High-Dimensional Data Analysis

高维数据分析

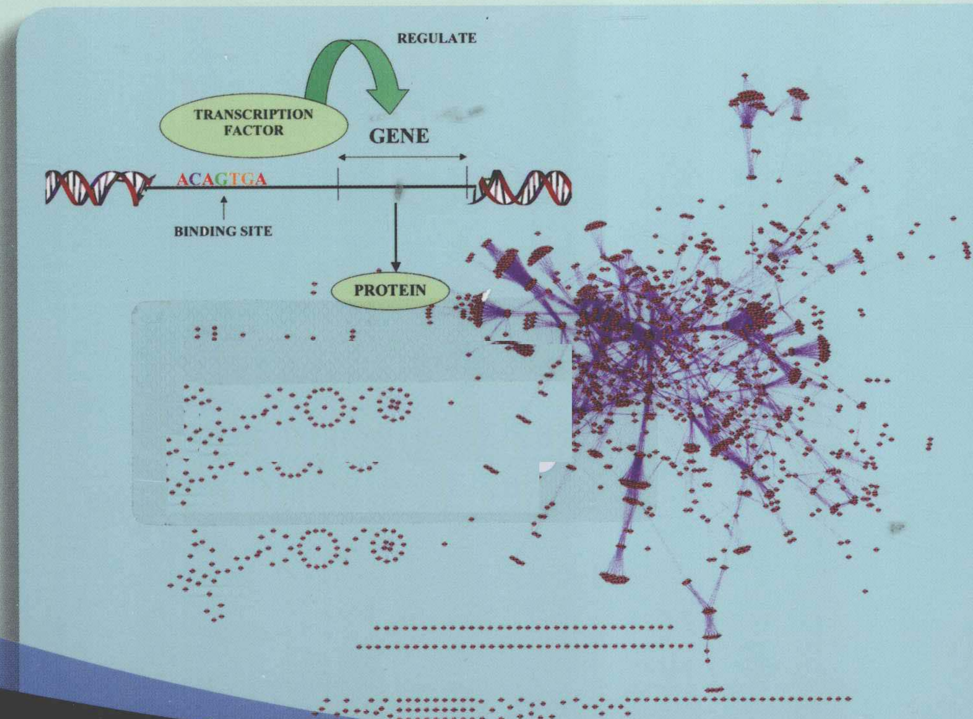
Editors

T. Tony Cai

Xiaotong Shen

Volume 2

Frontiers of Statistics



高等教育出版社
HIGHER EDUCATION PRESS

High-Dimensional Data Analysis

高维数据分析

Gaowei Shuju Fenxi

Editors

T. Tony Cai
University of Pennsylvania, USA

Xiaotong Shen
University of M

Volume 2

Frontiers of Statistics

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

World Scientific

NEW JERSEY · LONDON · SINGAPORE · BEIJING · SHANGHAI · HONG KONG · TAIPEI · CHENNAI

T. Tony Cai
Department of Statistics,
University of Pennsylvania,
Philadelphia, PA 19104, USA.

Xiaotong Shen
School of Statistics,
University of Minnesota,
Minneapolis, MN 55455, USA.

图书在版编目（CIP）数据

高维数据分析 = High-Dimensional Data Analysis: 英文 / 蔡天文, 沈晓彤编. —北京: 高等教育出版社, 2010.10
(统计前沿 / 范剑青, 马志明主编)
ISBN 978-7-04-029851-2

Ⅰ. ①高… Ⅱ. ①蔡… ②沈… Ⅲ. ①统计数据-统计分析 (数学) - 文集 - 英文 Ⅳ. ①O212.1-53

中国版本图书馆 CIP 数据核字 (2010) 第 173103 号

策划编辑	王丽萍	**责任编辑**	王丽萍	**封面设计**	张　楠
责任校对	刘　莉	**责任印制**	陈伟光		

出版发行	高等教育出版社	**购书热线**	010-58581118
社　　址	北京市西城区德外大街 4 号	**免费咨询**	800-810-0598
邮政编码	100120	**网　　址**	http://www.hep.edu.cn http://www.hep.com.cn
经　　销	蓝色畅想图书发行有限公司	**网上订购**	http://www.landraco.com http://www.landraco.com.cn
印　　刷	涿州市星河印刷有限公司	**畅想教育**	http://www.widedu.com
开　　本	787 × 1092　1/16		
印　　张	20	**版　　次**	2010 年 10 月第 1 版
字　　数	300 000	**印　　次**	2010 年 10 月第 1 次印刷
插　　页	1	**定　　价**	68.00 元

