High-Dimensional Classification Methods for Sparse Signals and Their Applications in Text and Data Mining

by

Dawit G. Tadesse

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 02, 2014

Keywords: Feature Selection, Fisher discriminant, High-Dimensional classification, Naive Bayes, Sparse Signals, Text Mining.

Approved by

Mark Carpenter, Chair, Professor of Mathematics and Statistics
Asheber Abebe, Associate Professor of Mathematics and Statistics
Xiaoyu Li, Assistant Professor of Mathematics and Statistics
Guanqun Cao, Assistant Professor of Mathematics and Statistics

Abstract

Classification using high-dimensional features arises frequently in many contemporary statistical studies such as tumor classification using microarray or other high-throughput data. In this dissertation we conduct a rigorous performance analysis of the two linear methods for high-dimensional classification, Independence Rule (or Naive Bayes) and Fisher discriminant both in theory and simulation. We know that, for the normal population model, when all the parameters are known Fisher is optimal and Naive Bayes is suboptimal. But in this dissertation we give the conditions under which Naive Bayes is optimal. Through theory and simulation, we further, show that Naive Bayes performs better than Fisher under broader conditions. We also study the associated feature selection methods. The two-sample t-test is a widely popular feature selection method. But it heavily depends on the normality assumption so we proposed a generalized feature selection algorithm which works regardless of the distribution. Our generalized feature selection is a special case of two-sample t-test, Wilcoxon-Mann Whitney Statistic and two-sample proportion statistic. We know that Singular Value Decomposition(SVD) is a popular dimension reduction method in text mining problems. Researchers take the first few SVDs which explain the largest variation. However, in this dissertation we argue that the first few SVDs are not necessarily the most important ones for classification. We then give a new feature selection algorithm for the data matrix in text mining problem in high-dimensional spaces.

Acknowledgments

This dissertation would not have been possible without the support of many people. I wish to express my sincere gratitude to my supervisor, Dr. Mark Carpenter who was abundantly helpful and offered invaluable assistance, support and guidance. It has been privilege working with you. Deepest gratitude are also due to the members of the supervisory committee, Drs. Asheber Abebe, Xiaoyu Li, Guanqun Cao, and Roy Hartfield without whose knowledge and assistance this study would not have been successful. I wish thank my mentor Prof. Charles Chidume of the African University of Science and Technology, Abuja, Nigeria who gave me the opportunity to study at Auburn. I also would like to thank my Auburn professors Drs. Erkan Nane, Huajun Huang, Tin-Yau Tam, Geraldo De Souza, and Narendra Govil for their support.

I would like to thank my father Gezahegn Tadesse who gave me the motivation to start the PhD and for the support he had been giving me throughout my study. My dear friends Nar Rawal, Mulugeta Woldu, Chabbi Adhikari, Bretford Griffin, Achard Bindele, Fasil Mulat, Dawit Befekadu, Simon Atsbeha, Lebanos Woldu, Eze Nwaeze, Seth Kermasour thank you all for the good and fun time i spent with you at Auburn.

I dedicate this dissertation to my late mother Belaynesh Yimam. I know you would have been very proud if you lived to see this day.

Table of Contents

Abstract ii

Acknowledgments iii

List of Figures vi

1 Introduction 1

1.1 Elements of Classification 2

1.2 Organization 2

1.3 Notations 2

2 On High-Dimensional Classification for Sparse Signals 4

2.1 High-Dimensional Classification 4

2.2 Classification with Sparse Signals 5

2.2.1 Sample Model 13

2.2.2 Sample Misclassification Error Rates for Naive Bayes 15

2.3 Univariate and Multivariate t distribution 19

2.4 Feature Selection: two-sample t-test 21

2.5 Simulation Results 30

2.6 Real Data Analysis: Leukemia Data 35

2.7 Conclusion 37

3 Generalized Feature Selection 38

3.1 Introduction 38

3.2 The Generalized Feature Selection 39

3.3 Conclusion 43

4 Applications of High-Dimensional Classification in Text Mining 44

4.1 Introduction 44

4.2 Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) 48
4.2.1 Comparing SVD and PCA 50
4.2.2 Sparse vectors for SVDs 51
4.2.3 Fisher and Naive Bayes Discriminant Functions for SVDs 51
4.3 Partitioning of the Data Matrix into Training and Validation Data Matrices 57
4.3.1 Singular Value Decomposition of the Training Data Matrix 57
4.3.2 Sorting Features Based on T-statistics on the Training Data 59
4.4 Overall Prediction Modeling 60
4.4.1 Feature Selection Algorithm 60
4.5 Real Data Analysis 60
4.5.1 NASA flight data set 60
4.5.2 DBWorld Email Messages 63
4.6 Conclusion 65
5 Summary and Future Work 66
5.1 Summary 66
5.2 Future Work 67
Bibliography 68

List of Figures

2.1 *The horizontal axis is the mean difference (α) and the vertical axis is the maximum misclassification error rates. The different lines are the maximum error rates of Fisher, $W(\delta_F, \boldsymbol{\theta}^{(m)})$, using the first $m = 90$ features for $\rho = 0, 0.1, \ldots, 0.9$ from bottom to top respectively.* 13

2.2 *The horizontal axis is difference of the means (α) and the vertical is error rates. The different lines are theoretical maximum error rate vs upper bound on the sample error rate for Naive Bayes when $\rho = 0.1, 0.5, 0.9$ from bottom to top respectively using the first $m = 90$ features.* 19

2.3 *Horizontal axis is the mean difference (α) and vertical is probability of getting all the important s features in the first $s, s+1, s+2, s+3, 3s/2, 2s$ t-statistics respectively. The results are for $\rho = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ from bottom to top respectively.* 25

2.4 *Horizontal axis is the number we go after s (call it, m') and the vertical axis is probability of getting all the important s features in the first $s+m'$ t-statistics for $\rho = 0.5, 0.9$ respectively. The different lines are for $\alpha = 1, 1.2, 1.4, 1.6, 1.8, 2$ from bottom to top respectively.* 26

2.5 *Horizontal axis is the mean difference (α). The vertical axis is testing misclassification error rate using the first $10, 30, 45$ features for Naive Bayes vs Fisher. The first and second figures are when $\rho = 0.5, 0.9$ respectively.* 31

2.6 *Horizontal axis is the mean difference (α) and the vertical axis is testing misclassification error rate.* 32

2.7 *Horizontal axis is the mean difference (α). The vertical axis is testing misclassification error rate using the first 10 and 45 features for Naive Bayes vs Fisher.* 33

2.8 *Horizontal axis is the number of features used divided by $s = 90$ and the vertical axis is testing misclassification error rate. The different lines (starting from up) are for $\alpha = 1.0, 1.5, 2.0, 2.5, 3.0$ respectively.* 34

2.9 *Horizontal axis is the mean difference (α) and the vertical axis is testing misclassification error rate using the first $90, 100, 135, 180$ and 270 features.* 34

2.10 *Horizontal axis is the number of genes used and the vertical axis is testing misclassification error rate for Fisher vs Naive Bayes.* 36

