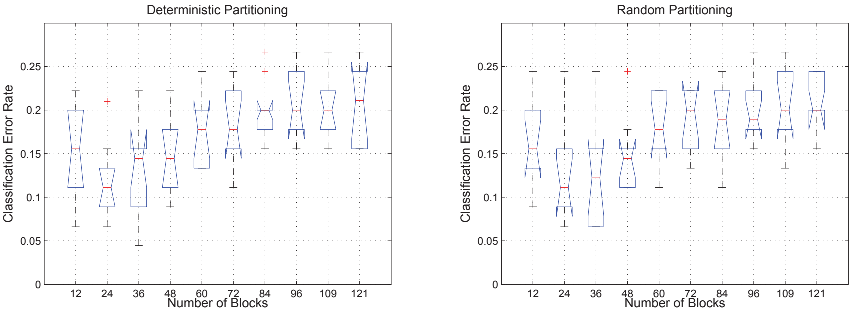


Figure 3.5: Evaluation of the block-diagonally regularized LDA on the Yale dataset

3.3.2 On Comparison with Other Regularization Methods

In this subsection, we compare the proposed block-diagonally regularized LDA with other types of regularized LDA. In the literature, the most commonly used type of regularized LDA is based on the Tikhonov regularization Ye et al. [2006] Guo et al. [2007]. Suppose the sample covariance matrix is $\widehat{\boldsymbol{\Sigma}}$, the Tikhonov regularized LDA optimizes the projection matrix by

$$\widehat{\mathbf{W}}_t^* = \arg\max_{\mathbf{W}} \Delta(\widehat{\boldsymbol{\Sigma}} + \rho\mathbf{I}, \widehat{\mathbf{S}}|\mathbf{W}), \tag{3.25}$$

where ρ is a tuning parameter to tradeoff between the approximation and the sample errors.

Another regularization method we intend to compare with is the banded estimation of the covariance matrix Bickel and Levina [2008b] Wagaman and Levina [2009]. It is based on the assumption that the population covariance matrix belongs to the approximately banded family

$$\{\boldsymbol{\Sigma} : \max_j \sum_i \{|\boldsymbol{\Sigma}_{ij}| : |i-j| > m\} \leq Cm^{-a} \text{ for all } k > 0,$$
$$\text{and } 0 < \epsilon \leq \lambda_{min}(\boldsymbol{\Sigma})\lambda_{max}(\boldsymbol{\Sigma}) \leq 1/\epsilon\},$$

where a, C, and ϵ are family parameters. Then, the banded estimator is given by

$$\widehat{\boldsymbol{\Sigma}}_B = \widehat{\boldsymbol{\Sigma}} \odot \mathbf{B}, \text{ with } \mathbf{B}_{ij} = 1(|i-j| > k), \tag{3.26}$$

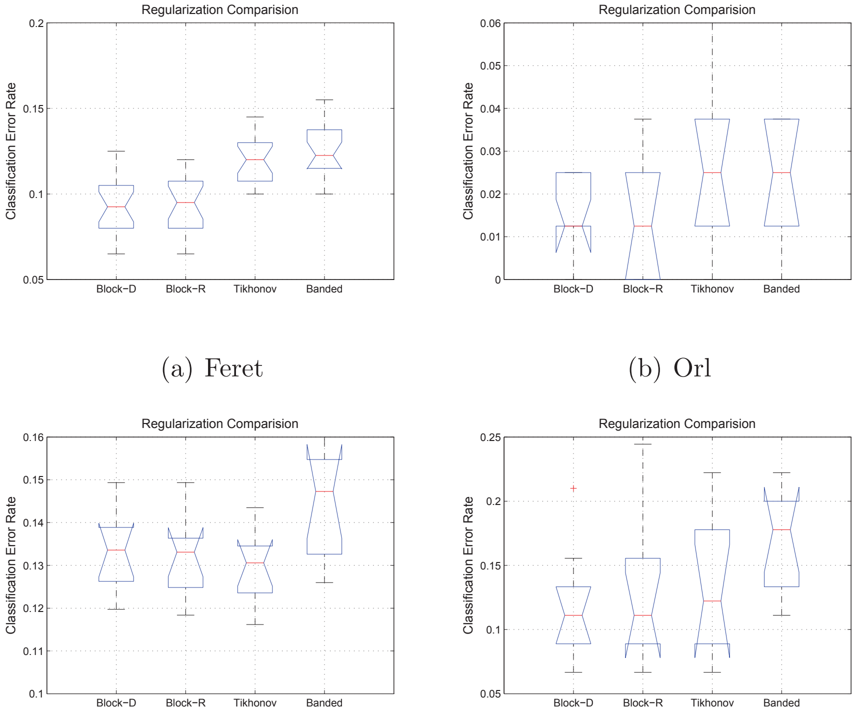
where $\odot$ denotes the Hadamard (entry-wise) product between two matrices. Note that in calculating $\mathbf{B}$, we need the order of variables. To this end, we use the Laplacian eigenmaps based method as described in Section 3.2.3 to get the order. Further, a problem with (3.26) is that $\widehat{\mathbf{\Sigma}}_B$ is not necessarily positive definite. Therefore, we modify $\mathbf{B}$ as

$$\mathbf{B}_{ij} = e^{-\frac{|i-j|^2}{\sigma^2}}, \tag{3.27}$$

where σ is a tuning parameter. Note that $\mathbf{B}$ defined by (3.27) is positive definite, and thus the Schur product theorem guarantees the banded estimator $\widehat{\mathbf{\Sigma}}_B = \widehat{\mathbf{\Sigma}} \odot \mathbf{B}$ to be positive definite. Accordingly, the LDA with banded regularization optimizes the projection matrix by

$$\widehat{\mathbf{W}}_b^* = \arg\max_{\mathbf{W}} \Delta(\widehat{\mathbf{\Sigma}}_B, \widehat{\mathbf{S}}|\mathbf{W}). \tag{3.28}$$

With the same experimental setting as in last subsection, we compare the proposed block-diagonal regularization with the Tiknonov regularization and the banded regularization. The parameters, k for the block-diagonal regularization, ρ for the Tikhnonov regularization, and σ for the banded regularization, are tuned with respect to the minimized test error. The classification performance on the four face image datasets is shown by Figure 3.6. One can see that on three out of the four datasets, i.e., Feret, Orl, and Yale, the proposed block-diagonal regularization outperforms the Tiknonov regularization and the banded regularization.



(a) Feret

(b) Orl

(c) Pie

(d) Yale

Figure 3.6: Comparison of different regularization methods on the Feret, Orl, Pie, and Yale dataset. Block-D, the block-diagonally regularized LDA with deterministic variable partitioning; Block-R, the block-diagonally regularized LDA with random variable partitioning; Tikhnonov, LDA with the Tikhonov regularization; and Banded, LDA with the banded regularization.

3.4 Appendixes

3.4.1 Proof of Theorem 3.1

Proof.

$$\begin{aligned} \mathcal{E}_a &= 2 \max_{\mathbf{W}} \left|\Delta(\mathbf{\Sigma}_r, \mathbf{S}|\mathbf{W}) - \Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W})\right| \\ &= 2 \max_{\mathbf{W}} \left|\mathrm{Tr}((\mathbf{W}^T\mathbf{\Sigma}_r\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S}\mathbf{W}) - \mathrm{Tr}((\mathbf{W}^T\mathbf{\Sigma}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S}\mathbf{W})\right| \end{aligned} \tag{3.29}$$

Recall that $\mathbf{X}^*$ simultaneously diagonalize $\mathbf{\Sigma}$ and $\mathbf{S}$. Then, by letting $\mathbf{W} =$

$\mathbf{X}^*\mathbf{V}$, $\mathbf{V}^T\mathbf{V}=\mathbf{I}$, we have

$$\begin{aligned}\mathcal{E}_a &= 2\max_{\mathbf{V}^T\mathbf{V}=\mathbf{I}}\left|\mathrm{Tr}\left(((\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V})^{-1}-\mathbf{I})\mathbf{V}^T\mathbf{\Lambda}\mathbf{V}\right)\right| \\ &\leq 2\max_{\mathbf{V}^T\mathbf{V}=\mathbf{I}}\max_i\left|\lambda_i((\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V})^{-1}-\mathbf{I})\right|\mathrm{Tr}(\mathbf{V}^T\mathbf{\Lambda}\mathbf{V}) \\ &\leq 2\max_{\mathbf{V}^T\mathbf{V}=\mathbf{I}}\max_i\left|\lambda_i((\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V})^{-1}-\mathbf{I})\right|\mathrm{Tr}(\mathbf{\Lambda}) \\ &= 2\max_{\mathbf{V}^T\mathbf{V}=\mathbf{I}}\max\{\lambda_{max}((\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V})^{-1})-1, \\ &\qquad\qquad 1-\lambda_{min}((\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V})^{-1})\}\mathrm{Tr}(\mathbf{\Lambda}) \\ &= 2\max_{\mathbf{V}^T\mathbf{V}=\mathbf{I}}\max\{\lambda_{min}^{-1}(\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V})-1, \\ &\qquad\qquad 1-\lambda_{max}^{-1}(\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V})\}\mathrm{Tr}(\mathbf{\Lambda}).\end{aligned} \tag{3.30}$$

On $\lambda_{max}(\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V})$, we have

$$\begin{aligned}\lambda_{max}(\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V}) &\leq \lambda_{max}(\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*) = \lambda_{max}(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}_r\mathbf{\Sigma}^{-\frac{1}{2}}) \\ &= \lambda_{max}(\mathbf{I}-\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}) \\ &\leq 1+\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\|.\end{aligned} \tag{3.31}$$

Similarly, on $\lambda_{max}(\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V})$, we have

$$\begin{aligned}\lambda_{min}(\mathbf{V}^T\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*\mathbf{V}) &\geq \lambda_{min}(\mathbf{X}^{*T}\mathbf{\Sigma}_r\mathbf{X}^*) = \lambda_{min}(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}_r\mathbf{\Sigma}^{-\frac{1}{2}}) \\ &= \lambda_{min}(\mathbf{I}-\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}) \\ &\geq 1-\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\|.\end{aligned} \tag{3.32}$$

Substituting (3.31) and (3.32) to (3.30), and noticing $\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\| = \rho < 1$, we get

$$\begin{aligned}&\mathcal{E}_a \leq 2\max\left\{\frac{\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\|}{1-\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\|}, \frac{\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\|}{1+\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\|},\right\}\mathrm{Tr}(\mathbf{\Lambda}) \\ &\leq \frac{2\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\|}{1-\|\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{\Sigma}-\mathbf{\Sigma}_r)\mathbf{\Sigma}^{-\frac{1}{2}}\|}\mathrm{Tr}(\mathbf{\Lambda}).\end{aligned} \tag{3.33}$$

By using the fact $\mathrm{Tr}(\mathbf{\Lambda}) = \Delta(\mathbf{\Sigma}, \mathbf{S}|\mathbf{W}^*)$, we complete the proof. □

3.4.2 Proof of Theorem 3.2

Proof. Suppose the eigendecompositions $\widehat{\mathbf{\Sigma}}_0 = \mathbf{U}\Lambda(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T$ and $\widehat{\mathbf{S}}_0 = \mathbf{V}\Lambda(\widehat{\mathbf{S}}_0)\mathbf{V}^T$, by using Lemma 2.1, we have

$$\Delta(\mathbf{\Sigma}_r, \mathbf{S}|\widehat{\mathbf{W}}^*) = \sum_{i=1}^{c} \boldsymbol{\delta}_i \boldsymbol{\lambda}_{ri}, \tag{3.34}$$

where

$$\boldsymbol{\delta}_i = \|\mathcal{R}^T(\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\mathbf{U}^T\mathbf{V}_{1:c})\mathbf{U}^T\mathbf{e}_i\|^2. \tag{3.35}$$

By using the same proofing steps as in the proof of Lemma 2.6, we have

$$\boldsymbol{\delta}_i \geq \max{}^2\{\cos(\theta_1 + \theta_2), 0\}, \tag{3.36}$$

and

$$\theta_1 = \arccos(\|\mathbf{V}_{1:c}^T\mathbf{e}_i\|), \tag{3.37}$$

$$\theta_2 = \arccos\left(\xi^T\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi\Big/\sqrt{\xi^T\Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\xi}\right), \tag{3.38}$$

where $\xi = \mathbf{U}^T\zeta$ and ζ is a unit-length random vector independent of $\mathbf{U}$.

For θ_1, note that Lemma 2.5 actually is valid even when $\gamma \in (0, \infty)$ because (2.92) holds for any $\gamma > 0$, and thus we have $\|\mathbf{V}_{1:c}^T\mathbf{e}_i\|^2 = \boldsymbol{\lambda}_{ri}/(\boldsymbol{\lambda}_{ri} + \gamma)$ and

$$\theta_1 = \arccos(\sqrt{\boldsymbol{\lambda}_{ri}/(\boldsymbol{\lambda}_{ri} + \gamma)}). \tag{3.39}$$

For θ_2, we need to calculate $\xi^T\Lambda^{-1}(\widehat{\mathbf{\Sigma}}_0)\xi$ and $\xi^T\Lambda^{-2}(\widehat{\mathbf{\Sigma}}_0)\xi$. By using Lemma 3.1, we have

$$\xi = \mathbf{U}^T\zeta = \mathrm{diag}(\mathbf{U}_1^T, \mathbf{U}_2^T, ..., \mathbf{U}_k^T)\mathbf{U}_r^*\zeta. \tag{3.40}$$

Letting $\mathbf{U}_r^* = [\mathbf{U}_{r1}^{*T}, \mathbf{U}_{r2}^{*T}, ..., \mathbf{U}_{rk}^{*T}]^T$, we have

$$\xi = [\xi_1^T, \xi_2^T, ..., \xi_k^T,]^T, \tag{3.41}$$

where

$$\xi_j = \mathbf{U}_j^T\zeta_j \text{ and } \zeta_j = \mathbf{U}_{rj}^*\zeta. \tag{3.42}$$

From Lemma 3.1, $\mathbf{U}_j^T$ follows the Haar distribution, and thus, conditioned on ζ_j, $\frac{1}{\|\zeta_j\|}\xi_j$ is a unit-length random vector on the sphere $\mathbb{S}^{D/k-1}$. Then, by using Lemma 2.3 k times, we have

$$\begin{aligned}
\xi^T \Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi &= \sum_{j=1}^{k} \xi_j^T \Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_{0j})\xi_j \\
&= \sum_{j=1}^{k} \|\zeta_j\|^2 (\xi_j/\|\zeta_j\|)^T \Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_{0j})(\xi_j/\|\zeta_j\|) \\
&= \sum_{j=1}^{k} \|\zeta_j\|^2 \frac{1}{1-\gamma/k} \\
&= \frac{1}{1-\gamma/k} \sum_{j=1}^{k} \|\mathbf{U}_{rj}^* \zeta\|^2 \\
&= \frac{1}{1-\gamma/k} \|\zeta\|^2 \\
&= \frac{1}{1-\gamma/k}.
\end{aligned} \tag{3.43}$$

With the similar procedure and by using Lemma 2.4, we have

$$\xi^T \Lambda^{-1}(\widehat{\boldsymbol{\Sigma}}_0)\xi = \frac{1}{(1-\gamma/k)^3}. \tag{3.44}$$

Substituting (3.43) and (3.44) into (3.38), we get

$$\theta_2 = \arccos(\sqrt{1-\gamma/k}). \tag{3.45}$$

This completes the proof. □

Part II

Algorithmic Extensions

Chapter 4

Max-Min Distance Analysis for Parametric SLDR

4.1 Introduction

Parametric methods consist of a major subcategory in SLDR. By modeling with certain probability distribution families, e.g., the homoscedastic or heteroscedastic Gaussian distributions, parametric SLDR generally has a low model complexity in density estimation and a low computational cost in optimizing the projection matrix. Due to these advantages, parametric SLDR has been applied to wide range of areas, from speech analysis spe [1998] and image retrieval He et al. [2008] to face recognition Belhumeur et al. [1997] Kim and Kittler [2005].

A major problem with existing parametric SLDR methods, including LDA and its many extensions, is that when the dimensionality of the learned subspace is low, close class pairs tend to merge. This is referred to as the "class separation" problem in the literature Tao et al. [2009]. A number of methods have been developed to address this problem in recent years, which be categorized into the following categories:

- The first category contains weighting scheme based methods. For example, Lotlikar and Kothari [2000] proposed the fractional step LDA (FS-LDA), which reduces the space dimensionality by using a series of fractional steps and meanwhile uses weighting functions to emphasize the discrimination power between close class pairs. Loog et al. [2001] proposed the approximate pairwise accuracy criterion (aPAC). Similar to FS-LDA, aPAC also puts higher weights on close class pairs, and additionally it provides a way to calculate the weight by referring to the Bayes error.

- The second category is general mean based methods. These methods are motivated by the fact that Fisher's criterion in LDA is equivalent to maximizing the arithmetic mean of the discrimination power of all class pairs. By replacing the arithmetic mean by other mean functions, e.g., the geometric mean and the harmonic mean, the obtained criteria automatically emphasize close class pairs and thus give rise to a better class separation than LDA. The geometric and the harmonic mean based subspace selection methods (GMSS and HMSS) were developed in Tao et al. [2009] and Bian and Tao [2008], respectively.