物料号 29851-00

本书海外版由 World Scientific 负责在中国大陆地区以外区域销售, ISBN 为 978-981-4824-85-4

Preface

Over the last few years, significant developments have been taking place in high-dimensional data analysis, which are driven primarily by a wide range of applications in many fields, such as genomics and signal processing. In particular, substantial advances have been made in the areas of feature selection, covariance estimation, classification and regression. This book intends to examine important issues arising from high-dimensional data analysis to explore key ideas for statistical inference and prediction. The book is structured around topics on multiple hypothesis testing, feature selection, regression, classification, dimension reduction, as well as applications in survival analysis and in biomedical research.

Fundamental statistical issues underlying data have changed, when moving from low-dimensional to high-dimensional analyses. For instance, certain structures such as sparsity need to be utilized in feature selection when the number of candidate features greatly exceeds that of the sample size. As a result of high-dimensionality, traditional statistical methods designed for low dimensional problems become inadequate or break down. To meet these challenges in high-dimensional analysis, statisticians have been developing new methods and introducing new concepts, where many issues emerge with regard to how to identify or utilize certain structures for dimension reduction in inference and prediction.

There exists a vast body of literature on high-dimensional analysis, especially for prediction, classification and regression. We do not intend to give an overview of each subject but would like to mention here only a few topics of interest—feature selection, basis/grouping pursuit, multiple hypothesis testing, effective dimension reduction and projection pursuit, sparsity, high-dimensional regression and classification.

Many classification problems are often high-dimensional. In Chapter 1, Fan, Fan and Wu review contemporary classification methods, including linear discriminant analysis, naive Bayes, and loss-based methods, as well as the impact of dimensionality on classification, with a special attention towards regularization and feature selection. In Chapter 2, Liu and Wu move further to the topic of large margin classification, where they examine various state-of-art methods for various of support vector machines, and their connection with probability estimation.

In Chapter 3, Cai and Sun consider large-scale multiple testing. They begin by reviewing methods for controlling family-wise error rates and false discovery rates (FDR), as well as other pertinent issues in multiple testing. Their main focus is on optimal multiple testing procedures minimizing the false nondiscovery rate while controls the FDR. Both independent and dependent cases are considered.

The topic of high-dimensional feature (variable) selection has been a focus in recent research. In Chapter 4, Yuan reviews several popular variable selection methods, and contrast classical methods such as stepwise selection with modern methods such as regularization. In Chapter 5, Zhu, Pan and Shen examine Bayesian model selection for networks, particularly gene networks where the number of genes in a network may greatly exceed the sample size.

In Chapter 6, Li describes a number of interesting applications in genomics studies involving networks and graphical models, where the dimension under consideration is ultra-high. Various regression techniques have been reviewed, where special structures of genomic data are considered.

Survival data analysis is an important subject in biostatistics. Analysis high-dimensional survival data requires power tools. In Chapter 7, Li and Ren focus on joint modeling for censored and longitudinal data. Various models are reviewed, subject to different types of censoring. In Chapter 8, Nan reviews the recent development of feature selection in penalized regression in survival analysis, which is a marriage between high-dimensional feature selection and survival analysis. Several methods are examined, particularly for high-dimensional covariates such as gene expressions, whereas various penalties such as grouped, hierarchical penalties are discussed.

For high-dimensional data analysis, dimension reduction is essential. In Chapter 9, Yin gives a comprehensive review on sufficient dimension reduction in regression. In Chapter 10, Chen and Yang discuss combining strategies.