2.11 *Horizontal axis is the number of genes used and the vertical axis is testing misclassification error rate for Naive Bayes.* 36

4.1 *The horizontal axis is s and vertical axis is $f(s)$.* 56

4.2 *Horizontal axis is the number of features used and the vertical axis is testing misclassification error rate for Fisher vs Naive Bayes.* 61

4.3 *Horizontal axis is the number of svds used and the vertical axis is testing misclassification error rate for Naive Bayes vs Fisher.* 62

4.4 Error rate vs number of svds for subjects. We can see that ranking the svds based on their t-statistic improves the error rate. We need also fewer number of svds. 64

4.5 Error rate vs number of svds for bodies. We can see that ranking the svds based on their t-statistic improves the error rate. We need also fewer number of svds. 64

Chapter 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, Naive Bayes classifier, k-th-nearest-neighbor classifier, support vector machines, and many others. When the dimensionality of the input feature space is large, things become complicated. Fan and Fan (2008) study 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 reduction of classification error. For example, for the Fisher discriminant analysis, one needs to estimate the class mean vectors and covariance matrix. Although individually each parameter can be estimated accurately, aggregated estimation error over many features can be very large and this could significantly increase the misclassification rate. This is another important reason that causes the bad performance of Fisher discriminant analysis in high dimensional setting. Greenshtein and Ritov (2004) and Greenshtein (2006) introduced and studied the concept of persistence, which places more emphasis on misclassification rates or expected loss rather than the accuracy of estimated parameters. In high dimensional classification, since we care much more about the misclassification rate instead of the accuracy of the estimated parameters, estimating the full covariance matrix and the class mean vectors will result in very high accumulation error and thus low classification accuracy.

1.1 Elements of Classification

Suppose we have some input space $\mathcal{X}$ and some output space $\mathcal{Y}$. Assume that there are independent training data $(\boldsymbol{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, n$ coming from some unknown distribution P, where Y_i is the i^{th} observation of the response variable and $\boldsymbol{X}_i$ is its associated feature or covariate vector. In classification problems, the response variable Y_i is qualitative and the set Y has only finite values. For example, in the cancer classification using gene expression data, each feature vector $\boldsymbol{X}_i$ 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 $\boldsymbol{X}$, 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. In this dissertation, we consider the case when $K = 2$.

1.2 Organization

This dissertation is organized as follows: in chapter 2 we will set up the classification problem, give some theories on misclassification error rates for Fisher and Naive Bayes, introduce the sample model and we give some bounds on the sample misclassification error rate of Naive Bayes, we study the associated feature selection methods, we will present some simulation results. In chapter 3 we study our new generalized feature selection method. In chapter 4, we present some results on the applications of high-dimensional classification in text mining. In chapter 5, we give the summary for the dissertation and we present our future work.

1.3 Notations

Here are some notations i used throughout this dissertation:

• $\boldsymbol{X}_1$ and $\boldsymbol{X}_0$ are two random variables from the corresponding classes $\mathcal{C}_1$ and $\mathcal{C}_0$ with sample sizes n_1 and n_0 respectively.

• p is the number of features (variables) and s is the number of features with non zero mean difference.

• n is the total sample size for the two classes.

• $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ are the mean vectors of classes $\mathcal{C}_1$ and $\mathcal{C}_0$ respectively. $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_0$ are the sample mean vectors of classes $\mathcal{C}_1$ and $\mathcal{C}_0$ respectively.

• Σ_1 and Σ_0 are the covariance matrices for classes $\mathcal{C}_1$ and $\mathcal{C}_0$ respectively. $\hat{\Sigma}_1$ and $\hat{\Sigma}_0$ are the sample covariance matrices for classes $\mathcal{C}_1$ and $\mathcal{C}_0$ respectively.

• Σ is the common covariance matrix. $\hat{\Sigma}$ is the sample common covariance matrix.

• $\boldsymbol{\rho}$ is the common correlation matrix. $\boldsymbol{\rho}^{(m)}$ is the truncated $m \times m$ common correlation matrix.

• D is the diagonal matrix for Σ. In other words, D is the variance matrix. $\hat{D}$ is the sample version.

• $\delta_{NB}(\cdot)$ and $\delta_F(\cdot)$ are the discriminant functions for Naive Bayes and Fisher respectively. $\hat{\delta}_{NB}(\cdot)$ and $\hat{\delta}_F(\cdot)$ are the sample discriminant functions for Naive Bayes and Fisher respectively.

• $W(\delta_{NB}, \cdot)$ and $W(\delta_F, \cdot)$ are the misclassification error rates for Naive Bayes and Fisher respectively. $W\left(\hat{\delta}_{NB}, \cdot\right)$ and $W\left(\hat{\delta}_F, \cdot\right)$ are the sample misclassification error rates for Naive Bayes and Fisher respectively.

• $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ are the smallest and largest eigenvalues.

• $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_0$ are two singular value decomposition (svd) random variables from the corresponding classes $\mathcal{C}_1$ and $\mathcal{C}_0$ with sample sizes n_1 and n_0 respectively.

Chapter 2

On High-Dimensional Classification for Sparse Signals

2.1 High-Dimensional Classification

Technological innovations have had deep impact on society and on various areas of scientific research. High-throughput data from microarray and proteomics technologies are frequently used in many contemporary statistical studies (see Dudoit et al. (2002)). In the case of microarray data, the dimensionality is frequently in thousands or beyond, while the sample size is typically in the order of tens. The large-p-small-n scenario poses challenges for the classification problems. We refer to Fan and Lv (2010) for an overview of statistical challenges associated with high dimensionality (Fan and et al. (2012)).

When the feature space dimension p is very high compared to the sample size n, the Fisher discriminant rule performs poorly due to diverging spectra as demonstrated by Bickel and Levina (2004). These authors showed that the independence rule in which the covariance structure is ignored performs better than the naive Fisher rule (NFR) in the high dimensional setting. Fan and Fan (2008) demonstrated further that even for the independence rules, a procedure using all the features can be as poor as random guessing due to noise accumulation in estimating population centroids in high-dimensional feature space. As a result, Fan and Fan (2008) proposed the Features Annealed Independence Rule (FAIR) that selects a subset of important features for classification. Dudoit et al. (2002) reported that for microarray data, ignoring correlations between genes leads to better classification results. But recent works try to show that the independence rule may lead to higher misclassification error rates when there is correlation among the variables (for example, Fan and et al (2012)). In this dissertation, we show that even under high correlations independence rule (or Naive Bayes) can still dominate the Fisher rule at the sample level using subset of the features.