- The methods in the third category are based on Bayes error minimization. In general, Bayes error minimization is intractable due to the hardness of calculating the Bayes error, though it is theoretically the most favorable for dimension reduction. However, exception exits in special cases. For example, Schervish [1984] proposed a method to select the one-dimensional Bayes optimal subspace in a three-class problem represented by homoscedastic Gaussian distributions. And most recently, Hamsici and Martinez [2008] generalized Schervish's method to multiclass problems, and showed that the one-dimensional Bayes optimal subspace is achievable given the order of the class means located in the one-dimensional Bayes optimal subspace. Since the order of locations of means are unknown in advance, enumeration of the order is needed in their method. Further, they also proposed a greedy algorithm to sequentially select orthogonal one-dimensional subspaces so as to compose a subspace of higher dimensionality.

In this chapter, we propose a new method for parametric SLDR based on the homoscedastic Gaussian assumption, termed max-min distance analysis (MMDA). MMDA optimizes the projection matrix by maximizing the minimum pairwise distance among all class pairs in the dimension reduced space. Thus, it duly considers the separation of all classes. Unfortunately, MMDA is hard to optimize directly due to the non-smoothness of the objective function and the orthonormal constraints. Therefore, we derive an approximate algorithm for MMDA by using the sequential convex relaxation technique. On both synthetic data experiments and face recognition experiments, we show that MMDA performs competitively compared with the state-of-the-art parametric SLDR methods.

Note that for a $c+1$-class problem represented by homoscedastic Gaussian distributions, the c-dimensional subspace learned by LDA contains the entire discrimination power. Thus, for any method based on the homoscedastic Gaussian assumption, including aforementioned MMDA, GMSS, HMSS, aPAC, FS-LDA, we can first perform LDA and then perform the method in the dimension reduced space of LDA. After performing LDA, the data dimensionality becomes c and the sample covariance becomes $\widehat{\mathbf{W}}^* \widehat{\mathbf{\Sigma}} \widehat{\mathbf{W}}^* = \mathbf{I}_c$, where $\widehat{\mathbf{W}}^*$ is the projection matrix learned by LDA. Thus, we can assume the sample means $\widehat{\boldsymbol{\mu}}_i \in \mathbb{R}^c$, $i = 1, 2, ..., c+1$, and the sample covariance $\widehat{\mathbf{\Sigma}} = \mathbf{I}_c$.

The rest of this chapter is organized as follow. In Section 4.2, we present the MMDA criterion, and discuss its relationship to other criterions. In Section 4.3, we derive an approximate algorithm to solve MMDA by using sequential convex relaxation technique. Section 4.4 presents empirical evaluations on both synthetic and real datasets.

4.2 Max-Min Distance Analysis

4.2.1 MMDA Criterion

Given the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \mathbf{I}_c$, we define the discrimination power between class ω_i and ω_j in the subspace $\mathbf{W} \in \mathbb{R}^{c \times d}$, $d < c$, by

$$\Delta(\omega_i, \omega_j | \mathbf{W}) = \mathrm{Tr}(\mathbf{W}^T \mathbf{D}_{ij} \mathbf{W}), \ 1 \leq i < j \leq c+1, \tag{4.1}$$

where

$$\mathbf{D}_{ij} = (\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j)(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j)^T \tag{4.2}$$

is called the distance matrix between ω_i and ω_j. Since $\mathrm{Tr}(\mathbf{W}^T \mathbf{D}_{ij} \mathbf{W}) = \|\mathbf{W}^T(\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_j)\|^2$, the pairwise discrimination power $\Delta(\omega_i, \omega_j | \mathbf{W})$ is the squared distance between the two class means in the subspace $\mathbf{W}$.

Then, the MMDA criterion is defined by

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_d} \min_{1\leq i<j\leq C} \Delta(\omega_i, \omega_j|\mathbf{W}). \tag{4.3}$$

In (4.3), the inner minimization chooses the minimum pairwise (squared) distance of all class pairs in $\mathbf{W}$, while the outer maximization maximizes this minimum pairwise distance. This explains the name of MMDA. Let the optimal value and solution of (4.3) be Δ_{opt} and $\mathbf{W}_{opt}$, respectively. We have

$$\Delta(\omega_i, \omega_j|\mathbf{W}_{opt}) \geq \Delta_{opt}, \text{for all } i \neq j, \tag{4.4}$$

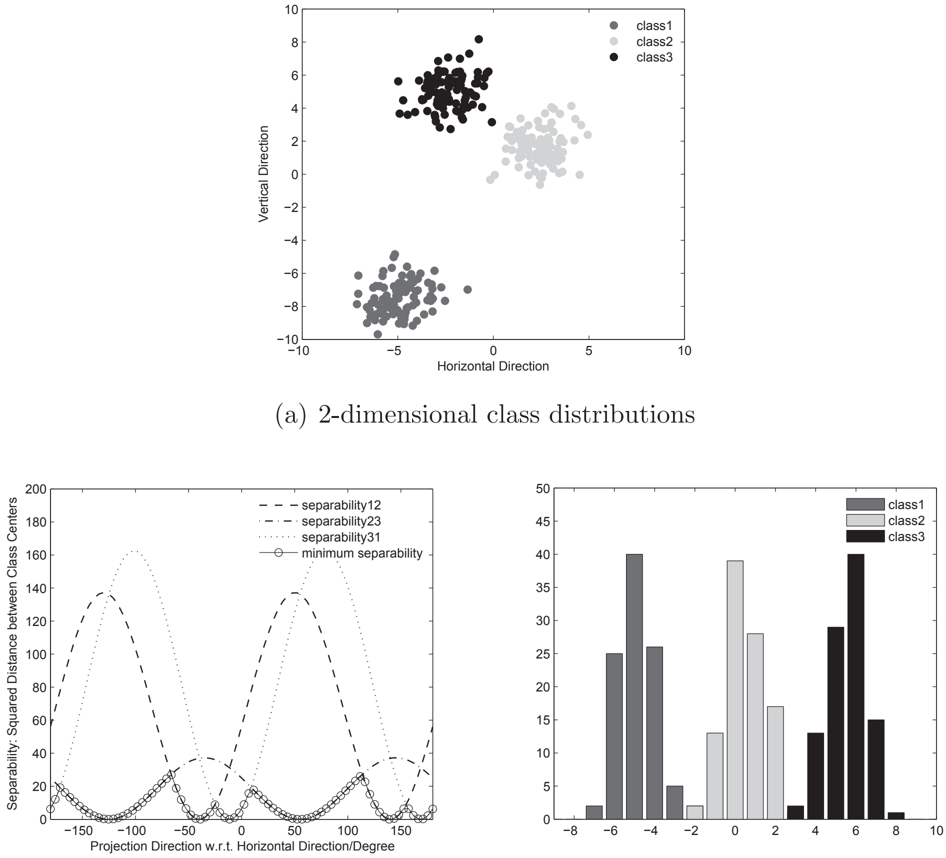
which guarantees the separation (as best as possible) of any class pairs in the selected low dimensional subspace.

Figure 4.1 shows the results of MMDA on a toy example, which is a three-class problem and requires a 1-dimensional subspace to separate these three classes. Varying the 1-dimensional subspace, by changing its angle with respect to the horizontal direction, the three pairwise distances change, and the minimum distance among the three classes is maximized at the direction about 115 degrees, i.e., the direction of MMDA subspace. The corresponding class distributions (histograms) after projected onto the MMDA subspace are shown in Figure 4.1 (c), which shows that all classes are well separated from the others.

Remark 4.1. MMDA criterion (4.3) can be interpreted by the minimax decision rule. For a binary classification, e.g., between classes ω_i and ω_j, we can define a loss function as $\ell(\omega_i, \omega_j|\mathbf{W}) = e^{-\Delta(\omega_i,\omega_j|\mathbf{W})}$. Thus, a minimax decision can be made by minimizing the maximum binary classification loss

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_d} \max_{1\leq i<j\leq C} \ell(\omega_i, \omega_j|\mathbf{W}). \tag{4.5}$$

Since $e^{-(\cdot)}$ is a monotonically decreasing function, minimax decision rule (4.5) is equivalent to the MMDA criterion (4.3).



(a) 2-dimensional class distributions

(b) Plots of pairwise distance and the minimum pairwise distance

(c) Histogram of three classes projected onto the MMDA direction

Figure 4.1: MMDA for three Gaussian distributions on the 2-dimensional space.

4.2.2 Relationships with Other Criteria

Here, we discuss the relationships between MMDA and other criteria, including, LDA Rao [1948], GMSS Tao et al. [2009], HMSS Bian and Tao [2008], and BLDA Hamsici and Martinez [2008].

4.2.2.1 MMDA vs. LDA, GMSS and HMSS

According to Loog et al. [2001], Fisher's criterion in LDA is equivalent to maximizing the sum of the pairwise discrimination powers of all class pairs, i.e.,

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_d} \sum_{1\leq i<j\leq c+1} \Delta(\omega_i, \omega_j|\mathbf{W}), \tag{4.6}$$

which follows the fact that $\mathbf{S} = \frac{1}{(c+1)^2}\sum_{1\leq i<j\leq c+1}\mathbf{D}_{ij}$. The criterion (4.6) puts equal weights on all class pairs, and thus close class pairs are sacrificed in the optimization and tend to be merged together in the selected low dimensional subspace.

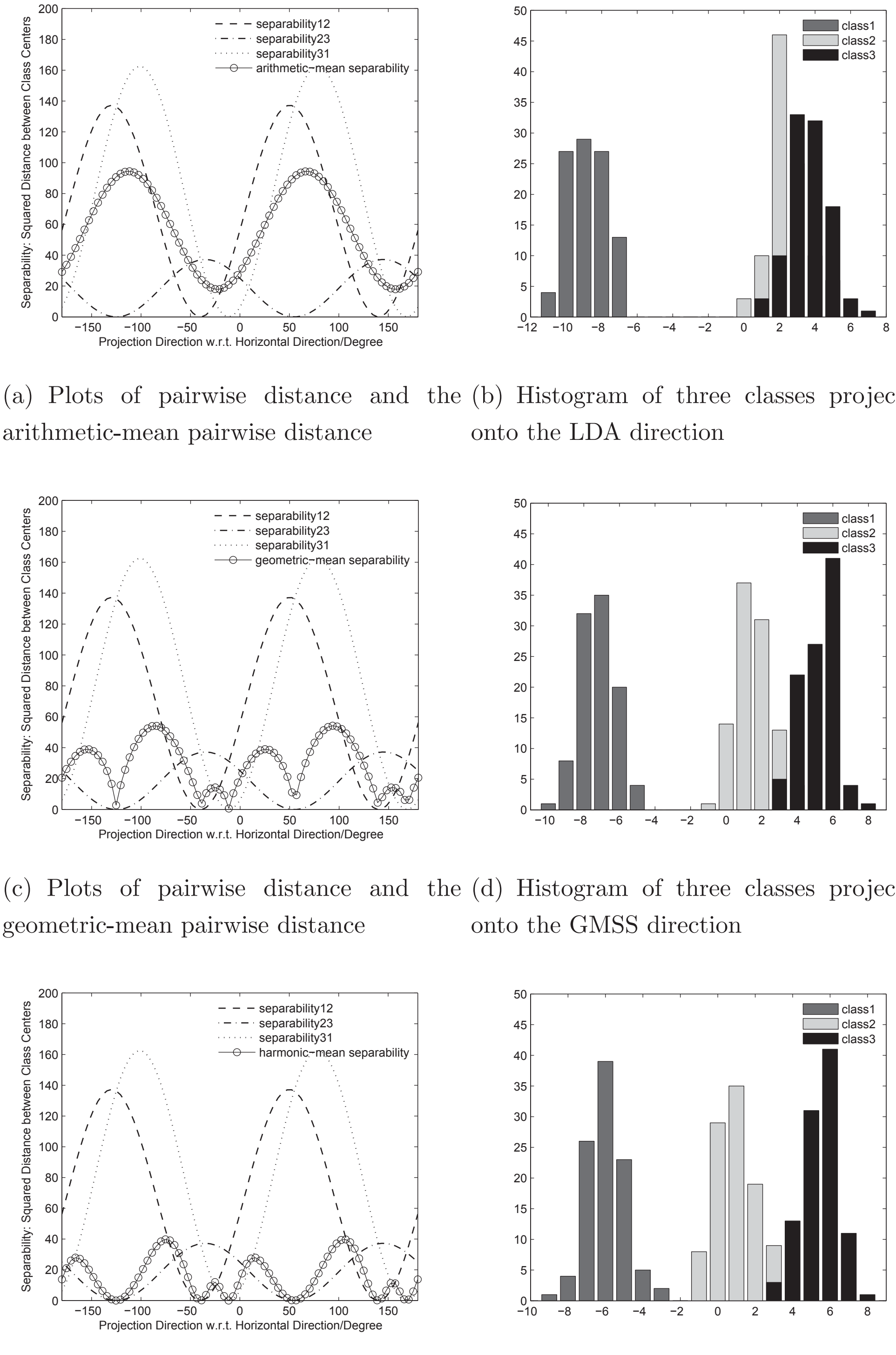
To improve the separation of close class pairs, GMSS and HMSS utilize the geometric and the harmonic means to replace the arithmetic mean in (4.6), which are defined as

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_d} \prod_{1\leq i<j\leq c+1} \Delta(\omega_i, \omega_j|\mathbf{W}), \tag{4.7}$$

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_d} \left[\sum_{1\leq i<j\leq c+1} \left(\Delta(\omega_i, \omega_j|\mathbf{W})\right)^{-1}\right]^{-1}, \tag{4.8}$$

Due to the characteristic of the geometric and the harmonic means, they adaptively put large weights on close class pairs, and thus can achieve better class separation than LDA.

However, all criteria above still suffer from the limitation that the separation of all class pairs cannot be guaranteed. Figure 4.2 shows the corresponding results of LDA, GMSS, and HMSS on the toy example shown in Figure 4.1. It can be observed that LDA tends to merge class 2 and class 3 together, and, by contrast, GMSS and HMSS give improved class separation. However, unlike MMDA, none of them is able to separate all class pairs.



(a) Plots of pairwise distance and the arithmetic-mean pairwise distance

(b) Histogram of three classes projected onto the LDA direction

(c) Plots of pairwise distance and the geometric-mean pairwise distance

(d) Histogram of three classes projected onto the GMSS direction

(e) Plots of pairwise distance and the harmonic-mean pairwise distance

(f) Histogram of three classes projected onto the HMSS direction

Figure 4.2: LDA, GMSS, and HMSS for three Gaussian distributions on the 2-dimensional space.

4.2.2.2 MMDA vs. the Bayes Optimal Criterion

Blow, we compare MMDA and the Bayes optimal criterion, i.e., Bayes error minimization. We apply the three-class problem in Schervish [1984], which contains three standard Gaussian distributions in the 2-dimensional space, and show the difference between MMDA 1-dimensional subspace and the Bayes optimal 1-dimensional subspace. According to Hamsici and Martinez [2008], the Bayes optimal 1-dimensional subspace can be obtained by

$$\min_{v} \Phi\left(\mathbf{v}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/2\right) + \Phi\left(\mathbf{v}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)/2\right), \tag{4.9}$$

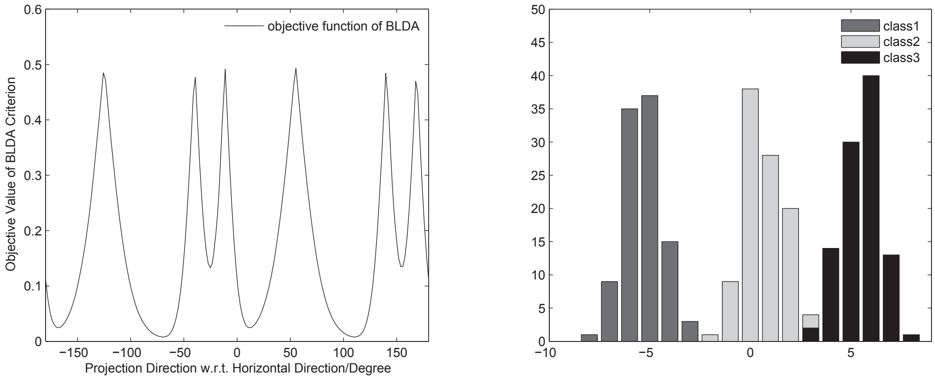
where $\boldsymbol{\mu}_i$, $i = 1, 2, 3$, are the class means, the vector $\mathbf{v}$ represents the 1-dimensional subspace, and $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian distribution. It is required that $\mathbf{v}^T\boldsymbol{\mu}_1 \leq \mathbf{v}^T\boldsymbol{\mu}_2 \leq \mathbf{v}^T\boldsymbol{\mu}_3$, according to [10]. Let $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = d_{12}$, $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3\| = d_{23}$, and the angle between vectors $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)$ be α_0. Then (4.9) can be rewritten as a maximization problem with respect to the angle α between $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)$ and $\mathbf{v}$, i.e.,

$$\max_{\alpha} \Phi\left(-d_{12}\cos(\alpha)/2\right) + \Phi\left(-d_{23}\cos(\alpha_0 - \alpha)/2\right). \tag{4.10}$$

By taking derivative of (4.10) with respect to α, and setting it to be zero, we arrive at the Bayes optimal solution α_{opt} and

$$\frac{\exp\left(-\left(d_{12}\cos(\alpha_{opt})\right)^2/8\right)}{\exp\left(-\left(d_{23}\cos(\alpha_0 - \alpha_{opt})\right)^2/8\right)} = \frac{d_{23}\sin(\alpha_0 - \alpha_{opt})}{d_{12}\sin(\alpha_{opt})} = \gamma_{opt} \tag{4.11}$$

In general, $\gamma_{opt} \neq 1$, and suppose $\gamma_{opt} > 1$ without loss of generality, which implies $d_{12}\cos(\alpha_{opt}) < d_{23}\cos(\alpha_0 - \alpha_{opt})$. It shows that the minimum pairwise distance is determined by $\min\{d_{12}\cos(\alpha_{opt}), d_{23}\cos(\alpha_0 - \alpha_{opt})\}$, and thus one can increase the minimum distance by moving α from α_{opt} to zero, until $d_{12}\cos(\alpha_{opt}) = d_{23}\cos(\alpha_0 - \alpha_{opt})$ for a certain α, or letting $\alpha = 0$ if this equation cannot be achieved. Therefore, the Bayes optimal criterion does not maximize the minimum distance of all class pairs, though it minimizes the over-all classification error. In contrast, MMDA duly concerns the separation of all classes. Figure 4.3 shows the corresponding results of BLDA on the toy example used in Figure 4.1.