Finally, we sincerely hope that this book can simulate further interest from statisticians, computer scientists and engineers, and promote further collaborations among them to attack important problems in high-dimensional data analysis.

T Tony Cai, Philadelphia
Xiaotong Shen, Minneapolis
May 18, 2010

Contents

Part III Model Building with Variable Selection

Part IV High-Dimensional Statistics in Genomics

Part V Analysis of Survival and Longitudinal Data

Part VI Sufficient Dimension Reduction in Regression

Part I

High-Dimensional Classification

Chapter 1

High-Dimensional Classification[*]

Jianqing Fan[†] Yingying Fan[‡] and Yichao Wu[§]

Abstract

In this chapter, we give a comprehensive overview on high-dimensional classification, which is prominently featured in many contemporary statistical problems. Emphasis is given on the impact of dimensionality on implementation and statistical performance and on the feature selection to enhance statistical performance as well as scientific understanding between collected variables and the outcome. Penalized methods and independence learning are introduced for feature selection in ultrahigh dimensional feature space. Popular methods such as the Fisher linear discriminant, Bayes classifiers, independence rules, distance-based classifiers and loss-based classification rules are introduced and their merits are critically examined. Extensions to multi-class problems are also given.

Keywords: Bayes classifier, classification error rates, distanced-based classifier, feature selection, impact of dimensionality, independence learning, independence rule, loss-based classifier, penalized methods, variable screening.

1 Introduction

Classification is a supervised learning technique. It arises frequently from bioinformatics such as disease classifications using high throughput data like micorarrays or SNPs and machine learning such as document classification and image recognition. It tries to learn a function from training data consisting of pairs of input features and categorical output. This function will be used to predict a class label of any valid input feature. Well known classification methods include (multiple) logistic regression, Fisher discriminant analysis, k-th-nearest-neighbor classifier, support vector machines, and many others. When the dimensionality of the input

[*] The authors are partly supported by NSF grants DMS-0714554, DMS-0704337, DMS-0906784, and DMS-0905561 and NIH grants R01-GM072611 and R01-CA149569.

[†] Department of ORFE, Princeton University, Princeton, NJ 08544, USA, E-mail: jqfan@princeton.edu

[‡] Information and Operations Management Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA, E-mail: fanyingy@marshall.usc.edu

[§] Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA, E-mail: wu@stat.ncsu.edu

feature space is large, things become complicated. In this chapter we will try to investigate how the dimensionality impacts classification performance. Then we propose new methods to alleviate the impact of high dimensionality and reduce dimensionality.

We present some background on classification in Section 2. Section 3 is devoted to study the impact of high dimensionality on classification. We discuss distance-based classification rules in Section 4 and feature selection by independence rule in Section 5. Another family of classification algorithms based on different loss functions is presented in Section 6. Section 7 extends the iterative sure independent screening scheme to these loss-based classification algorithms. We conclude with Section 8 which summarizes some loss-based multicategory classification methods.

2 Elements of classifications

Suppose we have some input space $\mathcal{X}$ and some output space $\mathcal{Y}$. Assume that there are independent training data $(\mathbf{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$ coming from some unknown distribution P, where Y_i is the i-th observation of the response variable and $\mathbf{X}_i$ is its associated feature or covariate vector. In classification problems, the response variable Y_i is qualitative and the set $\mathcal{Y}$ has only finite values. For example, in the cancer classification using gene expression data, each feature vector $\mathbf{X}_i$ represents the gene expression level of a patient, and the response Y_i indicates whether this patient has cancer or not. Note that the response categories can be coded by using indicator variables. Without loss of generality, we assume that there are K categories and $\mathcal{Y} = \{1, 2, \ldots, K\}$. Given a new observation $\mathbf{X}$, classification aims at finding a classification function $g : \mathcal{X} \to \mathcal{Y}$, which can predict the unknown class label Y of this new observation using available training data as accurately as possible.