2.2 Classification with Sparse Signals

We introduce the objective classification problem. We assume in what follows that the variability of data under consideration can be described reasonably well by mean vector $\boldsymbol{\mu}$ and variance-covariance matrix Σ. Suppose that the random variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_0$ representing two classes $\mathcal{C}_1$ with mean vector $\boldsymbol{\mu}_1$ and $\mathcal{C}_0$ with mean vector $\boldsymbol{\mu}_0$ follow $p-$variate distributions with densities $f(\boldsymbol{X}|\boldsymbol{\theta}_1)$ and $f(\boldsymbol{X}|\boldsymbol{\theta}_0)$ respectively with Σ the common covariance matrix where $\boldsymbol{\theta}_i \in \boldsymbol{\Theta} = \{(\boldsymbol{\mu}_i, \Sigma) : \boldsymbol{\mu}_i \in \mathbb{R}^p, \det(\Sigma) > 0,\ i = 0, 1\}$ is the parameter space consisting of the mean vectors and the common covariance matrix. In other words,

$$\boldsymbol{X}_i \sim f_i(\boldsymbol{X}|\boldsymbol{\theta}_i) = f(\boldsymbol{X}|\boldsymbol{\theta}_i), \quad i = 0, 1. \tag{2.1}$$

Suppose that

$$\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, \quad \boldsymbol{\mu}_a = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)/2, D = diag(\Sigma) \tag{2.2}$$

Let π_0 and π_1 be the class prior probabilities for classes $\mathcal{C}_0$ and $\mathcal{C}_1$ respectively. A new observation $\boldsymbol{X}$ is to be assigned to one of $\mathcal{C}_1$ or $\mathcal{C}_0$. The optimal classifier is the Bayes rule:

$$\delta(\boldsymbol{X}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_0) = \mathbf{1}\left\{\log \frac{f(\boldsymbol{X}|\boldsymbol{\theta}_1)}{f(\boldsymbol{X}|\boldsymbol{\theta}_0)} > \log \frac{\pi_0}{\pi_1}\right\}, \tag{2.3}$$

where $\mathbf{1}$ denotes the indicator function with value 1 corresponds to assigning $\boldsymbol{X}$ to $\mathcal{C}_1$ and 0 to class $\mathcal{C}_0$.

Unless specified, throughout this section we let that $\boldsymbol{X}_1 \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \Sigma)$ and $\boldsymbol{X}_0 \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \Sigma)$. Under these assumptions (2.3) becomes

$$\delta(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, \Sigma) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T \Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_a) > \frac{\pi_0}{\pi_1}\right\}. \tag{2.4}$$

We propose the family of discriminant functions given by

$$\delta(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, M) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T M^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_a) > \frac{\pi_0}{\pi_1}\right\}, \tag{2.5}$$

where M is a $p \times p$ symmetric positive definite matrix.

We define the misclassification error rate of $\delta(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, M)$ as the following sum of posterior probabilities

$$W(\delta, \boldsymbol{\theta}) = \pi_1 P(\delta(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, M) = 0 | \boldsymbol{X} \in \mathcal{C}_1) + \pi_0 P(\delta(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, M) = 1 | \boldsymbol{X} \in \mathcal{C}_0), \tag{2.6}$$

where $\boldsymbol{\theta} \in \{(\boldsymbol{\mu}_d, \Sigma), \boldsymbol{\mu}_d \in \mathbb{R}^p, \det(\Sigma) > 0\}$.

It is easy to show that the misclassification error rate of $\delta(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, M)$ (when $\pi_1 = \pi_0 = 1/2$) is given below

$$W(\delta, \boldsymbol{\theta}) = \bar{\Phi}\left(\frac{\boldsymbol{\mu}_d^T M^{-1} \boldsymbol{\mu}_d}{2(\boldsymbol{\mu}_d^T M^{-1} \Sigma M^{-1} \boldsymbol{\mu}_d)^{1/2}}\right), \tag{2.7}$$

where $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot), \Phi(\cdot)$ is a CDF for standard normal distribution.

Note that if $M = \Sigma$ and $\pi_1 = \pi_0 = 1/2$, then we have the Fisher discriminant rule

$$\delta_F(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, \Sigma) = \delta(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, \Sigma) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T \Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_a) > 0\right\}, \tag{2.8}$$

with corresponding misclassification error rate

$$W(\delta_F, \boldsymbol{\theta}) = \bar{\Phi}\left(\frac{(\boldsymbol{\mu}_d^T \Sigma^{-1} \boldsymbol{\mu}_d)^{1/2}}{2}\right). \tag{2.9}$$

Alternatively, assuming independence of components and replacing off-diagonal elements of Σ with zeros leads to a new covariance matrix

$$D = \text{diag}(\Sigma), \tag{2.10}$$

and a different discrimination rule, the Naive Bayes,

$$\delta_{NB}(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, D) = \delta(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, D) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T D^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_a) > 0\right\}, \tag{2.11}$$

whose misclassification error rate is

$$W(\delta_{NB}, \boldsymbol{\theta}) = \bar{\Phi}\left(\frac{\boldsymbol{\mu}_d^T D^{-1} \boldsymbol{\mu}_d}{2(\boldsymbol{\mu}_d^T D^{-1} \Sigma D^{-1} \boldsymbol{\mu}_d)^{1/2}}\right). \tag{2.12}$$

We define sparse vector and signal as follows:

Definition 1. *Suppose that* $\boldsymbol{\mu}_d = (\alpha_1, \alpha_2, \ldots, \alpha_s, 0, \ldots, 0)^T$ *is the* $p \times 1$ *mean difference vector where* $\alpha_j \in \mathbb{R}\backslash\{0\}, j = 1, 2, \ldots, s$. *We say that* $\boldsymbol{\mu}_d$ *is sparse if* $s = o(p)$. *Signal is defined as* $C_s = \boldsymbol{\mu}_d^T D^{-1} \boldsymbol{\mu}_d = \sum_{j=1}^{s} \frac{\alpha_j^2}{\sigma_j^2}$ *where* σ_j^2 *is the common variance for feature* j *in the two classes.*

Some Examples of Sparse situations in real life:

• Gene Expression data (take p genes from two group of patients, most of them are the same for the two group of patients and only s of them are different).

• Author Identification (two documents from two authors and they use equal proportion of many words and there are only s few words which separate them).

Note that Σ can be partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{11} = \Sigma^{(m)}$ is the $m \times m$ truncated covariance matrix. We denote $D_{kj} = \text{diag}(\Sigma_{kj})$ and D_{11} by $D^{(m)}$. Similarly, the sparse mean difference vector $\boldsymbol{\mu}_d$ can be partitioned as

$$\boldsymbol{\mu}_d = \begin{pmatrix} \boldsymbol{\mu}_d^{(m)} \\ \boldsymbol{\mu}_d^{(s-m)} \\ \mathbf{0}_{p-s} \end{pmatrix},$$

assuming that $m \leq s$.

Define $\boldsymbol{\rho}^{(m)} = D_{11}^{-1/2} \Sigma_{11} D_{11}^{-1/2}$ which is the $m \times m$ truncated correlation matrix. We say that $\boldsymbol{\rho}^{(m)}$ is the $m \times m$ equicorrelation matrix when $\boldsymbol{\rho}^{(m)} = \Big(\rho_{kj}\Big), \rho_{kj} = 1$ if $k = j$ and $\rho_{kj} = \rho \in [0, 1)$ otherwise.

In this dissertation we focused mainly on equicorrelation matrices as the they give us bounds for the general correlation matrices. Also, when the general correlation structure is unknown, we can design our experiments using equicorrelation structures.