(a) The objective of BLDA criterion (4.9)

(b) Histogram of three classes projected onto the BLDA direction

Figure 4.3: BLDA for three Gaussian distributions on the 2-dimensional space.

4.3 Sequential Convex Relaxation

The optimization problem in MMDA (4.3) is hard to solve directly. First, the inner minimization is over discrete variables i and j, which makes the objective function for the outer maximization nonsmooth. Second, the orthonormal constraints are also difficult to deal with in general, except for the spectrum analysis based problems, such as PCA Jolliffe [2002] and LDA Fisher [1936] Rao [1948]. In this section, we propose to solve (4.3) approximately, by developing a sequential convex relation based algorithm. Specifically, each convex relation leads to a semidefinite programming (SDP) problem, for which efficient solver exists when the size of the problem is moderate.

4.3.1 Global SDP Relaxation

Recall the optimization problem of MMDA

$$\begin{aligned} \max_{\mathbf{W}} \min_{1\le i<j\le c+1} \quad & \mathrm{Tr}(\mathbf{W}^T\mathbf{D}_{ij}\mathbf{W}) \\ \text{subject to} \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I}_d. \end{aligned} \tag{4.12}$$

By introducing a matrix variable $\mathbf{X} = \mathbf{W}\mathbf{W}^T$ and an auxiliary scale variable t, and further utilizing the invariant property of trace, $\mathrm{Tr}(\mathbf{W}^T\mathbf{D}_{ij}\mathbf{W}) = \mathrm{Tr}(\mathbf{D}_{ij}\mathbf{X})$, we can equivalently transform (4.12) into

$$\begin{aligned} \min \quad & -t \\ \text{subject to} \quad & \mathrm{Tr}(\mathbf{D}_{ij}\mathbf{X}) \geq t, \;\; 1 \leq i < j \leq c+1 \\ & \mathbf{X} = \mathbf{W}\mathbf{W}^T \\ & \mathbf{W}^T\mathbf{W} = \mathbf{I}_d. \end{aligned} \tag{4.13}$$

The objective function and the trace constraints in (4.13) are easy to deal with, since they are linear with respect to $(t, \mathbf{X})$. The difficulty still lies in the complex constraints on $\mathbf{X}$. Next, we introduce a theorem, with which we can relax these constraints into convex ones.

Theorem 4.1. *Overton and Womersley [1992] Let $\Omega_1 = \{\mathbf{X} : \mathbf{X} = \mathbf{W}\mathbf{W}^T, \mathbf{W}^T\mathbf{W} = \mathbf{I}_d\}$ and $\Omega_2 = \{\mathbf{X} : Tr(\mathbf{X}) = d, 0 \preceq \mathbf{X} \preceq \mathbf{I}_c\}$, wherein $\mathbf{W}$ is of the size c by d, and $\mathbf{X}$ has dimension of c by c. The second condition $0 \preceq \mathbf{X} \preceq \mathbf{I}$ means that both $\mathbf{X}$ and $\mathbf{I} - \mathbf{X}$ are positive semidefinite. Then Ω_2 is the convex hull of Ω_1, and Ω_1 is the set of extreme points[1] of Ω_2.*

According to Theorem 4.1, Ω_2 is a convex relaxation of Ω_1, and it is the tightest convex relaxation since Ω_2 is the convex hull of Ω_1. Using this result, we can relax (4.13) to the convex problem below

$$\begin{aligned} \min \quad & -t \\ \text{subject to} \quad & \mathrm{Tr}(\mathbf{D}_{ij}\mathbf{X}) \geq t, 1 \leq i < j \leq c+1 \\ & \mathrm{Tr}(\mathbf{X}) = d \\ & 0 \preceq \mathbf{X} \preceq \mathbf{I}_c. \end{aligned} \tag{4.14}$$

Problem (4.14) is a semidefinite programming (SDP) problem, though not in a canonical form. We call it the global SDP relaxation of MMDA, in contrast to the local SDP relaxation developed afterwards. Denote by $\mathbf{X}_{opt}$ the optimal solution of (4.14). If $\mathbf{X}_{opt}$ has rank d, then the eigenvectors associated to nonzero eigenvalues consist of the optimal solution of MMDA (4.12). However, in general, the rank of $\mathbf{X}_{opt}$ is not exactly d, and to get a $c \times d$ projection matrix $\mathbf{W}$, we

[1] A point x of a convex set S is an extreme point if only if it cannot be expressed as a convex combination of other points in $\bar{S}$, which is the closure of S.

eigendecompose $\mathbf{X}_{opt}$ and use its first d eigenvectors to construct an approximately optimal solution $\mathbf{W}_{app}$ of (4.12).

4.3.2 Local SDP Relaxation

The key point of the above global SDP relaxation is to relax the original feasible set Ω_1 to its convex hull. Though such relaxation is commonly used in many problems d'Aspremont et al. [2007], it could be relatively loose. This is because Ω_1 only contains extreme points of Ω_2, or in other words, too many infeasible points are contained in Ω_2. To address this problem, we develop a local SDP relaxation below, which makes relaxation only around a properly selected initial point $\mathbf{X}_0$ in Ω_1. The Theorem 4.2 below will be used in constructing the local SDP relaxation.

Theorem 4.2. *Let Ω_1 be the same as in Theorem 3.1, and $\Omega_3 = \{\mathbf{X} : Tr(\mathbf{X}) = d, 0 \preceq \mathbf{X} \preceq \mathbf{I}_c, \det(\mathbf{X} + \delta\mathbf{I}_c) = (1+\delta)^d\delta^{c-d}\}$. Then, Ω_3 is equivalent to Ω_1, for any $\delta > 0$.*

Proof. Let the eigenvalues of $\mathbf{X}$ be λ_i, $1 \le i \le c$, and then the constraints in Ω_3 is equivalent to $0 \le \lambda_i \le 1$, $\sum_i \lambda_i = d$, and $\prod_i (\lambda_i + \delta) = (1+\delta)^d\delta^{c-d}$. Both the left hand side and the right hand side of the last equation on λ_i are polynomial functions of δ, and the right hand side has only root of 0 or -1. Thus, λ_i should be either 0 or 1 to keep the equality between the two polynomial functions, which meets the orthonormal constraints in Ω_1. The equivalence between Ω_1 and Ω_3 is immediate. It can be further proved that $\det(\mathbf{X} + \delta\mathbf{I}_c) \le (1+\delta)^d\delta^{c-d}$, for $\mathrm{Tr}(\mathbf{X}) = d$, $0 \preceq \mathbf{X} \preceq \mathbf{I}_c$. This completes the proof. □

Based on Theorem 4.2, we consider a local convex relaxation around a given point $\mathbf{X}_0$ in Ω_3 (which is also in Ω_1), and particularly we set $\mathbf{X}_0 = \mathbf{W}_{app}\mathbf{W}_{app}^T$, wherein $\mathbf{W}_{app}$ is the projection matrix obtained from the global convex relaxation (4.14). The only nonconvex condition in Ω_3 is the determinant constraint, so we apply the first order Taylor expansion to approximate $\det(\mathbf{X} + \delta\mathbf{I})$, which gives

$$\det(\mathbf{X} + \delta\mathbf{I}_c) \approx \det(\mathbf{X}_0 + \delta\mathbf{I}_c) + \det(\mathbf{X}_0 + \delta\mathbf{I}_c)\mathrm{Tr}\left((\mathbf{X}_0 + \delta\mathbf{I}_c)^{-1}(\mathbf{X} - \mathbf{X_0})\right). \tag{4.15}$$

Since we are only interested in a point $\mathbf{X} \in \Omega_3$, i.e., in the original feasible set, it is reasonable to restrict

$$\det(\mathbf{X} + \delta \mathbf{I}_c) = \det(\mathbf{X}_0 + \delta \mathbf{I}_c) = (1+\delta)^d \delta^{c-d}, \tag{4.16}$$

which together with (4.15) implies

$$(1+\delta)^d \delta^{c-d} \left| \mathrm{Tr}\left((\mathbf{X}_0 + \delta \mathbf{I}_c)^{-1} (\mathbf{X} - \mathbf{X_0}) \right) \right| \le \varepsilon, \tag{4.17}$$

where ε is an upper bound on the high order infinitesimal of the expansion (4.15).

Rearranging (4.17), we get

$$\begin{cases} \mathrm{tr}\left((\mathbf{X}_0 + \delta \mathbf{I}_c)^{-1} \mathbf{X} \right) \le (1+\eta) \mathrm{tr}\left((\mathbf{X}_0 + \delta \mathbf{I}_c)^{-1} \mathbf{X}_0 \right) \\ \mathrm{tr}\left((\mathbf{X}_0 + \delta \mathbf{I}_c)^{-1} \mathbf{X} \right) \ge (1-\eta) \mathrm{tr}\left((\mathbf{X}_0 + \delta \mathbf{I}_c)^{-1} \mathbf{X}_0 \right), \end{cases} \tag{4.18}$$

where $\eta = \varepsilon / \delta^{c-d}$. In (4.18), the parameter δ specifies the normal direction of the hyper-plane given by the linear term $tr\left((\mathbf{X}_0 + \delta \mathbf{I})^{-1} \mathbf{X} \right)$: when δ approaches to zero, this direction becomes parallel to the eigenspace associated with the 0 eigenvalue of $\mathbf{X}_0$; when δ approaches to infinity, no particular direction is preferred. In this paper, we set $\delta = 1$, which implies we set a 2 to 1 weighting ratio on the eigenspaces associated with 0 and 1 eigenvalues of $\mathbf{X}_0$.

By combining (4.18) with (4.14), we get the local SDP relaxation of (4.12) around an initial point $\mathbf{X}_0$

$$\begin{aligned} \min \quad & -t \\ \text{subject to} \quad & tr\left(\mathbf{A}_{ij} \mathbf{X} \right) \ge t, \quad 1 \le i < j \le c+1 \\ & \mathrm{tr}(\mathbf{X}) = d \\ & 0 \preceq \mathbf{X} \preceq \mathbf{I}_c \\ & \mathrm{tr}\left((\mathbf{X}_0 + \mathbf{I}_c)^{-1} \mathbf{X} \right) \le (1+\eta) \mathrm{tr}\left((\mathbf{X}_0 + \mathbf{I}_c)^{-1} \mathbf{X}_0 \right) \\ & \mathrm{tr}\left((\mathbf{X}_0 + \mathbf{I}_c)^{-1} \mathbf{X} \right) \ge (1-\eta) \mathrm{tr}\left((\mathbf{X}_0 + \mathbf{I}_c)^{-1} \mathbf{X}_0 \right). \end{aligned} \tag{4.19}$$

In (4.19), the parameter η controls the volume of the feasible set determined by the local relaxation (i.e., the pair of inequalities). The smaller η is, the smaller the volume of the feasible set of (4.19) will be, and when η is exactly 0, the feasible set of (4.19) will degenerate to a single point $\mathbf{X}_0$. Alternatively, if $\eta = \infty$, the last two inequalities in (4.19) will be satisfied automatically, and in this case the local SDP relaxation reduces to the global relaxation.

4.3.3 Iterative Local SDP Relaxation

Further, we suggest using (4.19) iteratively. Suppose the optimal solution of (4.19) in the k-th iteration is $\mathbf{X}_{opt}^{(k)}$, we construct the projection matrix $\mathbf{W}_{app}^{(k)}$ by using the first d eigenvectors of $\mathbf{X}_{opt}^{(k)}$. Then, we set $\mathbf{X}_0^{(k+1)} = \mathbf{W}_{app}^{(k)}\mathbf{W}_{app}^{(k)T}$ in (4.19) and do the next iteration. In addition, the parameter $\eta(k)$ in the iterations is set as a decreasing sequence to 0, analogous to an "annealing" process. According to the previous discussion on the parameter η, the feasible set of (4.19) shrinks with the decreasing of $\eta(k)$. When $\eta(k)$ decreases to 0, the feasible set of (4.19) will converge to a particular single point $\mathbf{X}_0^{(k)}$, and thus we can get a converged projection matrix $\mathbf{W}_{app}^{(k)}$. In this chapter, we set $\eta(k)$ as a sequence starts from 10^{-2} and decreases to 10^{-6} in 20 iterations. Algorithm 1 summarizes the pseudo code for the iterative Local SDP relaxation procedure above.

Algorithm 1 Iterative Local SDP Relaxation for MMDA

```
Input: Sample means data $\widehat{\boldsymbol{\mu}}_i$, $i = 1, 2, ..., c+1$.
Output: Projection matrix $\mathbf{W} \in \mathbb{R}^{c \times d}$.

  Step 1. Calculate distance matrix $\mathbf{D}_{ij}$, $1 \le i < j \le c+1$, by (4.2).
  Step 2. Solve the global SDP relaxation (4.14), obtaining $\mathbf{X}_{opt}$. Construct $\mathbf{W}_{app}$ by the first $d$ eigenvectors of $\mathbf{X}_{opt}$. Let $\mathbf{X}_0^{(1)} = \mathbf{W}_{app}\mathbf{W}_{app}^T$.
  Step 3. Set $\eta(k)$, $k = 1, 2, ..., K$, as a decreasing sequence to 0.
  for $k = 1$ to $K$ do
    Solve the local SDP relaxation (4.19) with $\eta = \eta(k)$, and $\mathbf{X}_0 = \mathbf{X}_0^{(k)}$, obtaining $\mathbf{X}_{opt}^{(k)}$. Construct $\mathbf{W}_{app}^{(k)}$ by the first $d$ eigenvectors of $\mathbf{X}_{opt}^{(k)}$. Let $\mathbf{X}_0^{(k+1)} = \mathbf{W}_{app}^{(k)}\mathbf{W}_{app}^{(k)T}$.
  end for
  Step 5. Set $\mathbf{W} = \mathbf{W}_{app}^{(K)}$.
```

Remark 4.2. The original MMDA problem (4.12) is nonconvex and thus has multiple local optima. However, it is worth emphasizing that its feasible set (equivalent to Ω_1) only contains extreme points of its convex hull Ω_2, and thus has no interior point, which makes conventional gradient based methods difficult to be applied, because any finite step line search will exceed the feasible set.

Remark 4.3. Algorithm 1 requires solving several SDP problems and thus is generally time-consuming compared against conventional spectral decomposition based dimension reduction algorithms, such as PCA and LDA. The worst case computational complexity of interior-point methods for SDP is $\mathcal{O}(m_0^2 n_0^2)$, where m_0 is the number of variables and n_0 is the size of the problem. For the proposed local SDP relaxation (4.19), $m_0 = c(c+1)/2+1$, wherein $c(c+1)/2$ is the number of independent variables in the symmetric matrix $\mathbf{X}$ and 1 is for the variable t, and $n_0 = 2c+c(c+1)/2+4$, wherein $2m$ is for the two-side inequality $0 \preceq \mathbf{X} \preceq \mathbf{I}$, $c(c+1)/2$ is the number of trace inequalities $\mathrm{Tr}(\mathbf{D}_{ij}\mathbf{X}) \geq t$, $1 \leq i < j \leq c+1$, and 4 is for the trace equality $\mathrm{Tr}(\mathbf{X}) = d$, which leads to 2 inequalities, and the last 2 trace inequalities in (4.19). Thus, the computational complexity of (4.19) is $\mathcal{O}(c^8)$, determined by the class number $c+1$. Therefore, the SDP relaxation based algorithm is applicable for moderate class number $c+1$. Empirically, we found that for $c \leq 50$, our algorithm is acceptably efficient. In future work, we will consider developing more efficient algorithm for problems with large class number $c+1$. So far, we use off-the-shelf package to solve the SDP problem in MMDA. One of these solvers is SDPA-M Fujisawa et al. [2000], which can tackle SDP problems with size of thousands or even tens of thousands for low rank or sparse problems.[1] On face recognition experiments in Section 4.4.2, the training time of Algorithm 1 is within 10 minutes, all implemented on a PC with 3.4 GHz CPU frequency and 2 GB memory. Besides, for moderate size problems, e.g., on the synthetic dataset experiments in Section 4.4.1, the disciplined convex programming MATLAB software CVX Grant and Boyd Grant and Boyd [2010] is also applicable.