To access the accuracy of classification, a loss function is needed. A commonly used loss function for classification is the *zero-one loss*:

$$L(y, g(\mathbf{x})) = \begin{cases} 0, & g(\mathbf{x}) = y, \\ 1, & g(\mathbf{x}) \neq y. \end{cases} \tag{2.1}$$

This loss function assigns a single unit to all misclassifications. Thus the risk of a classification function g, which is the expected classification error for an new observation $\mathbf{X}$, takes the following form:

$$\begin{aligned} \overline{W}(g) = E[L(Y, g(\mathbf{X}))] &= E\left[\sum_{k=1}^{K} L(k, g(\mathbf{X}))P(Y = k|\mathbf{X})\right] \\ &= 1 - P(Y = g(\mathbf{x})|\mathbf{X} = \mathbf{x}), \end{aligned} \tag{2.2}$$

where Y is the class label of $\mathbf{X}$. Therefore, the optimal classifier in terms of minimizing the misclassification rate is

$$g^*(\mathbf{x}) = \arg\max_{k \in \mathcal{Y}} P(Y = k|\mathbf{X} = \mathbf{x}) \tag{2.3}$$

This classifier is known as the *Bayes classifier* in the literature. Intuitively, Bayes classifier assigns a new observation to the most possible class by using the posterior probability of the response. By definition, Bayes classifier achieves the minimum misclassification rate over all measurable functions:

$$\overline{W}(g^*) = \min_g \overline{W}(g). \tag{2.4}$$

This misclassification rate $\overline{W}(g^*)$ is called the Bayes risk. The Bayes risk is the minimum misclassification rate when distribution is known and is usually set as the benchmark when solving classification problems.

Let $f_k(\mathbf{x})$ be the conditional density of an observation $\mathbf{X}$ being in class k, and π_k be the prior probability of being in class k with $\sum_{i=1}^K \pi_i = 1$. Then by Bayes theorem it can be derived that the posterior probability of an observation $\mathbf{X}$ being in class k is

$$P(Y = k|\mathbf{X} = \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{i=1}^K f_i(\mathbf{x})\pi_i}. \tag{2.5}$$

Using the above notation, it is easy to see that the Bayes classifier becomes

$$g^*(\mathbf{x}) = \arg\max_{k\in\mathcal{Y}} f_k(\mathbf{x})\pi_k. \tag{2.6}$$

For the following of this chapter, if not specified we shall consider the classification between two classes, that is, $K = 2$. The extension of various classification methods to the case where $K > 2$ will be discussed in the last section.

The *Fisher linear discriminant analysis* approaches the classification problem by assuming that both class densities are multivariate Gaussian $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectively, where $\boldsymbol{\mu}_k$, $k = 1, 2$ are the class mean vectors, and $\boldsymbol{\Sigma}$ is the common positive definite covariance matrix. If an observation $\mathbf{X}$ belongs to class k, then its density is

$$f_k(\mathbf{x}) = (2\pi)^{-p/2}(\det(\boldsymbol{\Sigma}))^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right\}, \tag{2.7}$$

where p is the dimension of the feature vectors $\mathbf{X}_i$. Under this assumption, the Bayes classifier assigns $\mathbf{X}$ to class 1 if

$$\pi_1 f_1(\mathbf{X}) \geqslant \pi_2 f_2(\mathbf{X}), \tag{2.8}$$

which is equivalent to

$$\log\frac{\pi_1}{\pi_2} + (\mathbf{X} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geqslant 0, \tag{2.9}$$

where $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$. In view of (2.6), it is easy to see that the classification rule defined in (2.8) is the same as the Bayes classifier. The function

$$\delta_F(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{2.10}$$

is called the *Fisher discriminant function*. It assigns $\mathbf{X}$ to class 1 if $\delta_F(\mathbf{X}) \geqslant \log \frac{\pi_2}{\pi_1}$; otherwise to class 2. It can be seen that the Fisher discriminant function is linear in $\mathbf{x}$. In general, a classifier is said to be linear if its discriminant function is a linear function of the feature vector. Knowing the discriminant function δ_F, the classification function of Fisher discriminant analysis can be written as $g_F(\mathbf{x}) = 2 - I(\delta_F(\mathbf{x}) \geqslant \log \frac{\pi_2}{\pi_1})$ with $I(\cdot)$ the indicator function. Thus the classification function is determined by the discriminant function. In the following, when we talk about a classification rule, it could be the classification function g or the corresponding discriminant function δ.