Lemma 1. ***(a)*** *If $\boldsymbol{\rho}^{(m)}$ is the equicorrelation matrix defined above, the eigenvalues of $\boldsymbol{\rho}^{(m)}$ are $\lambda_1(\boldsymbol{\rho}^{(m)}) = \cdots = \lambda_{m-1}(\boldsymbol{\rho}^{(m)}) = 1-\rho$ and $\lambda_m(\boldsymbol{\rho}^{(m)}) = 1+(m-1)\rho$. Note that the smallest eigenvalue is $\lambda_{\min}(\boldsymbol{\rho}^{(m)}) = 1 - \rho$ and the largest eigenvalue is $\lambda_{\max}(\boldsymbol{\rho}^{(m)}) = 1 + (m-1)\rho$.*

(b) *Let $\mathbf{1} = (1, 1, \ldots, 1)^T$ is the $m \times 1$ vector of ones and define $B = \mathbf{1}\mathbf{1}^T/m$ which is the $m \times m$ idempotent matrix. Note that $\boldsymbol{\rho}^{(m)}$ can be written as $\boldsymbol{\rho}^{(m)} = (1 + (m-1)\rho)B + (1-\rho)(I_m - B)$. The inverse of $\boldsymbol{\rho}^{(m)}$ is $(\boldsymbol{\rho}^{(m)})^{-1} = \frac{1}{1+(m-1)\rho} B + \frac{1}{1-\rho}(I_m - B)$ where I_m is the $m \times m$ identity matrix.*

Proof of Lemma 1: For **(a)** and **(b)** see Abadir and Magnus (2005, p241). □

Lemma 2. *If $k\Sigma^{-1}\boldsymbol{\mu}_d = M^{-1}\boldsymbol{\mu}_d$ for some constant $k \neq 0$, then*

$$W(\delta_F, \boldsymbol{\theta}) = W(\delta, \boldsymbol{\theta}).$$

Proof of Lemma 2: From equations (2.7) and (2.9) and using the condition of lemma 2,

$$W(\delta, \boldsymbol{\theta}) = \bar{\Phi}\left(\frac{\boldsymbol{\mu}_d^T M^{-1}\boldsymbol{\mu}_d}{2(\boldsymbol{\mu}_d^T M^{-1}\Sigma M^{-1}\boldsymbol{\mu}_d)^{1/2}}\right) = \bar{\Phi}\left(\frac{k\boldsymbol{\mu}_d^T \Sigma^{-1}\boldsymbol{\mu}_d}{2k(\boldsymbol{\mu}_d^T \Sigma^{-1}\boldsymbol{\mu}_d)^{1/2}}\right) = \bar{\Phi}\left(\frac{(\boldsymbol{\mu}_d^T \Sigma^{-1}\boldsymbol{\mu}_d)^{1/2}}{2}\right) = W(\delta_F, \boldsymbol{\theta}).$$

□

Note that if Σ is an equicorrelation matrix and $M = D$ then lemma 2 means that Naive Bayes and Fisher have the same error rates if $\boldsymbol{\mu}_d$ is an eigenvector for Σ. Specifically, we have the following result.

Theorem 2.1. *If* $m \leq s, \boldsymbol{\mu}_d^{(m)} = (\alpha, \alpha, \ldots, \alpha)^T = \alpha\mathbf{1}, \alpha \neq 0$ *and* $\Sigma^{(m)}$ *is the truncated* $m \times m$ *equicorrelation matrix, then we have*

$$W(\delta_F, \boldsymbol{\theta}^{(m)}) = W(\delta_{NB}, \boldsymbol{\theta}^{(m)}),$$

where $\boldsymbol{\theta}^{(m)}$ *is the truncated parameter.*

Proof of Theorem 2.1: Let us assume that the off-diagonals for $\Sigma^{(m)}$ are each ρ. Note that $D^{(m)} = I_m$. We use the fact that $\mathbf{1}^T\mathbf{1} = m$.

If we take $k = 1 + (m-1)\rho$, using lemma 1 **(a)** and **(b)**

$$\begin{aligned} k(\Sigma^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)} &= k\alpha\left(\tfrac{1}{1+(m-1)\rho}\mathbf{1}\mathbf{1}^T/m + \tfrac{1}{1-\rho}(I_m - \mathbf{1}\mathbf{1}^T/m)\right)\mathbf{1} \\ &= \alpha\mathbf{1} = I_m\boldsymbol{\mu}_d^{(m)} \\ &= (D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)}. \end{aligned}$$

Using lemma 2, the result follows by taking $M = D^{(m)}$. □

We know that $\Sigma^{(m)} = \left(\sigma_{kj}\right), 1 \leq k \leq j \leq m$, let us consider the following $m \times m$ positive definite symmetric correlation matrix, $\boldsymbol{\rho}^{(m)} = (D^{(m)})^{-1/2}\Sigma^{(m)}(D^{(m)})^{-1/2}$, defined as $\boldsymbol{\rho}^{(m)} = \left(\rho_{kj}\right), \rho_{kj} = 1$ when $k = j$ and $\rho_{kj} = \rho_{jk} \in \mathbb{R}$ when $k \neq j$. Let us define

$\bar{\rho} = \sum_{k \neq j} \frac{\rho_{kj}}{m(m-1)}$ and $\rho_{\max} = \max_{k \neq j} |\rho_{kj}|$. Define $\bar{\boldsymbol{\rho}}^{(m)} = \left(\rho_{kj}\right), \rho_{kj} = 1$ when $k = j$ and $\rho_{kj} = \bar{\rho}$ when $k \neq j$ and $\boldsymbol{\rho}_{\max}^{(m)} = \left(\rho_{kj}\right), \rho_{kj} = 1$ when $k = j$ and $\rho_{kj} = \rho_{\max}$ when $k \neq j$. In other words, $\bar{\boldsymbol{\rho}}^{(m)}$ and $\boldsymbol{\rho}_{\max}^{(m)}$ are equicorrelation matrices with off diagonals the mean of the correlation coefficients and largest of the absolute values of the correlation coefficients, ρ_{kj} when $k \neq j$ respectively. We use the notation $\sigma_{kj} = \sigma_j^2$ when $k = j$.