Remark 4.4. Iterative using of the convex relaxation is a common used technique in solving problems with rank constraints Dattorro [2008] and low rank approximations Fazel et al. [2003]. MMDA (4.12) can be casted into a rank constrained problem by adding to (4.14) the constraint $rank(\mathbf{X}) = d$. With the rank

[1] Low rank or sparse SDP problem refers to the case where the constraint matrices are low rank or sparse.

constraint the set Ω_2 becomes equivalent to the set Ω_1, and thus (4.14) is no more a relaxation but an equivalent problem of (4.12). The iterative method used in this paper is different from those used in Dattorro [2008] and Fazel et al. [2003]. The proposed method approximately solves a max-min optimization problem, while that in Dattorro [2008] solves the feasible problem with a rank constraint and thus is not applicable for MMDA. The log-det heuristic in Fazel et al. [2003] is similar to what we have done in the local convex relaxation, but they are intrinsically different. It is utilized as the objective to encourage low rank solution in Fazel et al. [2003] while it serves as a local constraint in this paper. Besides, it is worth emphasizing that both methods Dattorro [2008] and Fazel et al. [2003], and the proposed method in this paper, cannot guarantee a global optimal solution in general cases. Indeed, the applying of SDP relaxation to rank constrained optimization is still under investigation.

4.4 Empirical Evaluations

4.4.1 Experiments on Synthetic Datasets

In this section, we conduct statistical experiments on synthetic datasets to demonstrate the effectiveness of MMDA.

4.4.1.1 Data Generation and Evaluation Methods

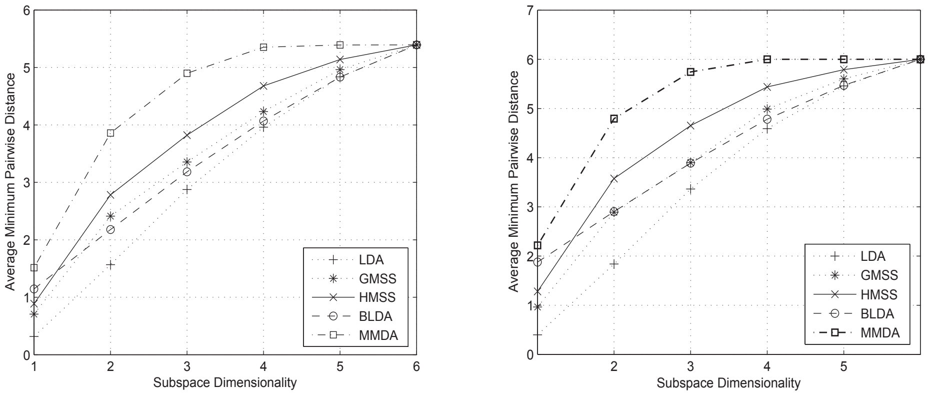
Two sets of synthetic data are generated. For the first set, we consider a 7-class classification problem represented by homoscedastic Gaussian distributions in $\mathbb{R}^{10}$. The common covariance matrix is set as $\mathbf{I}_{10}$, while the class means are randomly sampled from a 10-dimensional Gaussian distribution with zero mean and a covariance matrix $2\mathbf{I}_{10}$. We sample the class means 500 times, and for each time of their realizations we generate 200 samples for each of the 7 classes, 100 for training and the rest 100 for test. Thus, we have 500 independent groups of training and test samples. We refer to this dataset as the *uniformly distributed dataset*. For the second dataset, we use the same procedure to generate the data, except that when we sample the means of the 7 classes from the Gaussian distribution we add a bias of 10 to the first dimension of the means of the first three classes. Such bias enforces the first three classes to be distant from the rest four, which is useful to test whether the subspace selection methods will be affected by the nonuniform distribution of classes. And we refer to this dataset as the *nonuniformly distributed dataset*.

We compared MMDA against four other methods, i.e., the conventional LDA, and the recently developed GMSS Tao et al. [2009], HMSS Bian and Tao [2008], and BLDA Hamsici and Martinez [2008]. In all experiments, we first select the subspace with varying dimensionality from 1 to 6 by performing different methods on the training dataset, and then do classification on the test dataset by using the nearest mean (NM) classifier on the dimension reduced subspace. We evaluate the performance of different methods from three aspects: 1) the separation of all class pairs, i.e, the minimum pairwise distance in the selected low dimensional subspace; 2) the average classification error rate with the standard deviation; and 3) a 2-dimensional graphical representation of data distribution.

4.4.1.2 Results and Analyses

The minimum pairwise distances in the low dimensional subspace obtained by different methods are averaged over the 500 independent trials, and shown in Figure 4.4. It can be observed that, when the dimensionality is less than the class number (7), the minimum pairwise distances achieved by different methods are different. Specifically, LDA performs the worst, which gives the smallest minimum pairwise distance. GMSS and BLDA are comparable to each other, HMSS is relatively better, and MMDA performs the best. In addition, GMSS is still readily affected by the distributions of classes in the original space. For the uniformly distributed dataset, GMSS is superior to BLDA on most dimensions. When the classes are nonuniformly distributed by added bias, then the performance of GMSS degrades to be nearly overlapped with BLDA. Besides, for 1-dimensional subspace, the minimum pairwise distance achieved by the Bayes optimal method BLDA is larger than GMSS and HMSS. This implies that lower classification error generally requires larger minimum pairwise distance (though not necessarily maximized).

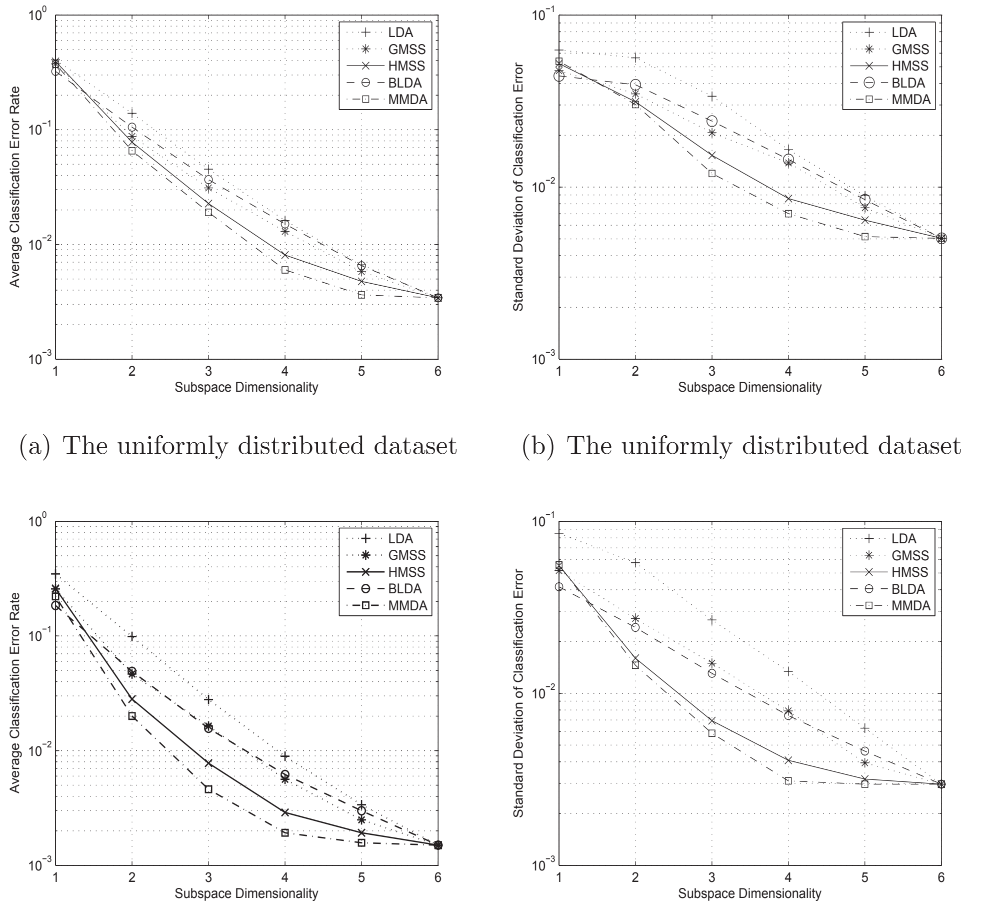
The test results by different methods on the two synthetic datasets are sum-



(a) The uniformly distributed dataset (b) The nonuniformly distributed dataset

Figure 4.4: Evaluation on synthetic datasets by minimum pairwise distance.

marized in Figure 4.5, including the average classification error rate and the standard deviation. To get a clear view of the differences among different methods, we plot these results in the log scale. In the 6-dimensional case, all methods perform equally, this is consistent with the fact that the c-dimensional subspace selected by LDA contains the entire discrimination power for a $c + 1$-class problem. However, when the dimensionality becomes lower, LDA performs worse than all other methods. One can see that except the 1-dimensional case, MMDA has the most competitive performance, with GMSS, HMSS and BLDA performing moderately. In the 1-dimensional case, BLDA performs the best since it utilizes Bayes optimal criterion. However, when dimensionality increases, its performance degrades, because it utilizes a greedy method to construct high-dimensional subspaces and thus is no more Bayes optimal.

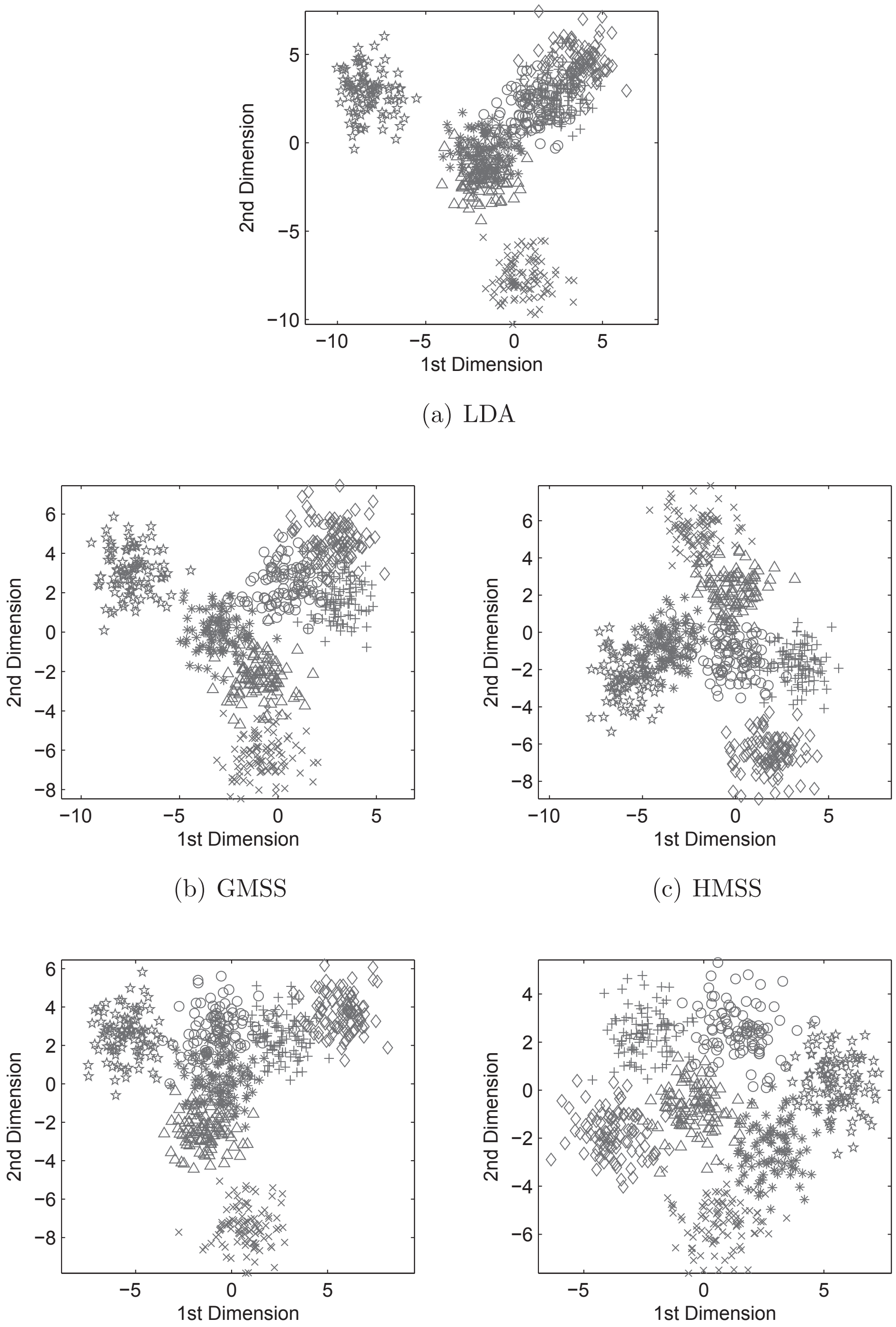


(a) The uniformly distributed dataset (b) The uniformly distributed dataset

(c) The nonuniformly distributed dataset (d) The nonuniformly distributed dataset

Figure 4.5: Evaluation on synthetic datasets by the average classification error rate and the standard deviation.

We randomly select one group of training data from each of the two synthetic datasets, and used them to demonstrate the capability of different methods in selecting a 2-dimensional subspace for the graphical representation of data. The corresponding results are shown by Figure 4.6 and Figure 4.7, respectively. From these graphs, one can see that LDA is unable to separate all classes, while GMSS, HMSS and BLDA only give improved separation. However, in both cases, MMDA clearly separates all classes. Besides, comparing Figure 4.6 and Figure 4.6, one can see that that LDA, GMSS and BLDA are more likely to be affected by the nonuniform distribution of classes, while HMSS and MMDA are more robust.



(a) LDA

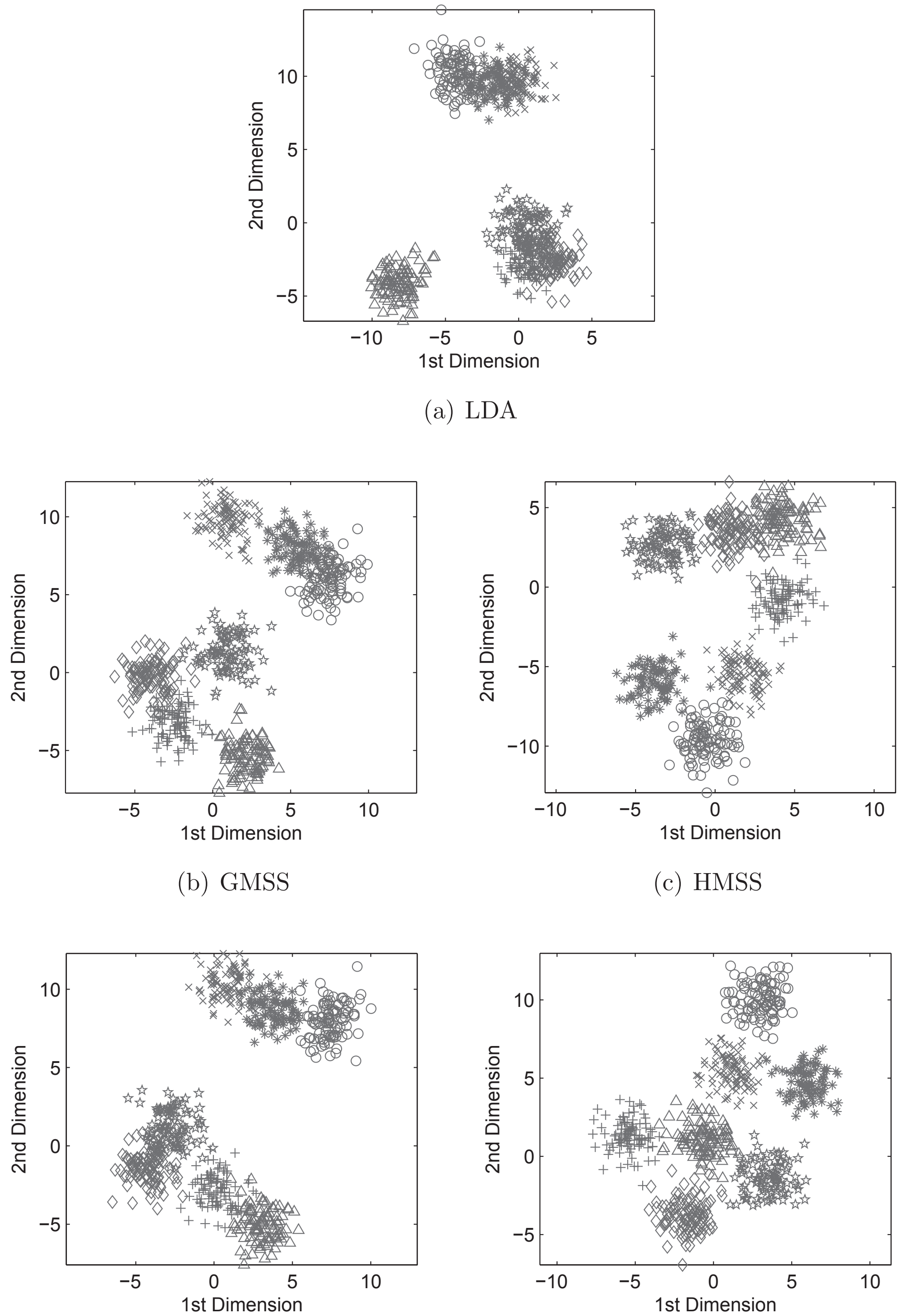
(b) GMSS

(c) HMSS

(d) BLDA

(e) MMDA

Figure 4.6: 2-Dimensional data representation on the uniformly distributed dataset.