Denote by $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ the parameters of the two Gaussian distributions $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Write $\overline{W}(\delta, \boldsymbol{\theta})$ as the misclassification rate of a classifier with discriminant function δ. Then the discriminant function δ_B of the Bayes classifier minimizes $\overline{W}(\delta, \boldsymbol{\theta})$. Let $\Phi(t)$ be the distribution function of a univariate standard normal distribution. If $\pi_1 = \pi_2 = \frac{1}{2}$, it can easily be calculated that the misclassification rate for Fisher discriminant function is

$$\overline{W}(\delta_F, \boldsymbol{\theta}) = \Phi\left(-\frac{d^2(\boldsymbol{\theta})}{2}\right), \tag{2.11}$$

where $d(\boldsymbol{\theta}) = \{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}^{1/2}$ and is named as the *Mahalanobis distance* in the literature. It measures the distance between two classes and was introduced by Mahalanobis (1930). Since under the normality assumption the Fisher discriminant analysis is the Bayes classifier, the misclassification rate given in (2.11) is in fact the Bayes risk. It is easy to see from (2.11) that the Bayes risk is a decreasing function of the distance between two classes, which is consistent with our common sense.

Let Γ be some parameter space. With a slight abuse of the notation, we define the maximum misclassification rate of a discriminant function δ over Γ as

$$\overline{W}_\Gamma(\delta) = \sup_{\boldsymbol{\theta} \in \Gamma} \overline{W}(\delta, \boldsymbol{\theta}). \tag{2.12}$$

It measures the worst classification result of a classifier δ over the parameter space Γ. In some cases, we are also interested in the *minimax regret* of a classifier, which is the difference between the maximum misclassification rate and the minimax misclassification rate, that is,

$$R_\Gamma(\delta) = \overline{W}_\Gamma(\delta) - \sup_{\boldsymbol{\theta} \in \Gamma} \min_{\delta} \overline{W}(\delta, \boldsymbol{\theta}). \tag{2.13}$$

Since the Bayes classification rule δ_B minimizes the misclassification rate $\overline{W}(\delta, \boldsymbol{\theta})$, the minimax regret of δ can be rewritten as

$$R_\Gamma(\delta) = \overline{W}_\Gamma(\delta) - \sup_{\boldsymbol{\theta} \in \Gamma} \overline{W}(\delta_B, \boldsymbol{\theta}). \tag{2.14}$$

From (2.11) it is easy to see that for classification between two Gaussian distributions with common covariance matrix, the minimax regret of δ is

$$R_\Gamma(\delta) = \overline{W}_\Gamma(\delta) - \sup_{\boldsymbol{\theta} \in \Gamma} \Phi\left(-\frac{1}{2} d(\boldsymbol{\theta})\right). \tag{2.15}$$

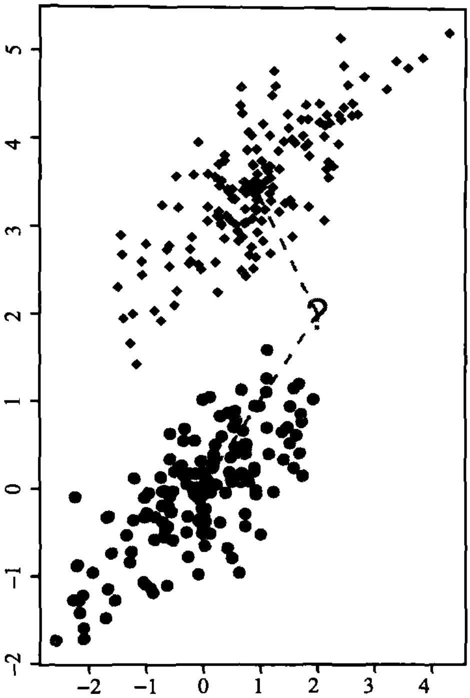


Figure 2.1 Illustration of distance-based classification. The centroid of each subsample in the training data is first computed by taking the sample mean or median. Then, for a future observation, indicated by query, it is classified according to its distances to the centroids.