Definition 2. *If A and B are positive definite matrices of the same size, then we will write $A \leq B$ if each entry of $B - A$ is non-negative. $|A|$ is the matrix we get from A by replacing each entry of A by their absolute values. The spectral radius of matrix A is $\lambda_{\max}(A)$.*

Lemma 3. *If $|A| \leq B$, then $\lambda_{\max}(A) \leq \lambda_{\max}(|A|) \leq \lambda_{\max}(B)$ which implies that $\lambda_{\max}(\boldsymbol{\rho}^{(m)}) \leq \lambda_{\max}(\boldsymbol{\rho}_{\max}^{(m)})$.*

Proof of Lemma 3: See Horn and Johnson (1985, p491). □

Lemma 4. *Let M be the truncated $m \times m$ positive definite matrix with eigenvalues $\lambda_1(M) \leq \lambda_2(M) \leq \cdots \leq \lambda_m(M)$ so that $\lambda_1(M) = \lambda_{\min}(M)$ and $\lambda_m(M) = \lambda_{\max}(M)$. Then, for $\boldsymbol{X} \in \mathbb{R}^m$, we have*

$$\lambda_{\max}(M) = \max_{\boldsymbol{X} \neq \boldsymbol{0}} \frac{\boldsymbol{X}^T M \boldsymbol{X}}{\boldsymbol{X}^T \boldsymbol{X}} \quad \text{and} \quad \lambda_{\min}(M) = \min_{\boldsymbol{X} \neq \boldsymbol{0}} \frac{\boldsymbol{X}^T M \boldsymbol{X}}{\boldsymbol{X}^T \boldsymbol{X}}$$

Proof of Lemma 4: See Johnson and Wichern (2007, p80). □

Theorem 2.2. *Suppose $\boldsymbol{\rho}^{(m)}$ is an $m \times m$ correlation matrix and $\boldsymbol{\mu}_d^{(m)}$ is an $m \times 1$ mean difference vector. Then, we have the following bounds on the error rates of Fisher and Naive Bayes which are given in equations* (2.9) *and* (2.12) *respectively:*

(a)

$$\bar{\Phi}\left(\frac{\sqrt{(\boldsymbol{\mu}_d^{(m)})^T (D^{(m)})^{-1} \boldsymbol{\mu}_d^{(m)}}}{2\sqrt{\lambda_{\min}(\boldsymbol{\rho}^{(m)})}}\right) \leq W(\delta_w, \boldsymbol{\theta}^{(m)}) \leq \bar{\Phi}\left(\frac{\sqrt{(\boldsymbol{\mu}_d^{(m)})^T (D^{(m)})^{-1} \boldsymbol{\mu}_d^{(m)}}}{2\sqrt{\lambda_{\max}(\boldsymbol{\rho}^{(m)})}}\right),$$

*(**b**) Suppose, further, that* $\lambda_{\min}(\boldsymbol{\rho}^{(m)}) \geq \lambda_{\min}(\bar{\boldsymbol{\rho}}^{(m)}) = 1 - \bar{\rho}$. *Then*

$$\bar{\Phi}\left(\frac{\sqrt{(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)}}}{2\sqrt{1-\bar{\rho}}}\right) \leq W(\delta_w, \boldsymbol{\theta}^{(m)}) \leq \bar{\Phi}\left(\frac{\sqrt{(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)}}}{2\sqrt{1+(m-1)\rho_{\max}}}\right)$$

where $w = F$ *or* $w = NB$ *for the truncated parameter* $\boldsymbol{\theta}^{(m)}$.

Proof of Theorem 2.2: (**a**) Note that the numerator for Fisher can be written as $(\boldsymbol{\mu}_d^{(m)})^T(\Sigma^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)} = (\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1/2}(\boldsymbol{\rho}^{(m)})^{-1}(D^{(m)})^{-1/2}\boldsymbol{\mu}_d^{(m)}$. Using lemma 4 we have,

$$\begin{aligned}\lambda_{\min}((\boldsymbol{\rho}^{(m)})^{-1})(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)} &\leq (\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1/2}(\boldsymbol{\rho}^{(m)})^{-1}(D^{(m)})^{-1/2}\boldsymbol{\mu}_d^{(m)} \\ &\leq \lambda_{\max}((\boldsymbol{\rho}^{(m)})^{-1})(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)}\end{aligned}$$

which implies that

$$\begin{aligned}\tfrac{1}{\lambda_{\max}((\boldsymbol{\rho}^{(m)})}(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)} &\leq (\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1/2}(\boldsymbol{\rho}^{(m)})^{-1}(D^{(m)})^{-1/2}\boldsymbol{\mu}_d^{(m)} \\ &\leq \tfrac{1}{\lambda_{\min}((\boldsymbol{\rho}^{(m)})}(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)}.\end{aligned}$$

Noting that $\bar{\Phi}(\sqrt{x}/2)$ is a decreasing function of x we have the inequalities for Fisher.

Similarly, the denominator which is inside the square root for Naive Bayes can be written $(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\Sigma^{(m)}(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)} = (\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1/2}\boldsymbol{\rho}^{(m)}(D^{(m)})^{-1/2}\boldsymbol{\mu}_d^{(m)}$. Using lemma 4 we have,

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\rho}^{(m)})(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)} &\leq (\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1/2}(\boldsymbol{\rho}^{(m)})^{-1}(D^{(m)})^{-1/2}\boldsymbol{\mu}_d^{(m)} \\ &\leq \lambda_{\max}(\boldsymbol{\rho}^{(m)})(\boldsymbol{\mu}_d^{(m)})^T(D^{(m)})^{-1}\boldsymbol{\mu}_d^{(m)}.\end{aligned}$$

Noting that $\bar{\Phi}(K/\sqrt{x})$ is an increasing function of x when $K > 0$ we have the inequalities for Naive Bayes.

(**b**) Note that $\lambda_{\min}(\bar{\boldsymbol{\rho}}^{(m)}) = 1 - \bar{\rho}$ and using lemma 3 we have

$$\lambda_{\min}(\bar{\boldsymbol{\rho}}^{(m)}) \leq \lambda_{\min}(\boldsymbol{\rho}^{(m)}) \leq \lambda_{\max}(\boldsymbol{\rho}^{(m)}) \leq \lambda_{\max}(\boldsymbol{\rho}_{\max}^{(m)}).$$

Using the bounds in part (**a**) and noting that $\bar{\Phi}(K/\sqrt{x})$ is an increasing function of x when $K > 0$ we have the inequalities in part (**b**). □

We need the bounds in theorem 2.2 (**b**), because in experimental designs we do not know the actual correlations, if we know the maximum correlation, we can give the maximum error rate for Fisher.

Corollary 1. *If $m \leq s$, then*

$$\max_{\boldsymbol{\theta}^{(m)}} W(\delta_F, \boldsymbol{\theta}^{(m)}) = W(\delta_F, \boldsymbol{\theta}_2^{(m)}) = W(\delta_{NB}, \boldsymbol{\theta}_2^{(m)})$$

where $\boldsymbol{\theta}_2^{(m)}$ is the parameter which consists of the equal mean difference vectors and equicorrelation matrix.

Proof of Corollary 1: Using theorem 2.2 (**a**), it is easy to see that the upper bound for Fisher is achieved when we have equal mean difference vectors and equicorrelation matrix. Therefore, $\max_{\boldsymbol{\theta}^{(m)}} W(\delta_F, \boldsymbol{\theta}^{(m)}) = W(\delta_F, \boldsymbol{\theta}_2^{(m)})$ where $\boldsymbol{\theta}_2^{(m)}$ is the parameter which consists of the equal mean difference vectors and equicorrelation matrix. But using theorem 2.1, $W(\delta_F, \boldsymbol{\theta}_2^{(m)}) = W(\delta_{NB}, \boldsymbol{\theta}_2^{(m)})$. Hence, $\max_{\boldsymbol{\theta}^{(m)}} W(\delta_F, \boldsymbol{\theta}^{(m)}) = W(\delta_{NB}, \boldsymbol{\theta}_2^{(m)})$. □

Note that since Fisher is optimal, from corollary 1, the misclassification error rate for Naive Bayes can be taken as a minimax estimator for the misclassification error rate for Fisher over the parameter $\boldsymbol{\theta}^{(m)}$. Also, the maximum error rates for both Fisher and Naive Bayes occur when we have equal mean difference vector and equicorrelation matrix.