(a) LDA

(b) GMSS

(c) HMSS

(d) BLDA

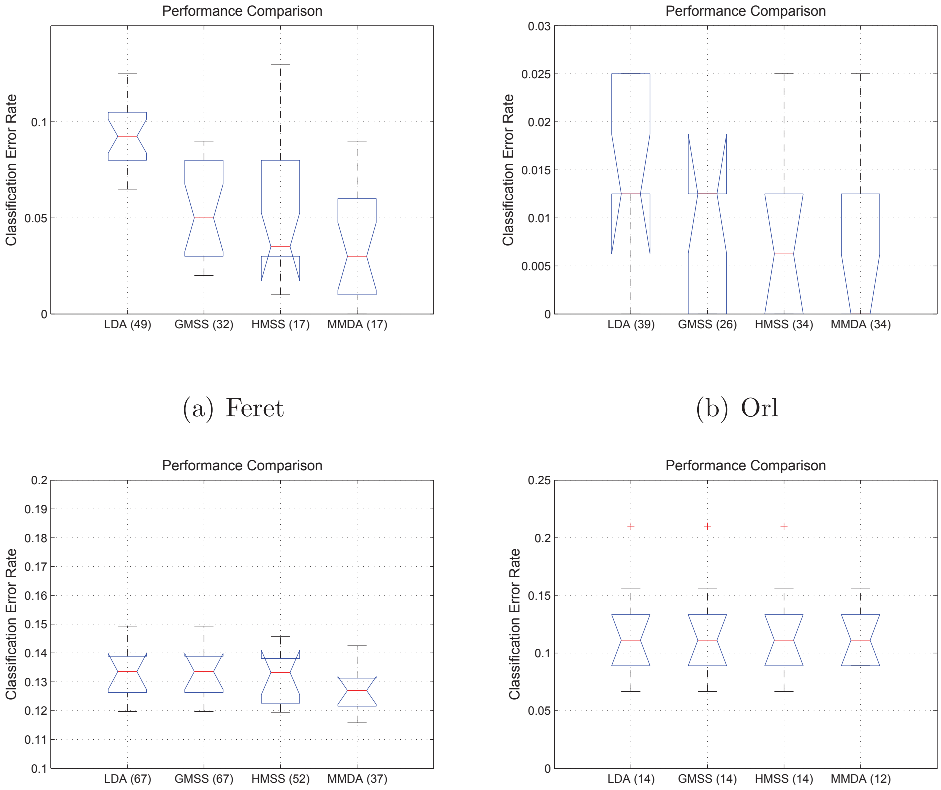
(e) MMDA

Figure 4.7: 2-Dimensional data representation on the uniformly distributed dataset.

4.4.2 Experiments on Face Recognition

We apply MMDA to face recognition and compare it against LDA, GMSS and HMSS. We do not perform BLDA in this experiments, as it needs too many computational costs due to the enumeration of the order of different class centers' locations. Here, we use the same four benckmark datasets as used Chapter 3, i.e., "Feret", "Orl", "Pie", and "Yale", and adopt the same experimental setting. For a brief description of these datasets and the experimental setting, please refer to Section 3.3.

On each dataset, we first perform LDA with the block-diagonal regularization (using the deterministic variable partition), and use it as the baseline method for comparison. Then, we perform GMSS, HMSS, and MMDA, for further dimension reduction. The performance of different methods are shown by Figure 4.8, where the best dimensionality for each method is tuned according to the average performance over 20 random trials. From these results, we have two observations: 1) though all based on homoscedastic Gaussian assumption, GMSS, HMSS and MMDA show improved performance over LDA; 2) MMDA achieves the best performance because it duly guarantees the separation of all classes.



(a) Feret

(b) Orl

(c) Pie

(d) Yale

Figure 4.8: Evaluation on face recognition experiments

Chapter 5

Minimizing Asymptotic Nearest Neighbor Classification Error for Nonparametric SLDR

5.1 Introduction

In previous chapters, all the SLDR methods studied, including LDA Fisher [1936] Rao [1948] and its extensions, e.g., aPAC Loog et al. [2001], GMSS Tao et al. [2009], HMSS Bian and Tao [2008] and MMDA, assume data are sampled from homoscedastic Gaussian distributions. Certainly, as we have already seen, such assumption provides considerable feasibility for both theoretical analysis and algorithmic design. However, in practice, real world data are usually distributed by more general distributions. If homoscedastic Gaussian distributions fit the data well, though not exactly, above mentioned methods are still preferable and may have acceptable performance. This is because simple models generally have better generalization ability with finite training samples. But when the fitness is highly invalid, these methods will not be applicable any more.

One popular way to extent above methods is based on the heteroscedastic Gaussian assumption, i.e., allowing different classes of data to have distinct covariance matrices. A number of SLDR methods have been proposed in this direction, using different approaches to measure the discrimination power among heteroscedastic Gaussian distributions. For example, Decell and Mayekar [1977], De la Torre and Kanade [2005], and Tao et al. [2009] use the Kullback-Leibler divergence as the measurement, Loog and Duin [2004] and Saon and Padmanabhan [2001] use the Chernoff distance or the Bhattacharyya bound, and Nenadic [2007] defines the so-called μ-measure for the same purpose.

In order to deal with more general data distributions, nonparametric method has been introduced to SLDR. Fukunaga and Mantock [1983] proposed the first nonparametric SLDR method, called nonparametric discriminant analysis (NDA). Maximizing mutual information (MMI) is another nonparametric SLDR method proposed in Torkkola [2003]. In contrast to the heuristic treatment in NDA, MMI follows a more principal approach. First, it uses the kernel method to estimate the probability density of each class, and then it optimizes the projection matrix by maximizing the mutual information between class labels and the transformed samples in the dimension reduced space. The motivation behind MMI is that maximizing the mutual information is equivalent to minimizing the conditional entropy, and the latter can be deemed as a proxy of the Bayes optimal criterion since the conditional entropy provides an upper bound of Bayes error.

This chapter proposes a new method for nonparametric SLDR, which optimizes the projection matrix by minimizing the asymptotic nearest neighbor classification error (MNNE). According to Cover and Hart [1967], the asymptotic nearest neighbor classification error (briefly NN error) upper bounds Bayes error by a factor of at most 2. Therefore, MNNE can be regarded as a proxy of the Bayes optimal criterion for SLDR. One of our contributions is that we prove MNNE is superior to MMI and minimizing the Bhattacharyya in terms of the closeness to the the Bayes optimal criterion. We derive an algorithm for MNNE, based on kernel density estimation and a gradient descent method on the Grassmann manifold. Empirical evaluations on real datasets show the promising performance of MNNE compared with to the state-of-the-art nonparametric SLDR methods.

The rest of this chapter is organized as follows. In Section 5.2, we present MNNE and show its superiority as a proxy of the Bayes optimal criterion. In Section 5.3, we derive an algorithm for MNNE. Section 5.4 reports experimental results on real datasets.

5.2 Minimizing NN Error as a Proxy of the Bayes Optimal Criterion

Suppose the joint probability distribution of the problem to be study is $p(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^D$ and $y \in \{1, 2, ..., c\}$. The conditional density of the i-th class is given by $p_i(\mathbf{x}) \triangleq p(\mathbf{x}|y = i)$, and the corresponding prior probability is given by $\pi_i \triangleq p(y = i)$. We intend to learn a projection matrix $\mathbf{W} \in \mathbb{R}^{D\times d}$, $d < D$, such that after the transform $\mathbf{z} = \mathbf{W}^T\mathbf{x}$, the conditional densities, $p(\mathbf{z}|y = i)$, $i = 1, 2, ..., c$, are well separated from one another. For convenience, we still use $p_i(\mathbf{z})$ to denote $p(\mathbf{z}|y = i)$. Then, the posterior probability of y after observing $\mathbf{z}$ is given by

$$\eta_i(\mathbf{z}) \triangleq p(y = i|\ \mathbf{z}) = \frac{\pi_i p_i(\mathbf{z})}{p(\mathbf{z})}, \tag{5.1}$$

where

$$p(\mathbf{z}) = \sum_{i=1}^{m} \pi_i p_i(\mathbf{z}) \tag{5.2}$$

is the marginal density.

According to Cover and Hart [1967], the NN error on $p(\mathbf{z}|y = i)$, $i = 1, 2, ..., c$, is given by

$$P_{nn}(\mathbf{W}) = 1 - \sum_{i=1}^{c} \mathbb{E}\eta_i^2(\mathbf{z}), \tag{5.3}$$

where the expectation is taken with respect to $p(\mathbf{z})$. Note that the NN error $P_{nn}(\mathbf{W})$ is a function of the projection matrix $\mathbf{W}$. It can be shown that $P_{nn}(\mathbf{W})$ provides a lower and an upper bound of Bayes error Cover and Hart [1967]

$$P^* \triangleq 1 - \mathbb{E}\left(\max_i \eta_i(\mathbf{z})\right), \tag{5.4}$$

i.e.,

$$P^* \leq P_{nn} \leq P^* \left(2 - \frac{c}{c-1} P^*\right). \tag{5.5}$$

Thus, we can optimize $\mathbf{W}$ by minimizing $P_{nn}(\mathbf{W})$, i.e., treating NN error minimization as a proxy of the Bayes optimal criterion.

5.2.1 Compared to Mutual Information

Information theory plays an important role in statistical pattern recognition For SLDR, Torkkola [2003] proposed the maximizing mutual information criterion (MMI), which optimizes the projection matrix $\mathbf{W}$ by maximizing the mutual information $I(Y;Z)$ between the class label y and the dimension reduced sample $\mathbf{z}$.

According Cover and Thomas [1991], the mutual information can be decomposed into

$$I(Y;Z) = H(Y) - H(Y|Z), \tag{5.6}$$

wherein $H(Y)$ is the entropy of the prior probability $p(y = i)$, and

$$H(Y|Z) = -\int p(\mathbf{z}) \left(\sum_y p(y|\ \mathbf{z}) \log \left(p(y|\ \mathbf{z}) \right) \right) d\mathbf{z} \tag{5.7}$$

is the (averaged) conditional entropy of y given $\mathbf{z}$. Thus, maximizing the mutual information $I(Y;Z)$ is equivalent to minimizing the conditional entropy $H(Y|Z)$. It has been shown that $H(Y|Z)$ is upper and lower bounded by Bayes error P^* Hellman and Raviv [1970] Fano [1961], i.e.,

$$P^* \leq \frac{1}{2}H(Y|Z) \leq \frac{1}{2}H(P^*) + \frac{1}{2}P^* \log(m-1), \tag{5.8}$$

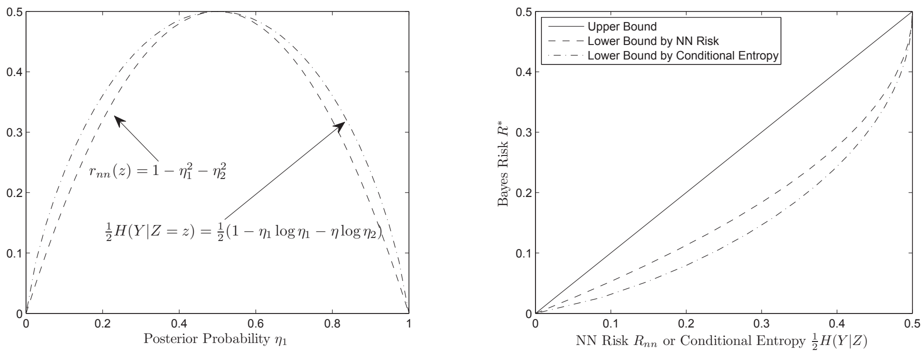
where the righthand side inequality is known as the Fano's inequality Fano [1961], and $H(P^*) = -P^* \log(P^*) - (1-P^*)\log(1-P^*)$ is the entropy of P^*. Thus, (5.6) and (5.8) suggest that maximizing mutual information $I(Y;Z)$ is equivalent to minimizing an upper bound of Bayes error P^*.

From above discussions, we see that both P_{nn} and $\frac{1}{2}H(|Z)$ provide an upper bound of P^*. However, by the following Theorem 5.1, we show that the bound provided by P_{nn} is tighter.

Theorem 5.1. *Given P^*, P_{nn}, and $\frac{1}{2}H(Y|Z)$, the following inequalities hold*

$$P^* \leq P_{nn} \leq \frac{1}{2}H(Y|Z). \tag{5.9}$$

Proof. The lefthand side inequality in (5.9) is due to the optimality of Bayes



(a) (b)

Figure 5.1: (a) The conditional NN error $r_{nn}(z)$ and half the posterior entropy $\frac{1}{2}H(Y|Z=z)$ in binary classification. (b) Upper and lower bound of Bayes error P^* by the NN error P_{nn} and half the conditional entropy $\frac{1}{2}H(Y|Z)$ in binary classification.

error, and thus we only need to prove the righthand side inequality. According to (5.7), the conditional entropy $H(Y|Z)$ can be rewritten as

$$H(Y|Z) = \mathbb{E}\left(H(Y|Z=\mathbf{z})\right). \tag{5.10}$$

By (5.10) and (5.3), it is sufficient to show that

$$1 - \sum_{i=1}^{c} \eta_i^2(\mathbf{z}) \leq \frac{1}{2}H(Y|Z=\mathbf{z}). \tag{5.11}$$

We prove (5.11) using mathematical induction. When $c=2$, it holds

$$1 - \eta_1^2 - \eta_2^2 \leq \frac{1}{2}(-\eta_1 \log \eta_1 - \eta_2 \log \eta_2). \tag{5.12}$$

Figure 5.1 (a) gives an illustration of (5.12), and the rigorous proof is given in Section 5.5. Next, we suppose (5.11) holds for $c=m$ and prove the case for $c=m+1$.

According to Ash [1965], it holds

$$\begin{aligned} H(\eta_1, \eta_2, ..., \eta_m, \eta_{m+1}) = & H(\eta_1, \eta_2, ..., \eta_{m-1}, \eta_m + \eta_{m+1}) \\ & + (\eta_m + \eta_{m+1}) H\left(\frac{\eta_m}{\eta_m + \eta_{m+1}}, \frac{\eta_{m+1}}{\eta_m + \eta_{m+1}}\right). \end{aligned} \tag{5.13}$$

By combining (5.11) and (5.13), we have

$$\begin{aligned} & 1 - \eta_1^2 - \eta_2^2 -, ..., -\eta_c^2 + \eta_{c+1}^2 \\ = & 1 - \eta_1^2 - \eta_2^2 -, ..., -(\eta_c + \eta_{c+1})^2 + 2\eta_m\eta_{m+1} \\ \leq & \frac{1}{2} H(\eta_1, \eta_2, ..., \eta_{c-1}, \eta_m + \eta_{m+1}) + 2\eta_m\eta_{m+1} \\ = & \frac{1}{2} H(\eta_1, \eta_2, ..., \eta_m, \eta_{m+1}) \\ & \qquad - \frac{1}{2}(\eta_m + \eta_{m+1}) H\left(\frac{\eta_m}{\eta_m + \eta_{m+1}}, \frac{\eta_{m+1}}{\eta_m + \eta_{m+1}}\right) + 2\eta_m\eta_{m+1}. \end{aligned} \tag{5.14}$$

Since $\eta_m + \eta_{m+1} \leq 1$, we have

$$\begin{aligned} 2\eta_m\eta_{m+1} \leq & \frac{2\eta_m\eta_{m+1}}{\eta_m + \eta_{m+1}} \\ = & (\eta_m + \eta_{m+1}) \left(2\frac{\eta_m}{\eta_m + \eta_{m+1}} \frac{\eta_{m+1}}{\eta_m + \eta_{m+1}}\right) \\ = & (\eta_m + \eta_{m+1}) \left(1 - \left(\frac{\eta_m}{\eta_m + \eta_{m+1}}\right)^2 - \left(\frac{\eta_{m+1}}{\eta_c y + \eta_{m+1}}\right)^2\right) \\ \leq & \frac{1}{2}(\eta_m + \eta_{m+1}) H\left(\frac{\eta_m}{\eta_c y + \eta_{m+1}}, \frac{\eta_{m+1}}{\eta_m + \eta_{m+1}}\right), \end{aligned} \tag{5.15}$$

where the last inequality is obtained by treating $\frac{\eta_m}{\eta_m + \eta_{m+1}}$ and $\frac{\eta_{m+1}}{\eta_m + \eta_{m+1}}$ as the posterior probabilities in case of $c = 2$ and applying (5.12). Finally, (5.14) and (5.15) give

$$1 - \eta_1^2 - \eta_2^2 -, ..., -\eta_m^2 - \eta_{m+1}^2 \leq \frac{1}{2} H(\eta_1, \eta_2, ..., \eta_m, \eta_{m+1}). \tag{5.16}$$

□

We have proved that P_{nn} provides a tighter upper bound of P^* than $\frac{1}{2}H(Y|Z)$.