The Fisher discriminant rule can be regarded as a specific method of distance-based classifiers, which have attracted much attention of researchers. Popularly used distance-based classifiers include support vector machine, naive Bayes classifier, and k-th-nearest-neighbor classifier. The distance-based classifier assigns a new observation $\mathbf{X}$ to class k if it is on average closer to the data in class k than to the data in any other classes. The "distance" and "average" are interpreted differently in different methods. Two widely used measures for distance are the Euclidean distance and the Mahalanobis distance. Assume that the center of class i distribution is $\boldsymbol{\mu}_i$ and the common convariance matrix is $\boldsymbol{\Sigma}$. Here "center" could be the mean or the median of a distribution. We use $\text{dist}(\mathbf{x}, \boldsymbol{\mu}_i)$ to denote the distance of a feature vector $\mathbf{x}$ to the centriod of class i. Then if the Euclidean distance is used,

$$\text{dist}_E(\mathbf{x}, \boldsymbol{\mu}_i) = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_i)^T(\mathbf{x} - \boldsymbol{\mu}_i)}, \tag{2.16}$$

and the Mahalanobis distance between a feature vector $\mathbf{x}$ and class i is

$$\text{dist}_M(\mathbf{x}, \boldsymbol{\mu}_i) = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)}. \tag{2.17}$$

Thus the distance-based classifier places a new observation $\mathbf{X}$ to class k if

$$\arg\min_{i\in\mathcal{Y}} \operatorname{dist}(\mathbf{X}, \boldsymbol{\mu}_i) = k. \tag{2.18}$$

Figure 2.1 illustrates the idea of distanced classifier classification.

When $\pi_1 = \pi_2 = 1/2$, the above defined Fisher discriminant analysis has the interpretation of distance-based classifier. To understand this, note that (2.9) is equivalent to

$$(\mathbf{X} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_1) \leqslant (\mathbf{X} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_2). \tag{2.19}$$

Thus δ_F assigns $\mathbf{X}$ to class 1 if its Mahalanobis distance to the center of class 1 is smaller than its Mahalanobis distance to the center of class 2. We will introduce in more details about distance-based classifiers in Section 4.

3 Impact of dimensionality on classification

A common feature of many contemporary classification problems is that the dimensionality p of the feature vector is much larger than the available training sample size n. Moreover, in most cases, only a fraction of these p features are important in classification. While the classical methods introduced in Section 2 are extremely useful, they no longer perform well or even break down in high dimensional setting. See Donoho (2000) and Fan and Li (2006) for challenges in high dimensional statistical inference. The impact of dimensionality is well understood for regression problems, but not as well understood for classification problems. In this section, we discuss the impact of high dimensionality on classification when the dimension p diverges with the sample size n. For illustration, we will consider discrimination between two Gaussian classes, and use the Fisher discriminant analysis and independence classification rule as examples. We assume in this section that $\pi_1 = \pi_2 = \frac{1}{2}$ and n_1 and n_2 are comparable.

3.1 Fisher discriminant analysis in high dimensions

Bickel and Levina (2004) theoretically studied the asymptotical performance of the sample version of Fisher discriminant analysis defined in (2.10), when both the dimensionality p and sample size n goes to infinity with p much larger than n. The parameter space considered in their paper is

$$\Gamma_1 = \{\boldsymbol{\theta} : d^2(\boldsymbol{\theta}) \geqslant c^2, c_1 \leqslant \lambda_{\min}(\boldsymbol{\Sigma}) \leqslant \lambda_{\max}(\boldsymbol{\Sigma}) \leqslant c_2, \boldsymbol{\mu}_k \in B, k = 1, 2\}, \tag{3.1}$$