For example, let $\boldsymbol{\mu}_1 = (\boldsymbol{\alpha}_{90}, \mathbf{0}_{4410})^T, \boldsymbol{\mu}_0 = (\mathbf{0}_{4500})^T, \sigma_{kk} = 1, \sigma_{kj} = \rho \in [0, 1), k \neq j$. Then we have the following graphs for the error rates of Fisher in equation (2.9)=Naive Bayes in equation (2.12) using the first $m = 90$ features against the mean difference α for several ρ.

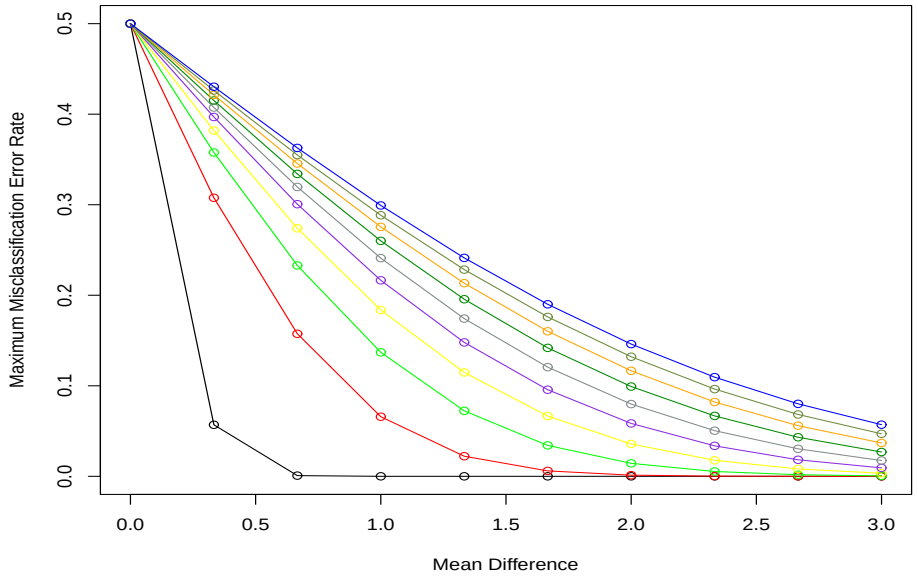


Figure 2.1: *The horizontal axis is the mean difference (α) and the vertical axis is the maximum misclassification error rates. The different lines are the maximum error rates of Fisher, $W(\delta_F, \boldsymbol{\theta}^{(m)})$, using the first $m = 90$ features for $\rho = 0, 0.1, \ldots, 0.9$ from bottom to top respectively.*

As we can see from the above figure, even at $\rho = 0.7$ with $\alpha = 1$ standard deviation difference, the maximum misclassification error rate is around 30%. Using theorem 2.2, we know that any correlation matrix with maximum absolute value correlation coefficient of 0.7 would give lower than 30% misclassification error rate.

2.2.1 Sample Model

Suppose that $\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}$ are random samples coming from a distribution with density function $f(\boldsymbol{X}|(\boldsymbol{\mu}_1, \Sigma))$ and $\boldsymbol{X}_{01}, \boldsymbol{X}_{02}, \ldots, \boldsymbol{X}_{0n_0}$ are random samples coming from a distribution with density function $f(\boldsymbol{X}|(\boldsymbol{\mu}_0, \Sigma))$.

The independence rule depends on the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_0$ and $D = \text{diag}\{\sigma_1^2, \ldots, \sigma_p^2\}$ and Fisher depends on the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_0$ and Σ. They can easily be estimated from the samples

$$\hat{\boldsymbol{\mu}}_i = \sum_{k=1}^{n_i} \frac{\boldsymbol{X}_{ik}}{n_i}, i = 0, 1, \quad \hat{\boldsymbol{\mu}}_a = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_0)/2, \quad \hat{\boldsymbol{\mu}}_d = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0 \tag{2.13}$$

$$\hat{\Sigma} = (\hat{\Sigma}_1 + \hat{\Sigma}_0)/2, \tag{2.14}$$

where $\hat{\Sigma}_i = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (\boldsymbol{X}_{ik} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{X}_{ik} - \hat{\boldsymbol{\mu}}_i)^T$.

$$\hat{D} = \text{diag}(\hat{\Sigma}) = \text{diag}\{(S_{1j}^2 + S_{0j}^2)/2, j = 1, \ldots, p\}, \tag{2.15}$$

where $S_{ij}^2 = \sum_{k=1}^{n_i} \frac{(X_{ikj} - \bar{X}_{ij})^2}{n_i - 1}$ is the sample variance of the j^{th} feature in class i and $\bar{X}_{ij} = \sum_{k=1}^{n_i} \frac{X_{ikj}}{n_i}$. Assuming that we have comparable sample sizes and under normality assumptions the plug-in discrimination functions can be written as

$$\hat{\delta}_F(\boldsymbol{X}, \hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\mu}}_a, \hat{\Sigma}) = \mathbf{1}\left\{\hat{\boldsymbol{\mu}}_d^T \hat{\Sigma}^{-1}(\boldsymbol{X} - \hat{\boldsymbol{\mu}}_a) > 0\right\}, \tag{2.16}$$

and

$$\hat{\delta}_{NB}(\boldsymbol{X}, \hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\mu}}_a, \hat{D}) = \mathbf{1}\left\{\hat{\boldsymbol{\mu}}_d^T \hat{D}^{-1}(\boldsymbol{X} - \hat{\boldsymbol{\mu}}_a) > 0\right\}. \tag{2.17}$$

If we have a new observation $\boldsymbol{X}$ from class $\mathcal{C}_0$, then the misclassification error rate of $\hat{\delta}_F(\boldsymbol{X}, \hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\mu}}_a, \hat{\Sigma})$ is

$$W(\hat{\delta}_F, \boldsymbol{\theta}_0) = P(\hat{\delta}_F(\boldsymbol{X}, \hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\mu}}_a, \hat{\Sigma}) = 1 | \boldsymbol{X}_{ik}, k = 1, \ldots, n_i, i = 0, 1) = \bar{\Phi}(\Psi_F) \tag{2.18}$$

where

$$\Psi_F = \frac{(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_a)^T \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_d}{\sqrt{\hat{\boldsymbol{\mu}}_d^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_d}}. \tag{2.19}$$

Similarly, the misclassification error rate of $\hat{\delta}_{NB}(\boldsymbol{X}, \hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\mu}}_a, \hat{D})$ is

$$W(\hat{\delta}_{NB}, \boldsymbol{\theta}_0) = P(\hat{\delta}_{NB}(\boldsymbol{X}, \hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\mu}}_a, \hat{D}) = 1 | \boldsymbol{X}_{ik}, k = 1, \ldots, n_i, i = 0, 1) = \bar{\Phi}(\Psi_{NB}) \tag{2.20}$$

where

$$\Psi_{NB} = \frac{(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_a)^T \hat{D}^{-1} \hat{\boldsymbol{\mu}}_d}{\sqrt{\hat{\boldsymbol{\mu}}_d^T \hat{D}^{-1} \Sigma \hat{D}^{-1} \hat{\boldsymbol{\mu}}_d}}. \tag{2.21}$$

In the next section, we always consider the misclassification error rate of observations from $\mathcal{C}_0$, since the misclassification error rate of observations from $\mathcal{C}_1$ can be easily obtained by interchanging n_0 with n_1 and $\boldsymbol{\mu}_0$ with $\boldsymbol{\mu}_1$.