Reversely, by taking binary classification for example, we show that the uncertainty of P^* given P_{nn} is smaller than given $\frac{1}{2}H(Y|Z)$. As P^* cannot exceed P_{nn} and $\frac{1}{2}H(Y|Z)$, we are interested in an lower bound of P^*. To this end, we convert[1] (5.5) and (5.8) to lower bounds of P^*, and plot them as functions of P_{nn} and $\frac{1}{2}H(Y|Z)$ in Figure 5.1 (b). One can see that given P_{nn} the largest uncertainty of P^* (i.e., the largest deviation from the upper bound to the lower bound of P^*) is 0.123, while the largest uncertainty of P^* is 0.157 given $\frac{1}{2}H(Y|Z)$.

5.2.2 Compared to Bhattacharyya Bound

For binary classes, the Bhattacharyya Bound is defined by

$$B = \sqrt{\pi_1 \pi_2} \int \sqrt{p_1(\mathbf{z}) p_2(\mathbf{z})} d\mathbf{z}. \tag{5.17}$$

In multiple classes problem, it can be extended to Saon and Padmanabhan [2001]

$$B = \sum_{1 \le i < j \le c} \sqrt{\pi_i \pi_j} \int \sqrt{p_i(\mathbf{z}) p_j(\mathbf{z})} d\mathbf{z}. \tag{5.18}$$

It has been proved that the Bhattacharyya Bound B in (5.18) provides an upper bound for the Bayes error P^*. By the following Theorem 5.2, we show that P_{nn} is tighter than B in bounding P^*.

Theorem 5.2. *Given P^*, P_{nn} and B, the following inequality holds*

$$P^* \le P_{nn} \le B. \tag{5.19}$$

[1] Since the Fano's inequality cannot be convert to lower bound of P^* analytically, we used numerical method.

Proof. We only need to show $P_{nn} \leq B$. By (5.18), we have

$$\begin{aligned} B &= \sum_{1 \leq i < j \leq c} \sqrt{\pi_i \pi_j} \int \sqrt{p_i(\mathbf{z}) p_j(\mathbf{z})} d\mathbf{z} \\ &= \int \sum_{1 \leq i < j \leq c} \sqrt{\frac{\pi_i \pi_j p_i(\mathbf{z}) p_j(\mathbf{z})}{p(\mathbf{z}) p(\mathbf{z})}} p(\mathbf{z}) d\mathbf{z} \\ &= \mathbb{E}\left(\sum_{1 \leq i < j \leq c} \sqrt{\eta_i(\mathbf{z}) \eta_j(\mathbf{z})} \right). \end{aligned} \tag{5.20}$$

Further, by (5.3), we have

$$P_{nn} = \mathbb{E}\left(1 - \sum_{i=1}^{c} \eta_i^2(\mathbf{z})\right) = \mathbb{E}\left(2 \sum_{1 \leq i < j \leq c} \eta_i(\mathbf{z}) \eta_j(\mathbf{z})\right). \tag{5.21}$$

Thus, it is sufficient to show

$$\sum_{1 \leq i < j \leq c} \sqrt{\eta_i(\mathbf{z}) \eta_j(\mathbf{z})} \geq 2 \sum_{1 \leq i < j \leq c} \eta_i(\mathbf{z}) \eta_j(\mathbf{z}). \tag{5.22}$$

Actually, this holds according to the following arguments.

$$\begin{aligned} &\frac{\sum_{1 \leq i < j \leq c} \eta_i(\mathbf{z}) \eta_j(\mathbf{z})}{\sum_{1 \leq i < j \leq c} \sqrt{\eta_i(\mathbf{z}) \eta_j(\mathbf{z})}} \leq \max_{1 \leq i < j \leq c} \frac{\eta_i(\mathbf{z}) \eta_j(\mathbf{z})}{\sqrt{\eta_i(\mathbf{z}) \eta_j(\mathbf{z})}} \\ &= \max_{1 \leq i < j \leq m} \sqrt{\eta_i(\mathbf{z}) \eta_j(\mathbf{z})} \leq \max_{1 \leq i < j \leq c} \frac{\eta_i(\mathbf{z}) + \eta_j(\mathbf{z})}{2} \\ &\leq \max_{1 \leq i < j \leq c} \frac{1}{2} = \frac{1}{2}. \end{aligned} \tag{5.23}$$

□

5.3 Algorithm

In this section, we derive an algorithm for MNNE, i.e., minimizing the NN error P_{nn} (5.3). First, we rewrite P_{nn} as below

$$\begin{aligned} P_{nn}(\mathbf{W}) =& 1 - \sum_{i=1}^{c} \mathbb{E}\eta_i^2(\mathbf{z}) = 1 - \sum_{i=1}^{c} \int \eta_i^2(\mathbf{z})p(\mathbf{z})d\mathbf{z} \\ =& 1 - \sum_{i=1}^{c} \int \eta_i(\mathbf{z}) \frac{p(\mathbf{z}, y = i)}{p(\mathbf{z})} p(\mathbf{z})d\mathbf{z} \\ =& 1 - \sum_{i=1}^{c} \int \eta_i(\mathbf{z})p(\mathbf{z}, y = i)d\mathbf{z} \\ =& 1 - \mathbb{E}_{p(\mathbf{z},y)}\eta_y(\mathbf{z}). \end{aligned} \tag{5.24}$$

Accordingly, MNNE can be formulated as the optimization below,

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_d} J(\mathbf{W}) = \mathbb{E}_{p(\mathbf{z},y)}\eta_y(\mathbf{z}). \tag{5.25}$$

where the orthonormal constraints ensures $\mathbf{W}$ not to be degenerated.

5.3.1 Kernel Density Estimation

In practice, we have no access to the true joint distribution $p(\mathbf{z}, y)$ and the posterior probability $\eta_i(\mathbf{z})$. Thus, to implement MNNE (5.25), we have to estimate $J(\mathbf{W})$ empirically. Suppose we have a training dataset $\mathcal{D} = \{(\mathbf{x}_j, y_j)|j = 1, 2, ..., n\}$. With projection matrix $\mathbf{W}$, we have $\mathbf{z}_j = \mathbf{W}^T\mathbf{x}_j$, $j = 1, 2..., n$. By using the plug-in method, an empirical estimate of $J(\mathbf{W})$ is given by

$$\hat{J}(\mathbf{W}) = \frac{1}{n}\sum_{j=1}^{n} \hat{\eta}_{y_j}(\mathbf{z}_j) = \frac{1}{n}\sum_{j=1}^{n} \frac{\hat{p}(\mathbf{z}_j|y = y_j)\hat{p}(y = y_j)}{\hat{p}(\mathbf{z}_j)}. \tag{5.26}$$

In (5.26), $\hat{p}(y = y_j)$ is easy to be estimated by using frequency of each class in the training dataset, i.e.,

$$\hat{p}(y = y_j) = \frac{n_{y_j}}{n}, \tag{5.27}$$

where n_{y_j} is the number of training samples in the class that $(\mathbf{x}_j, y_j)$ belongs to.

For $\hat{p}(\mathbf{z}_j|y = y_j)$ and $\hat{p}(\mathbf{z}_j)$, we apply kernel based density estimation. Specifically, we use the Gaussian kernel

$$K_h(\mathbf{u}) = h^{-d}(2\pi)^{-d/2} \exp\left\{-\frac{1}{2h^2}\mathbf{u}^T\mathbf{u}\right\}, \ \mathbf{u} \in \mathbb{R}^d, \tag{5.28}$$

where h is a bandwidth to be determined. Using the leave-one-out principle, we have the following kernel based estimates Parzen [1962]

$$\hat{p}(\mathbf{z}_j|y = y_j) = \frac{1}{n_{y_j} - 1} \sum_{y_k = y_j, k \neq j} K_h(\mathbf{z}_j - \mathbf{z}_k), \tag{5.29}$$

$$\hat{p}(\mathbf{z}_j) = \frac{1}{n - 1} \sum_{k \neq j} K_h(\mathbf{z}_j - \mathbf{z}_k). \tag{5.30}$$

Now, two problems still remain: first, how to select the bandwidth h; second, how to minimize $\hat{J}(\mathbf{W})$ under the orthonormal constraints $\mathbf{W}^T\mathbf{W} = \mathbf{I}_d$. We will address these two problem in following subsections.

5.3.2 Bandwidth Selection

We derive a convenient formula to select the bandwidth h used in the estimates given in (5.29) and (5.30). In particular, we use the asymptotic mean squared error (AMISE) criterion Silverman [1986], which trades the squared bias off the variance of the estimator,

$$AMISE(\mathbf{H}) = \frac{1}{4}\mu_2^2(K) \int [\text{tr}\{\mathbf{H}^T H_p(\mathbf{u})\mathbf{H}\}]^2 d\mathbf{u} + \frac{1}{n\text{det}(\mathbf{H})}\|K\|_2^2, \tag{5.31}$$

where $\mathbf{H}_p(\mathbf{u})$ is the Hessian of the density $p(\mathbf{u})$ to be estimated, K is the kernel function, $\mathbf{H}$ is the bandwidth matrix, and n is the sample size. In our case, K is the Gaussian kernel and the bandwidth matrix is $\mathbf{H} = h\mathbf{I}_d$. Since the true density $p(\mathbf{z}|y)$ is unknown, we adopt the commonly used rule-of-thumb Silverman [1986] to replace the unknown true density by a reference density $q(\mathbf{z})$, e.g., a Gaussian distribution with its covariance matrix equal to the sample covariance. To simplify calculation, we conduct whitening preprocessing before applying MNNE so that the sample covariance is identity matrix, and thus $q(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Based on above discussions, AMISE (5.31) for the conditional density $p(\mathbf{z}|y = i)$ can be simplified as,

$$AMISE(h) = \frac{2d + d^2}{2^{d+4}\pi^{d/2}} h^4 + \frac{1}{2^d \pi^{d/2} n_i h^d}, \tag{5.32}$$

where n_i is the training sample number of the i-th class. Taking the derivative with respect to h and setting it as 0, we get the optimal h_{opt} that

$$h_{opt} = \left(\frac{4}{d+2}\right)^{1/d+4} n_i^{-1/(d+4)}. \tag{5.33}$$

Moreover, since the same bandwidth is used for all conditional densities $p(\mathbf{z}|y = i)$, $i = 1, 2, ..., c$, we further replace n_i in (5.33) by n/c to balance the training sample number of different classes, which gives the final bandwidth

$$h_{opt} = \left(\frac{4m}{(d+2)n}\right)^{1/d+4}. \tag{5.34}$$

Note that 5.34 is obtained only from a theoretical viewpoint and involves many approximations. For better performance, one can use h_{opt} as an initial guess and find the best bandwidth for the problem at hand by using cross-validation around h_{opt}. However, in all experiments in this chapter, we found that h_{opt} works favorably.

5.3.3 Optimization on the Grassmann Manifold

Empirical, MNNE needs to solve the following optimiazation

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_d} \hat{J}(\mathbf{W}) = \frac{1}{n}\sum_{j=1}^{n} \frac{\hat{p}(\mathbf{z}_j|y = y_j)\hat{p}(y = y_j)}{\hat{p}(\mathbf{z}_j)}. \tag{5.35}$$

A basic algorithm for solving (5.35) is the gradient descent method combined with projection onto the feasible set $\mathbf{W}^T\mathbf{W} = \mathbf{I}_d$. Specifically, in each iteration, we first update

$$\mathbf{W}_{k+1} \leftarrow \mathbf{W}_k + \tau \nabla \hat{J}(\mathbf{W})\big|_{\mathbf{W}=\mathbf{W}_k}, \tag{5.36}$$

where τ is the learning rate, and then conduct Gram-Schmidt orthogonalization on $\mathbf{W}_{k+1}$. Such algorithm is commonly used in many works Tao et al. [2009] Torkkola [2003]. However, it has a key drawback that the objective value $\hat{J}(\mathbf{W}_k)$ is not necessarily monotonically decreasing due to the orthogonalization.

Note that the orthonormal constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}_d$ actually defines a Grassmann manifold $\mathcal{G}_{D,d}$ Edelman et al. [1998]. If we perform gradient descent on $\mathcal{G}_{D,d}$, then the orthonormal constraint can be automatically satisfied. Suppose the current solution from the k-th iteration is $\mathbf{W}_k$, and the corresponding gradient $\nabla J(\mathbf{W})\big|_{\mathbf{W}=\mathbf{W}_k}$ is given by

$$\begin{aligned}\nabla J(\mathbf{W})\big|_{\mathbf{W}=\mathbf{W}_k} = \frac{1}{n}\sum_{j=1}^{n}\frac{\hat{p}(y=y_j)}{\hat{p}(\mathbf{z}_j)}\nabla\hat{p}(\mathbf{z}_j|y=y_j)\big|_{\mathbf{W}=\mathbf{W}_k}\\ -\frac{1}{n}\sum_{j=1}^{n}\frac{\hat{p}(y=y_j)\hat{p}(\mathbf{z}_j|y=y_j)}{\hat{p}^2(\mathbf{z}_j)}\nabla\hat{p}(\mathbf{z}_j)\big|_{\mathbf{W}=\mathbf{W}_k},\end{aligned} \tag{5.37}$$

where

$$\begin{aligned}&\nabla\hat{p}(\mathbf{z}_j|y=y_j)\big|_{\mathbf{W}=\mathbf{W}_k}\\ &=\frac{1}{(n_{y_j}-1)h^2}\sum_{y_k=y_j,k\neq j}K_h(\mathbf{z}_j-\mathbf{z}_k)(\mathbf{z}_j-\mathbf{z}_k)(\mathbf{z}_j-\mathbf{z}_k)^T\mathbf{W}_k,\end{aligned} \tag{5.38}$$

and

$$\nabla\hat{p}(\mathbf{z}_j)\big|_{\mathbf{W}=\mathbf{W}_k} = \frac{1}{(n-1)h^2}\sum_{k\neq j}K_h(\mathbf{z}_j-\mathbf{z}_k)(\mathbf{z}_j-\mathbf{z}_k)(\mathbf{z}_j-\mathbf{z}_k)^T\mathbf{W}_k. \tag{5.39}$$

Instead of performing line search in the direction of $-\nabla J(\mathbf{W})\big|_{\mathbf{W}=\mathbf{W}_k}$, we first projection $\nabla J(\mathbf{W})\big|_{\mathbf{W}=\mathbf{W}_k}$ onto $\mathcal{G}_{D,d}$, which is given by Edelman et al. [1998]

$$\mathbf{T}_k = \nabla\hat{J}(\mathbf{W})\big|_{\mathbf{W}=\mathbf{W}_k} - \mathbf{W}_k\mathbf{W}_k^T\nabla\hat{J}(\mathbf{W})\big|_{\mathbf{W}=\mathbf{W}_k}, \tag{5.40}$$

and then, perform search along the geodetic determined by $\mathbf{W}_k$ and $\mathbf{T}_k$ on $\mathcal{G}_{D,d}$. According to Edelman et al. [1998], the geodetic is given by

$$g(t) = \mathbf{W}_k\mathbf{V}\cos(\mathbf{\Sigma}t)\mathbf{V}^T + \mathbf{U}\sin(\mathbf{\Sigma}t)\mathbf{V}^T, \tag{5.41}$$

where $\mathbf{U\Sigma V}^T$ is the compact singular value decomposition (SVD) of $\mathbf{T}_k$. Suppose the minimum point of (5.41) is t_{min}, we can update the projection matrix by $\mathbf{W}_{k+1} = g(t_{min})$.

Algorithm 2 summarizes the pseudo code of the Grassmann manifold based gradient descent method.

Algorithm 2 Grassmann Manifold based Gradient Descent Method for MNNE

```
Initialization: $\mathbf{W}_0$ such that $\mathbf{W}_0^T\mathbf{W}_0 = \mathbf{I}_d$.  , $\mathbf{T}_0 = \mathbf{G}_0 - \mathbf{W}_0\mathbf{W}_0^T\mathbf{G}_0$, and $\mathbf{H}_0 = -\mathbf{G}_0$.
for k=0,2,... do
    Step. 1 Calculate gradient $\nabla\hat{J}(\mathbf{W})\big|_{\mathbf{W}=\mathbf{W}_k}$ by (5.37) and its projection $\mathbf{T}_k$ by (5.40).
    Step. 2 Perform compact SVD $\mathbf{T}_k = \mathbf{U\Sigma V}$.
    Step. 3 Minimize $\hat{J}(g(t))$ over $t$, where
                $g(t) = \mathbf{W}_k\mathbf{V}\cos(\mathbf{\Sigma}t)\mathbf{V}^T + \mathbf{U}\sin(\mathbf{\Sigma}t)\mathbf{V}^T.$
    Step. 4 Update $\mathbf{W}_{k+1} = g(t_{min})$.
    Step. 5 Stop if $|\hat{J}(\mathbf{W}_{k+1} - \hat{J}(\mathbf{W}_k)|/|\hat{J}(\mathbf{W}_k)| < \epsilon$.
end for
```

5.4 Empirical Evaluations

In this section, we evaluate the performance of MNNE and compare it with other nonparametric SLDR methods.