where c, c_1 and c_2 are positive constants, $\lambda_{\min}(\boldsymbol{\Sigma})$ and $\lambda_{\max}(\boldsymbol{\Sigma})$ are the minimum and maximum eigenvalues of $\boldsymbol{\Sigma}$, respectively, and $B = B_{\mathbf{a},d} = \{\mathbf{u} : \sum_{j=1}^{\infty} a_j u_j^2 < d^2\}$ with d some constant, and $a_j \to \infty$ as $j \to \infty$. Here, the mean vectors $\boldsymbol{\mu}_k$, $k = 1, 2$ are viewed as points in l_2 by adding zeros at the end. The condition on eigenvalues ensures that $\frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})} \leqslant \frac{c_2}{c_1} < \infty$, and thus both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ are not ill-conditioned. The condition $d^2(\boldsymbol{\theta}) \geqslant c^2$ is to make sure that the Mahalanobis distance between two classes is at least c. Thus the smaller the value of c, the harder the classification problem is.

Given independent training data $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, the common covariance matrix can be estimated by using the sample covariance matrix

$$\widehat{\mathbf{\Sigma}} = \sum_{k=1}^{K} \sum_{Y_i = k} (\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_k)^T / (n - K). \tag{3.2}$$

For the mean vectors, Bickel and Levina (2004) showed that there exist estimators $\widetilde{\boldsymbol{\mu}}_k$ of $\boldsymbol{\mu}_k$, $k = 1, 2$ such that

$$\max_{\Gamma_1} E_{\boldsymbol{\theta}} \|\widetilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|^2 = o(1). \tag{3.3}$$

Replacing the population parameters in the definition of δ_F by the above estimators $\widetilde{\boldsymbol{\mu}}_k$ and $\widehat{\mathbf{\Sigma}}$, we obtain the sample version of Fisher discriminant function $\hat{\delta}_F$.

It is well known that for fixed p, the worst case misclassification rate of $\hat{\delta}_F$ converges to the worst case Bayes risk over Γ_1, that is,

$$\overline{W}_{\Gamma_1}(\hat{\delta}_F) \to \overline{\Phi}(c/2), \text{ as } n \to \infty, \tag{3.4}$$

where $\overline{\Phi}(t) = 1 - \Phi(t)$ is the tail probability of the standard Gaussian distribution. Hence, $\hat{\delta}_F$ is asymptotically optimal for this low dimensional problem. However, in high dimensional setting, the result is very different.

Bickel and Levina (2004) studied the worst case misclassification rate of $\hat{\delta}_F$ when $n_1 = n_2$ in high dimensional setting. Specifically they showed that under some regularity conditions, if $p/n \to \infty$, then

$$\overline{W}_{\Gamma_1}(\hat{\delta}_F) \to \frac{1}{2}, \tag{3.5}$$

where the Moore-Penrose generalized inverse is used in the definition of $\hat{\delta}_F$. Note that 1/2 is the misclassification rate of random guessing. Thus although Fisher discriminant analysis is asymptotically optimal and has Bayes risk when dimension p is fixed and sample size $n \to \infty$, it performs asymptotically no better than random guessing when the dimensionality p is much larger than the sample size n. This shows the difficulty of high dimensional classification. As have been demonstrated by Bickel and Levina (2004) and pointed out by Fan and Fan (2008), the bad performance of Fisher discriminant analysis is due to the diverging spectra (e.g., the condition number goes to infinity as dimensionality diverges) frequently encountered in the estimation of high-dimensional covariance matrices. In fact, even if the true covariance matrix is not ill conditioned, the singularity of the sample covariance matrix will make the Fisher discrimination rule inapplicable when the dimensionality is larger than the sample size.

3.2 Impact of dimensionality on independence rule

Fan and Fan (2008) studied the impact of high dimensionality on classification. They pointed out that the difficulty of high dimensional classification is intrinsically caused by the existence of many noise features that do not contribute to the