2.2.2 Sample Misclassification Error Rates for Naive Bayes

Fan and Fan (2008) gave an upper bound for the sample misclassification error of the Naive Bayes given in equation (2.20) using all the p features. Similar with them we give lower and upper bounds on its misclassification error using only the important m features.

Suppose our parameter space is

$$\Gamma_m = \{(\boldsymbol{\mu}_d^{(m)}, \Sigma^{(m)}) : (\boldsymbol{\mu}_d^{(m)})^T (D^{(m)})^{-1} \boldsymbol{\mu}_d^{(m)} \geq C_m, \lambda_{\min}(\boldsymbol{\rho}^{(m)}) \geq a_0, \lambda_{\max}(\boldsymbol{\rho}^{(m)}) \leq b_0, \min_{1 \leq j \leq m} \sigma_j^2 > 0\}, \tag{2.22}$$

where C_m is the minimum signal which depends only on the dimensionality m, a_0, b_0 are positive constants and $\lambda_{\min}(\boldsymbol{\rho}^{(m)}), \lambda_{\max}(\boldsymbol{\rho}^{(m)})$ are the smallest and largest eigenvalues of the $m \times m$ correlation matrix $\boldsymbol{\rho}^{(m)}$ respectively. Let $n = n_1 + n_0$ be the total sample size. Fan and Fan (2008) stated their result as follows.

Theorem 2.3. *Suppose that* $\log p = o(n), n = o(p)$ *and* $nC_p \to \infty$. *Then:*

The classification error $W(\hat{\delta}_{NB}, \boldsymbol{\theta})$ *with* $\boldsymbol{\theta} \in \Gamma_p$ *is bounded above as*

$$W(\hat{\delta}_{NB}, \boldsymbol{\theta}) \leq \bar{\Phi} \left(\frac{[n_1 n_0/(pn)]^{\frac{1}{2}} \boldsymbol{\mu}_d^T D^{-1} \boldsymbol{\mu}_d (1 + o_P(1)) + \sqrt{p/(nn_1n_0)}(n_1 - n_0)}{2\sqrt{\lambda_{\max}(\boldsymbol{\rho})}\{1 + n_1 n_0/(pn) \boldsymbol{\mu}_d^T D^{-1} \boldsymbol{\mu}_d (1 + o_P(1))\}^{\frac{1}{2}}} \right)$$

Proof of Theorem 2.3: See Fan and Fan (2008). □

Theorem 2.4. *Suppose* $\boldsymbol{\rho}^{(m)}$ *is an* $m \times m$ *correlation matrix and* $\boldsymbol{\mu}_d^{(m)}$ *is an* $m \times 1$ *mean difference vector. Suppose also that* $m \leq s, \log m = o(n), n = o(m)$ *and* $nC_m \to \infty$.

(***a***) *Then, the classification error* $W(\hat{\delta}_{NB}, \boldsymbol{\theta}^{(m)})$ *with* $\boldsymbol{\theta}^{(m)} \in \Gamma_m$ *is bounded below and above as*

$$\bar{\Phi}(\Psi_n(\lambda_{\min}(\boldsymbol{\rho}^{(m)}))) \leq W(\hat{\delta}_{NB}, \boldsymbol{\theta}^{(m)}) \leq \bar{\Phi}(\Psi_n(\lambda_{\max}(\boldsymbol{\rho}^{(m)}))).$$

(***b***) *Assume, further, that* $\lambda_{\min}(\boldsymbol{\rho}^{(m)}) \geq 1 - \bar{\rho}$. *Then we have:*

$$\bar{\Phi}(\Psi_n(1 - \bar{\rho})) \leq W(\hat{\delta}_{NB}, \boldsymbol{\theta}^{(m)}) \leq \bar{\Phi}(\Psi_n(1 + (m-1)\rho_{\max})),$$

where

$$\Psi_n(x) = \frac{[n_1 n_0/(mn)]^{\frac{1}{2}} (\boldsymbol{\mu}_d^{(m)})^T (D^{(m)})^{-1} \boldsymbol{\mu}_d^{(m)} (1 + o_P(1)) + \sqrt{m/(nn_1n_0)}(n_1 - n_0)}{2\sqrt{x}\{1 + n_1 n_0/(mn) (\boldsymbol{\mu}_d^{(m)})^T (D^{(m)})^{-1} \boldsymbol{\mu}_d^{(m)} (1 + o_P(1))\}^{\frac{1}{2}}} \tag{2.23}$$

Proof of Theorem 2.4:

(**a**) It can be shown that $\Psi_{NB}^{(m)}$ which is the truncated version in equation (2.21) can be bounded as

$$\Psi_{NB}^{(m)} \geq \frac{(\boldsymbol{\mu}_0^{(m)} - \hat{\boldsymbol{\mu}}_a^{(m)})^T (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)}}{\sqrt{b_0 (\hat{\boldsymbol{\mu}}_d^{(m)})^T (\hat{D}^{(m)})^{-1} D^{(m)} (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)}}} \tag{2.24}$$

because using lemma 4 we note that

$$\begin{aligned}
(\hat{\boldsymbol{\mu}}_d^{(m)})^T (\hat{D}^{(m)})^{-1} \Sigma^{(m)} (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)} &= (\hat{\boldsymbol{\mu}}_d^{(m)})^T (\hat{D}^{(m)})^{-1} (D^{(m)})^{\frac{1}{2}} \boldsymbol{\rho}^{(m)} (D^{(m)})^{\frac{1}{2}} (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)} \\
&\leq (\hat{\boldsymbol{\mu}}_d^{(m)})^T (\hat{D}^{(m)})^{-1} (D^{(m)})^{\frac{1}{2}} \lambda_{\max}(\boldsymbol{\rho}^{(m)}) (D^{(m)})^{\frac{1}{2}} (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)} \\
&\leq b_0 (\hat{\boldsymbol{\mu}}_d^{(m)})^T (\hat{D}^{(m)})^{-1} D^{(m)} (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)}
\end{aligned}$$

where b_0 is from equation (2.22) and the remainder of the proof is similar with the one in Fan and Fan (2008) for theorem 2.3 where they give asymptotic results for the numerator and denominator for the right side of equation (2.24) using all the p features.