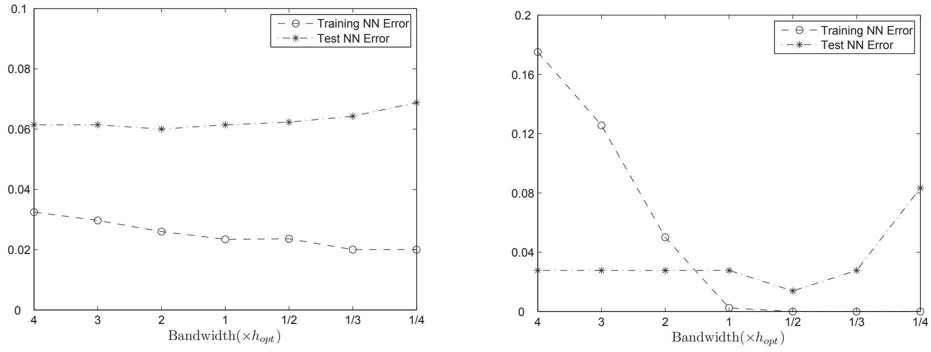
5.4.1 Bandwidth Selection

We evaluation the proposed bandwidth selection method for MNNE on two datasets, “BreastCancer” and “‘Wine”, both from the UCI machine learning repository Blake and Merz [1998]. The BreastCancer dataset contains 699 instances from 2 classes in $\mathbb{R}^9$, while the Wine dataset contains 178 instances from 3 classes in $\mathbb{R}^{13}$. On each dataset, we randomly select 80% samples for training and use the rest 20% for test. On each dataset, we first reduce the dimensionality to 2 by performing MNNE, and then use the NN rule as a classifier for classification. The bandwidth h_{opt} is calculated by using (5.33), where

$d = 2$, $c = 2$ and $n = 560$ for the BreastCancer dataset, and m =3 and $n = 142$ for the Wine dataset. We train MNNE with different bandwidths $h = \{4h_{opt}, 3h_{opt}, 2h_{opt}, h_{opt}, h_{opt}/2, h_{opt}/3, h_{opt}/4\}$. The training NN error $\hat{P}_{nn}$ is calculated by

$$\hat{P}_{nn} = 1 - \hat{J}(\mathbf{W}_{opt}), \tag{5.42}$$

where $\hat{J}(\mathbf{W}_{opt})$ is the optimal objective value of (5.26). Figure 5.2 shows the training NN error and test NN error under different bandwidth h. One can see that, on both datasets, as bandwidth h varies from $4h_{opt}$ to $h_{opt}/4$, the training NN error $\hat{P}_{nn}$ generally decreases. Especially, on the Wine dataset, the training NN error becomes nearly zero when the bandwidth h is smaller than $h_{opt}/2$. This reflects the fact that small bandwidth will make the model better fit the training data. However, the performance on the test data is not necessarily improved by using a small bandwidth. The test performance on the BreastCancer dataset is quite stable around h_{opt}, while the test performance on the Wine dataset becomes slightly worse when the bandwidth h is smaller than $h_{opt}/2$, which implies overfitting may have occurred. These results confirm the validity of the proposed bandwidth selection method.



(a) BreastCancer

(b) Wine

Figure 5.2: Evaluation of MNNE with respect to bandwidth selection.

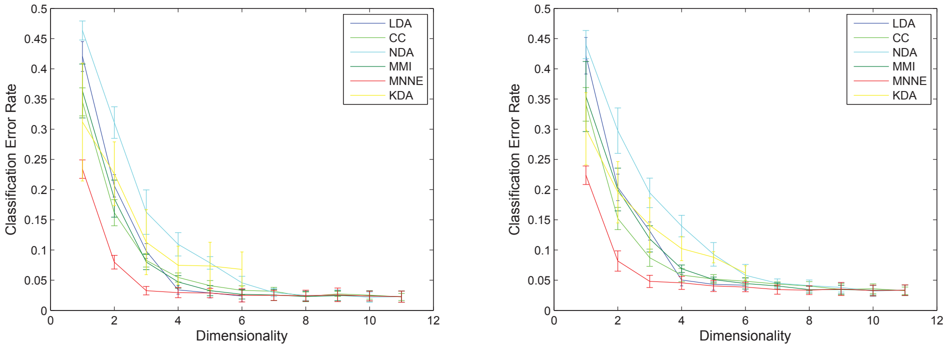
5.4.2 Experiments on UCI Machine Learning Repository

We evaluate the effectiveness of MNNE for SLDR by using data classification and visualization experiments on six datasets from the UCI machine learning repository Blake and Merz [1998]. We compare MNNE with CC Loog and Duin [2004], NDA Fukunaga and Mantock [1983], MMI Torkkola [2003], kernel discriminant analysis (KDA) Mika et al. [1999], and LDA Fisher [1936] Rao [1948]. NDA and MMI are nonparametric methods and thus distribution-free. KDA is also a distribution-free method, but instead of using nonparametric method it utilizes the kernel trick, which assumes that data follow homoscedastic Gaussian distributions in a high-dimension feature space induced by a reproducing kernel. CC is a parametric method based on heteroscedastic Gaussian assumption.The parameters of NDA and MMI are determined by 10-fold cross-validation on the training dataset, while the parameter of KDA is determined by the homoscedastic criterion in You et al. [2011].

5.4.2.1 On the ImageSeg Dataset

The “ImageSeg” dataset is commonly used for performance evaluation of SLDR methods. It contains 2,310 samples from 7 classes in $\mathbb{R}^{19}$. We randomly split the whole dataset into training and test sets at a ratio of 80% to 20%. The average performance over ten independent random splits is used for performance evaluation. Principal component analysis (PCA) Jolliffe [2002] is performed in each training round, keeping 99.9% of the total variance. Whitening preprocessing is applied before all SLDR methods. The NN rule and SVM with the Gaussian kernel are used as classifier for classification in the dimension reduced space.

Figure 5.3 shows the experimental results. A first observation is that LDA performs nearly optimal on dimensionality 6, after which all methods have almost equal performance. This implies that the dataset can be properly modeled by homoscedastic Gaussian distributions. However, when the dimensionality is lower than 6, these methods perform differently. In particular, from dimensionality 1 to 3, MNNE performs better than the other methods, because its closeness to the Bayes optimal criterion. NDA does not work well on this dataset, while CC, KDA and MMI only performs moderately.

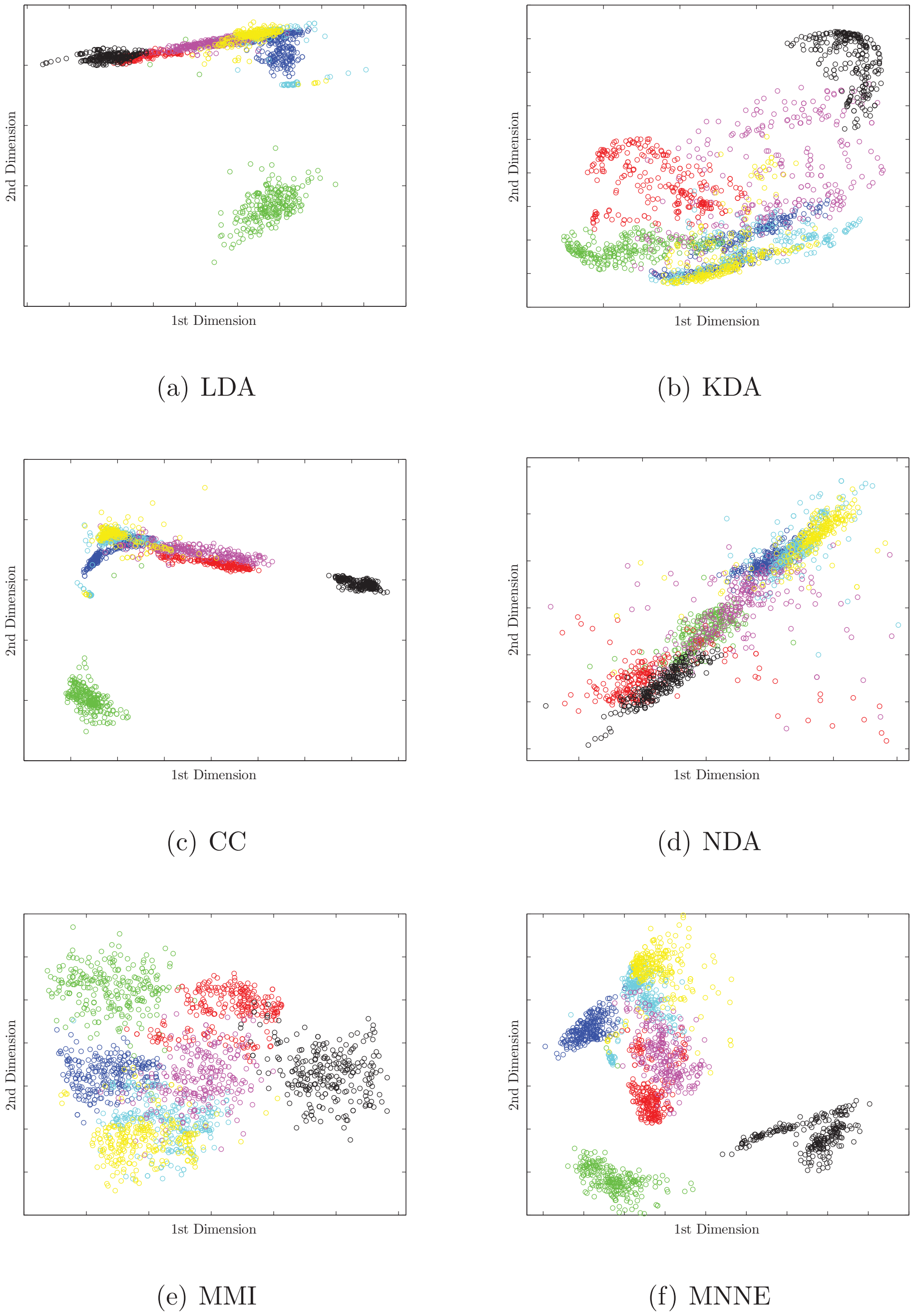


(a) By NN rule

(b) By SVM

Figure 5.3: Performance evaluation on the ImageSeg dataset. Lines denote average classification error rates and bars denote standard deviations.

Figure 5.4 shows 2-dimensional visualizations of the dataset (represented by a training dataset) obtained by using different methods. One can see that MMI and MNNE perform better than other methods in terms of class separation. Between MMI and MNNE, while both separate most classes, MNNE provides more concentrated visual results, i.e., data are more concentratively distributed with their respective classes in the 2-dimensional visualization obtained by MNNE than in that obtained by MMI. As a result, MNNE has a classification error rate of 0.08, which is lower than 0.13 of MMI.



(a) LDA

(b) KDA

(c) CC

(d) NDA

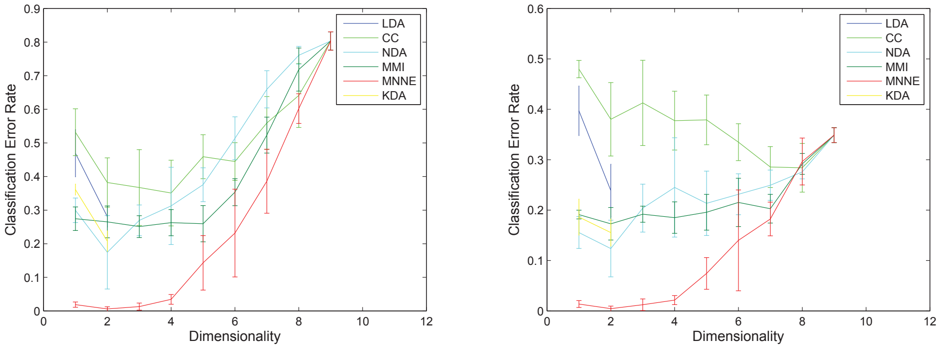
(e) MMI

(f) MNNE

Figure 5.4: 2 dimensional visualization of data from the ImageSeg dataset.

5.4.2.2 On the TicTacToe Dataset

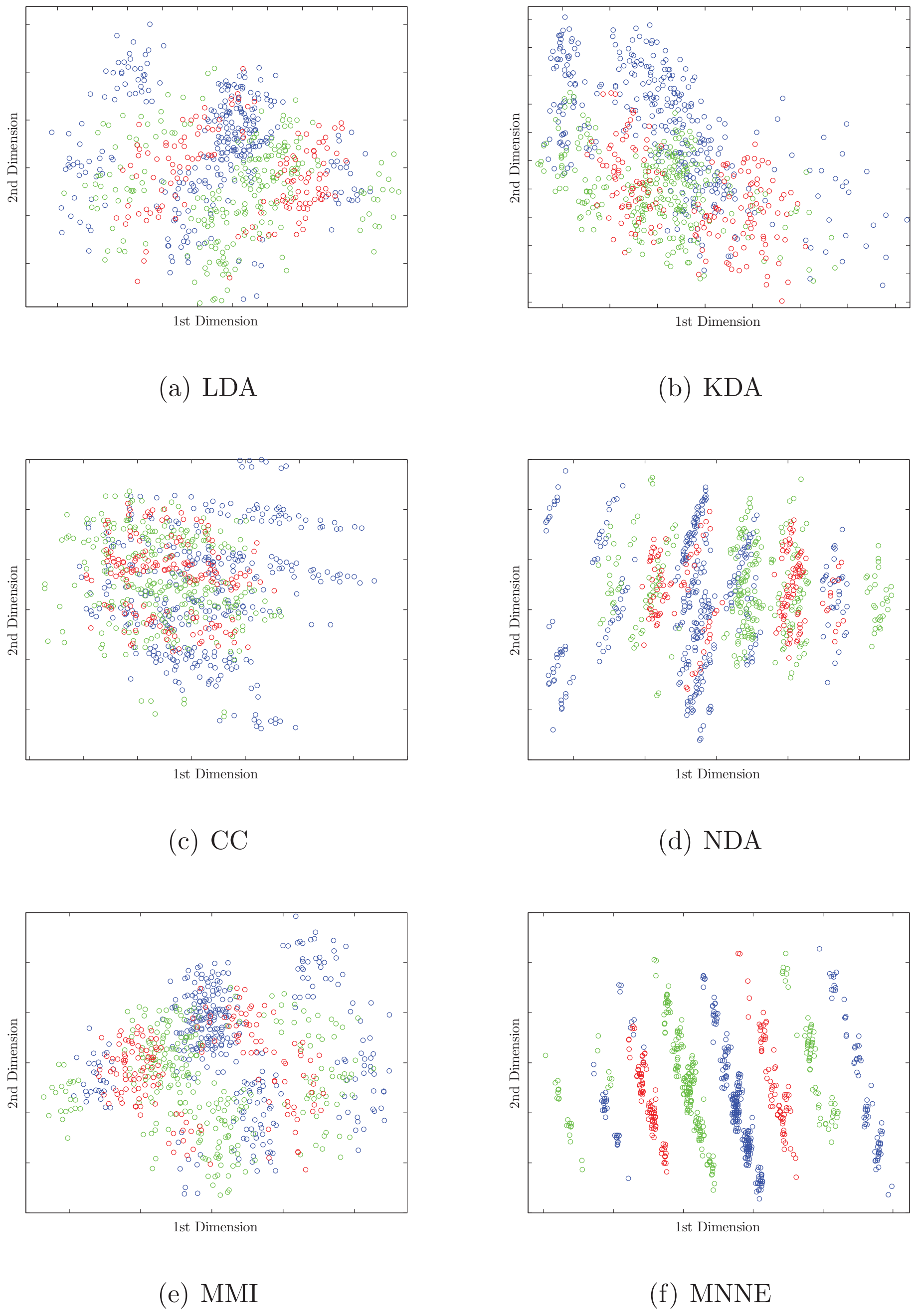
The "TicTacToe" dataset contains 958 samples from 3 classes in $\mathbb{R}^9$. Different from ImageSeg, the TicTacToe dataset is highly non-Gaussian, and all 3 classes have multi-modal. We randomly split the whole dataset into training and test sets at a ratio of 80% to 20%. The average performance over ten independent random splits is used for performance evaluation. Figure 5.5 shows the performance of different methods. Due to the multi-modal property, LDA which is based on the homoscedastic Gaussian assumption does not perform well. CC also fails due to the parametric (heteroscedastic Gaussian) assumption. NDA and MMI show improved performance over LDA and CC due to the nonparametric characteristic. However, they are inferior to MNNE in terms of lower classification error rate and higher confidence (smaller standard deviation). This can be explained via the 2 dimensional visualizations obtained by using NDA, MMI and MNNE. Figure 5.6 shows that NDA, MMI, and MNNE are able to reveal the multi-modal feature of the dataset, but the data are only concentratively distributed with their respective classes in the result obtained by MNNE. The concentrative property helps give a lower classification error with and a higher confidence.



(a) By NN rule

(b) By SVM

Figure 5.5: Performance evaluation on the TicTacToe Dataset. Lines denote average classification error rates and bars denote standard deviations.



(a) LDA

(b) KDA

(c) CC

(d) NDA

(e) MMI

(f) MNNE

Figure 5.6: 2 dimensional visualization of data from the TicTacToe dataset.

5.4.2.3 On More Datasets

Four more datasets from the UCI machine learning repository Blake and Merz [1998] are used in our experiments, including "BalanceScale", containing 625 samples from 3 classes in $\mathbb{R}^4$, "Car", containing 1728 samples from 4 classes in $\mathbb{R}^6$, "Ecoli", containing 336 samples from 8 classes in $\mathbb{R}^8$, and "Wine", containing 178 samples from 3 classes in $\mathbb{R}^{13}$. Similar to the above two experiments, we randomly split each dataset into training/test (80%/20%) datasets. The average performances over ten independent splits are shown in Table 5.1. One can see that, on the first two datasets, nonparametric methods (NDA, MMI, and MNNE) generally outperform parametric methods (LDA, CC). On the Ecoli dataset, LDA performs competitively, which implies the dataset can be well modeled by homoscedastic Gaussian distributions. On all dataset, MNNE shows promising performance.

Table 5.1: Performance evaluation on the BalanceScale, Car, Ecoli, and Wine datasets: average classification error rate (standard deviation, best dimensionality)

(a) By NN rule

	BalanceScale	Car	Ecoli	Wine
LDA	0.1246 (0.0254, 2)	0.0855 (0.0390, 3)	0.1832 (0.0320, 6)	0.0351 (0.0223, 2)
CC	0.1183 (0.0366, 1)	0.0448 (0.0108, 6)	0.1982 (0.0350, 7)	0.0162 (0.0342, 10)
NDA	0.1174 (0.0313, 3)	0.0341 (0.0155, 4)	0.1866 (0.0296, 6)	**0.0054** (0.0114, 7)
MMI	0.1138 (0.0368, 2)	0.0364 (0.0120, 5)	0.1892 (0.0304, 5)	0.0108 (0.0140, 6)
MNNE	**0.1111** (0.0297, 3)	**0.0337** (0.0101, 5)	0.1868 (0.0336, 4)	0.0088 (0.0228, 9)
KDA	0.1300 (0.0428, 2)	0.0607 (0.0420, 3)	0.2162 (0.0361, 7)	0.0889 (0.0316, 2)

(b) By SVM

	BalanceScale	Car	Ecoli	Wine
LDA	0.1040 (0.0281, 2)	0.0806 (0.0217, 3)	0.1359 (0.0202, 7)	0.0378 (0.0228, 2)
CC	0.1056 (0.0254, 1)	0.0260 (0.0059, 6)	0.1349 (0.0183, 6)	0.0297 (0.0199, 2)
NDA	0.0892 (0.0183, 3)	0.0263 (0.0117, 4)	0.1337 (0.0278, 6)	**0.0102** (0.0228, 7)
MMI	0.0889 (0.0216, 2)	0.0247 (0.0092, 4)	0.1359 (0.0202, 7)	0.0297 (0.0153, 6)
MNNE	**0.0679** (0.0113, 2)	**0.0239** (0.0092, 5)	0.1359 (0.0202, 7)	0.0216 (0.0171, 10)
KDA	0.1092 (0.0388, 2)	0.0554 (0.0277, 3)	0.1544 (0.0373, 6)	0.0507 (0.0214, 2)

5.5 Appendix

Proof of (5.12): Since $\eta_1 + \eta_2 = 1$, we have $1 - \eta_1^2 - \eta_2^2 = 2\eta_1\eta_2$. Letting $x = \eta_1$ and $1 - x = \eta_2$, (5.12) is equivalent to

$$f(x) = x \log\left(\frac{1}{x}\right) + (1-x)\log\left(\frac{1}{1-x}\right) - 4x(1-x) \geq 0, x \in [0,1]. \tag{5.43}$$

Due to the symmetry of (5.43) with respect to x and $1-x$, it is sufficient to prove that $f(x) \geq 0$ on the half interval $[0, 1/2]$. First, it is straightforward to calculate that $f(0) = 0$ and $f(1/2) = 0$. Then, we show that $f(x)$ first increases and then decreases on $[0, 1/2]$, which together with $f(0) = 0$ and $f(1/2) = 0$ guarantees $f(x)$ is nonnegative on $[0, 1/2]$. This can be done by checking

$$f'(x) = \log(1-x) - \log x - 4(1-2x) \text{ and } f''(x) = 8 - \frac{1}{x(1-x)}. \tag{5.44}$$

Note that $f''(x)$ monotonically increases from $f''(0) = -\infty$ to $f''(1/2) = 4$ on $[0, 1/2]$. Thus $f'(x)$ must first decreases and then increases on $[0, 1/2]$. In addition, $f'(0) = +\infty$ and $f'(1/2) = 0$, and thus there must exist $x^* \in (0, 1/2)$ such that $f'(x)$ is positive on $[0, x^*)$ and negative on $(x^*, 1/2)$, with $f'(x^*) = 0$. This means $f(x)$ first increases and then decreases on $[0, 1/2]$, and thus completes the proof.

Chapter 6

Conclusions

6.1 Summary of This Thesis

In this thesis, we have contributed to supervised linear dimension reduction (SLDR) from both theoretical and algorithmic aspects. To summarize, we have the following conclusions:

1. When the dimensionality D and the training sample number N are both large (actually 100 is sufficient according to empirical study), the generalization performance of LDA for a fixed problem is only affected by the dimensionality to training sample number ratio $\gamma = D/N$. Specifically, if the population discrimination power of the problem at hand is sufficient (larger than 10), then $\gamma = 0.2$ is enough for LDA to preserve about 70% of the discrimination power.

2. When the training sample number N is deficient compared to the dimensionality D, i.e., the ratio $\gamma = D/N$ is large, the generalization performance of LDA can be improved by block-diagonal regularization. In particular, the sample error of the block-diagonally regularized LDA decreases as the number of variable groups increases. Empirically, the block-diagonally regularized LDA performs competitively compared with other types of regularized LDA, e.g., with the Tikhonov regularization and the banded regularization.

3. A major problem of existing parametric SLDR methods, including LDA and its extensions, is that the dimensionality of the learned subspace is low

close classes cannot be well separated. This problem can be overcome by the max-min distance analysis (MMDA), which optimizes the projection matrix by maximizing the minimum pairwise distance among all class pairs in the dimension reduced space and thus duly considers the separation of all classes.

4. In nonparametric SLDR, it usually needs to define a proxy criterion to approximate the Bayes optimal criterion for projection matrix optimization. In the literature, the state-of-the-art proxy criterion for nonparametric SLDR is maximizing mutual information (MMI). However, minimizing the asymptotic nearest neighbor classification error (MNNE) is better than MMI in terms of the closeness to the Bayes optimal criterion.

6.2 Future Works

6.2.1 SLDR for Structured Data

In this thesis, we have studied SLDR from both theoretical and algorithmic perspectives. These studies are considerably general, which do not make assumptions or utilize prior knowledge on data structure. In one direction of future work, we will consider SLDR for structured data. Spatial and/or time structured data are common in practical applications, e.g., face images, traffic trajectories and time series. How to explore the structural information of data to improve the accuracy and the robustness of SLDR is a fundamental and valuable problem. There have been relevant studies in the literature, e.g., on functional data James and Hastie [2001] and stationary times series Shumway and Unger [1974]. However, a comprehensive study, especially from the theoretical viewpoint, has not been conducted. A number of important issues to be addressed are: 1) how to estimate data structure robustly with finite training samples, 2) how to integrate structural information into SLDR, and 3) how to justify the performance of a structural SLDR method theoretically.

6.2.2 SLDR for Compressed Data

Compressed sensing (CS) Candés et al. [2006] Donoho [2006] has been an emerging research direction in the signal processing field, which departs from the conventional transformation-based signal processing theories and techniques. It proves that under the (approximate) sparsity assumption, which can be satisfied by most natural signals, the information of a signal can be recovered from a relatively small number of random measurements. The practical application of CS, though still under development, will change not only the fields of signal processing and communications but also pattern recognition and other related areas. In one of our future works, we will try to establish new theories of SLDR for CS based data analysis.

Bibliography

Parameterisation of a Stochastic Model for Human Face Identification, Sarasota FL, December 1994. 61

Linear Discriminant Analysis for Speechreading, 1998. 71

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, New York, NY, second edition, 1984. 12, 13, 16

R. Ash. *Information Theory.* Wiley, New York, 1965. 98

Z.D. Bai and J.W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345, 1998. 13

P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 52, 53, 71

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003. 3, 60

W. Bian and D. Tao. Harmonic mean for subspace selection. In *19th International Conference on Pattern Recognition*, pages 1–4, 2008. 72, 75, 86, 93

Peter J. Bickel and Elizaveta Levina. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004. 36

P.J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008a. 53

P.J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008b. 53, 64

P. Billingsley. *Convergence of Probability Measures*, volume 175 of *Wiley Series in Probability and Statistics: Probability and Statistics*. John Wiley & Sons Inc., 1999. 43

C.M. Bishop. *Pattern recognition and machine learning*. springer New York, 2006. 1

C.M. Bishop, M. Svensén, and C.K.I. Williams. Gtm: The generative topographic mapping. *Neural computation*, 10(1):215–234, 1998. 3

C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. 31, 105, 107, 112

D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. 2

E.J. Candés, J.K. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006. 117

L. Cayton. Algorithms for manifold learning. Technical report, Department of Computer Science, University of California at San Diego, 2005. 3

L.F. Chen, H.Y.M. Liao, M.T. Ko, J.C. Lin, and G.J. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern recognition*, 33(10):1713–1726, 2000. 54

T.M. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. 94, 95

T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991. 96

S. Dasgupta. Learning mixtures of gaussians. In *IEEE Annual Symposium on Foundations of Computer Science*, pages 634–644. IEEE Computer Society, 1999. 19

A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation of sparse PCA using semidefinite programming. *SIAM Review*, 49(3), 2007. 81

J. Dattorro. *Convex Optimization and Euclidean Distance Geometry.* Meboo Publishing, 2008. 84, 85

F. De la Torre and T. Kanade. Multimodal oriented discriminant analysis. In *Proceedings of the 22nd international conference on Machine learning*, pages 177–184, 2005. 94

H.P. Decell and S.M. Mayekar. Feature combinations and the divergence criterion. *Computers and Math. with Applications*, 3(4):71–76, 1977. 5, 94

A.P. Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972. 53

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer Verlag, 1996. 1

D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 117

D.L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *PNAS*, 100(10):5591–5596, May 2003. 3

R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification.* John Wiley & Sons, 2001. 1

A. Edelman. *Eigenvalues and condition numbers of random matrices.* PhD thesis, Massachusetts Institute of Technology, 1989. 40

A. Edelman and N.R. Rao. Random matrix theory. *Acta Numerica*, 14(233-297): 139, 2005. 13, 22, 42, 46

A. Edelman, T.A. Arias, and S.T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl*, 20:303–353, 1998. 104

N. EL Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, 36:2757–2790, 2008. 13

R. Fano. *Transmission of Information: A Statistical theory of Communications*. Wiley, New York, 1961. 96

M. Fazel, H. Hindi, and S. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *Proceedings of the American Control Conference*, pages 2156–2162, 2003. 84, 85

R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen.*, 7:179–188, 1936. 2, 4, 12, 79, 93, 107

I.K. Fodor. A survey of dimension reduction techniques. Technical Report UCRL-ID-148494, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, 2002. 3

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432, 2008. 53

K. Fujisawa, Y. Futakata, M. Kojima, S. Matsuyama, S. Nakamura, K. Nakata, and M. Yamashita. Sdpa-m (semidefinite programming algorithm in matlab) user's manual — version 6.2.0. Technical Report B-359, Dept. Math. & Comp. Sciences, Tokyo Institute of Technology, 2000. 84

K. Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition*. Academic Press, September 1990. 15

K. Fukunaga and J. Mantock. Nonparametric discriminant analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 5:671–678, 1983. 2, 5, 94, 107

R.L. Gorsuch. *Factor analysis*. Hillsdale, NJ: Erlbaum, 1983. 2

M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control (a tribute to M. Vidyasagar), V. Blondel, S. Boyd, and H. Kimura, editors*, pages 95–110. Springer. 84

M. Grant and S. Boyd. Cvx: Matlab software for disciplined convex programming (web page and software). 2010. URL `http://cvxr.com/cvx`. 84

Y. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, January 2007. 52, 64

O.C. Hamsici and A.M. Martinez. Bayes optimality in linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4): 647–657, 2008. 72, 75, 78, 86

X. He, S. Yan, Y. Hu, and P. Niyogi. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005. 62

X. He, D. Cai, and J. Han. Learning a maximummargin subspace for image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, (2):189–201, 2008. 71

M.E. Hellman and J. Raviv. Probability of error, equivocation and the chernoff bound. *IEEE Transactions on Information Theory*, 16:368–372, 1970. 96

T. Hofmann. Probabilistic Latent Semantic Analysis. In *Proceedings of Uncertainty in Artificial Intelligence*, 1999. 2

G.M. James and T.J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B*, 63(3): 533–550, 2001. 116

I.T. Jolliffe. *Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed.* Springer, NY, 2002. 2, 79, 107

T. Kim and J. Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):318–327, 2005. 71

C. Lee and D.A. Landgrebe. Decision boundary feature extraction for nonparametric classification. *IEEE Transactions on Systems, Man and Cybernetics*, 23 (2):433–444, 1993. 6

M. Loog and R.P.W. Duin. Linear dimensionality reduction via a heteroscedastic extension of lda: The chernoff criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):732–739, 2004. 5, 6, 94, 107

M. Loog, R.P.W. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(7):762–766, 2001. 72, 76, 93

R. Lotlikar and R. Kothari. Fractional-step dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):623–627, 2000. 71

Juwei Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern Recognition Letter*, 26:181–191, January 2005. 52

J. Lv. *High Dimensional Variable Selection and Covariance Matrix Estimation*. PhD thesis, Princeton University, Department of Mathematics, 2007. 53

V.A Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1:457, 1967. 13, 39

S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.R. Mullers. Fisher discriminant analysis with kernels. In *Proceedings of IEEE Neural Networks for Signal Processing Workshop*, 1999. 2, 3, 107

Z. Nenadic. Information discriminant analysis: Feature extraction with an information-theoretic objective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1394–1407, 2007. 94

M.L. Overton and R.S. Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM J. Matrix Anal. Appl.*, 13(1):41–45, 1992. 80

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. 102

R. Pless and R. Souvenir. A survey of manifold learning for images. *IPSJ Transactions on Computer Vision and Applications*, 1:83–94, March 2009. 3

C.R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society series B: Methodological*, 10:159–203, 1948. 2, 4, 12, 75, 79, 93, 107

S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000. 3

G. Saon and M. Padmanabhan. Minimum bayes error feature selection for continuous speech recognition. In *Advances in Neural Information Processing Systems 13*, pages 800–806. MIT Press, 2001. 6, 94, 99

L.K. Saul, K.Q. Weinberger, J.H. Ham, F. Sha, and D.D. Lee. Spectral methods for dimensionality reduction. *Semi-Supervised Learning*. 3

M.J. Schervish. Linear discrimination for three known normal populations. *Journal of Statistical Planning and Inferencethe*, 10:167–175, 1984. 72, 78

B. Schölkopf, A. Smola, and K.R. Müller. Kernel principal component analysis. In *Artificial Neural Networks'97*, pages 583–588. Springer, 1997. 3

R.H. Shumway and A.N. Unger. Linear discriminant functions for stationary time series. *Journal of the American Statistical Association*, pages 948–956, 1974. 116

B.W. Silverman. *Density Estimation for Statistics and Data Analysis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1 edition, April 1986. 102

G.W. Stewart and J. Sun. *Matrix perturbation theory*, volume 175. Academic press New York, 1990. 22

D. Tao, X. Li, X. Wu, and S.J. Maybank. Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):260–274, 2009. 5, 71, 72, 75, 86, 93, 94, 104

T. Tao. *Topics in Random Matrix Theory*. American Mathematical Society, 2012. 41, 50

J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000. 3

K. Torkkola. Feature extraction by non parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003. 6, 94, 96, 104, 107

A.M. Tulino and S. Verdú. *Random matrix theory and wireless communications*, volume 1. Now Publishers Inc, 2004. 13, 42

V. Vapnik. *Statistical learning theory*. Wiley, New York, 1998. 1

A.S. Wagaman and E. Levina. Discovering sparse covariance structures with the isomap. *Journal of Computational and Graphical Statistics*, 18(3):551–572, 2009. 64

X. Wang and X. Tang. Dual-space linear discriminant analysis for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–569, 2004. 54

E.P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *The Annals of Mathematics*, 62(3):548–564, 1955. 13

E.P. Wigner. On the distribution of the roots of certain symmetric matrices. *The Annals of Mathematics*, 67(2):325–327, 1958. 13

J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6: 483–502, 2005. 54

J. Ye and Q. Li. A two-stage linear discriminant analysis via qr-decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):929–941, 2005. 54

J. Ye, R. Janardan, C.H. Park, and H. Park. An optimization criterion for generalized discriminant analysis on undersampled problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):982–994, 2004. 54

J. Ye, T. Xiong, Q. Li, R. Janardan, J. Bi, V. Cherkassky, and C. Kambhamettu. Efficient model selection for regularized linear discriminant analysis. In *International Conference on Information and Knowledge Management*. ACM, 2006. 52, 64

Y.Q. Yin, Z.D. Bai, and P.R. Krishnaiah. On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probability Theory Related Fields*, 78:509–521, 1988. 13

D. You, O.C. Hamsici, and A.M. Martinez. Kernel optimization in discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 (3):631–638, 2011. 107

Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal of Scientific Computing*, 26(1): 313–338, 2004. 3