Similarly,

$$\Psi_{NB}^{(m)} \leq \frac{(\boldsymbol{\mu}_0^{(m)} - \hat{\boldsymbol{\mu}}_a^{(m)})^T (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)}}{\sqrt{b_0 (\hat{\boldsymbol{\mu}}_d^{(m)})^T (\hat{D}^{(m)})^{-1} D^{(m)} (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)}}} \tag{2.25}$$

because using lemma 4 we note that

$$\begin{aligned}
(\hat{\boldsymbol{\mu}}_d^{(m)})^T (\hat{D}^{(m)})^{-1} \Sigma^{(m)} (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)} &= (\hat{\boldsymbol{\mu}}_d^{(m)})^T (\hat{D}^{(m)})^{-1} (D^{(m)})^{\frac{1}{2}} \boldsymbol{\rho}^{(m)} (D^{(m)})^{\frac{1}{2}} (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)} \\
&\geq (\hat{\boldsymbol{\mu}}_d^{(m)})^T (\hat{D}^{(m)})^{-1} (D^{(m)})^{\frac{1}{2}} \lambda_{\min}(\boldsymbol{\rho}^{(m)}) (D^{(m)})^{\frac{1}{2}} (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)} \\
&\geq a_0 (\hat{\boldsymbol{\mu}}_d^{(m)})^T (\hat{D}^{(m)})^{-1} D^{(m)} (\hat{D}^{(m)})^{-1} \hat{\boldsymbol{\mu}}_d^{(m)}
\end{aligned}$$

the second inequality is where we have used the assumption that $\lambda_{\min}(\boldsymbol{\rho}^{(m)}) \geq a_0$ from equation (2.22). The rest will be similar to the proof given in Fan and Fan (2008) for theorem 2.3.

(**b**) Using part (**a**) we have

$$\bar{\Phi}(\Psi_n(\lambda_{\min}(\boldsymbol{\rho}^{(m)}))) \leq W(\hat{\delta}_{NB}, \boldsymbol{\theta}^{(m)}) \leq \bar{\Phi}(\Psi_n(\lambda_{\max}(\boldsymbol{\rho}^{(m)}))),$$

where Ψ_n is given in equation (2.23). Note that $\lambda_{\min}(\bar{\boldsymbol{\rho}}^{(m)}) = 1 - \bar{\rho}$ and $\lambda_{\max}(\boldsymbol{\rho}_{\max}^{(m)}) = 1 + (m-1)\rho_{\max}$. Using lemma 3 we have

$$\lambda_{\min}(\bar{\boldsymbol{\rho}}^{(m)}) \leq \lambda_{\min}(\boldsymbol{\rho}^{(m)}) \leq \lambda_{\max}(\boldsymbol{\rho}^{(m)}) \leq \lambda_{\max}(\boldsymbol{\rho}_{\max}^{(m)}).$$

Using the bounds in part (**a**) and noting that $\bar{\Phi}(\Psi_n(x))$ is an increasing function of x then we have the inequalities in part (**b**). □

For example, let $\boldsymbol{\mu}_1 = (\boldsymbol{\alpha}_{90}, \mathbf{0}_{4410})^T, \boldsymbol{\mu}_0 = (\mathbf{0}_{4500})^T, \sigma_{kk} = 1, \sigma_{kj} = \rho \in [0, 1), k \neq j, n_1 = n_0 = 30$. The following compares the upper bound on the sample error rate given in theorem 2.2 (**a**) and the theoretical error rates for Naive Bayes given in equation (2.12), which is the same as the theoretical upper bound given in theorem 2.2 (**a**), for several correlation structures using the first $m = 90$ features.

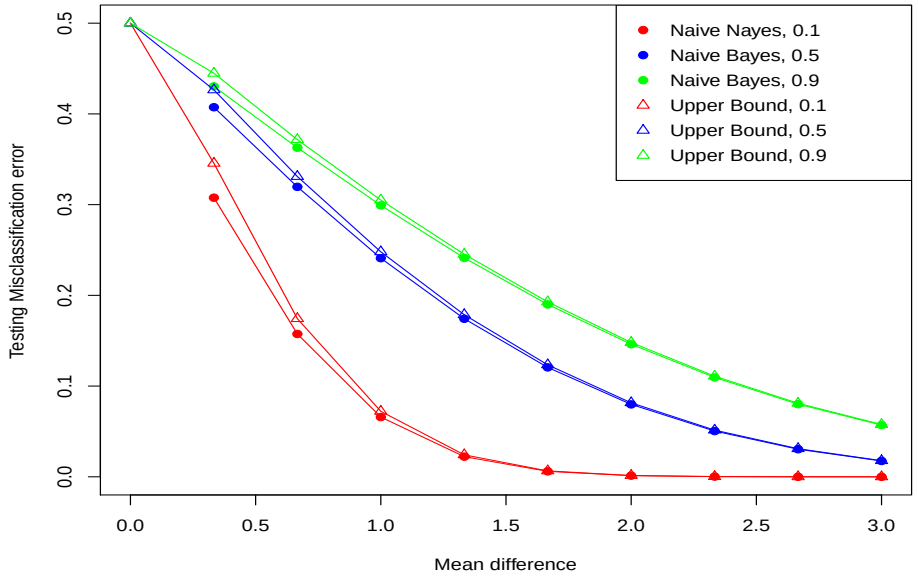


Figure 2.2: *The horizontal axis is difference of the means* (α) *and the vertical is error rates. The different lines are theoretical maximum error rate vs upper bound on the sample error rate for Naive Bayes when* $\rho = 0.1, 0.5, 0.9$ *from bottom to top respectively using the first* $m = 90$ *features.*

As we can see from the above figure, the upper bound for the sample misclassification error rates are close to the maximum theoretical error rates when we use only the subset of the features for the equicorrelation and equal mean difference case we consider.

2.3 Univariate and Multivariate t distribution

Definition 3. *Univariate t-distribution with* ν *degrees of freedom and noncentrality parameter* $\mu_Z/\sqrt{V/\nu}$ *can be defined as the distribution of the random variable* T *with*

$$T = \frac{Z + \mu_Z}{\sqrt{V/\nu}}$$

where Z is normally distributed with expected value 0 and variance 1; V has a chi-squared distribution with ν degrees of freedom; Z and V are independent. If $\mu_Z \neq 0$ it is called non-central t-distribution otherwise it is central (Student's) t-distribution (Wikipedia, 2013).

Student's t distribution can be generalized to a three parameter location-scale family, introducing a location parameter μ and a scale parameter σ, through the relation

$$X = \mu + \sigma T$$

The resulting **non-standardized Student's t-distribution** has a density denoted by, $\boldsymbol{t}_1(\mu, \sigma^2, \nu)$, has the form

$$f(x|\mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{(\pi\nu\sigma^2)^{1/2}\Gamma(\nu/2)[1 + \frac{(x-\mu)^2}{\nu\sigma^2}]^{(\nu+1)/2}}, \quad (-\infty < x < \infty)$$

where $-\infty < \mu < \infty, \sigma^2 > 0,$ and $\nu > 0$.

Here, σ does not correspond to a standard deviation: it is not the standard deviation of the scaled t distribution, which may not even exist; nor is it the standard deviation of the underlying normal distribution, which is unknown.

Generalization of the univariate student-t distribution to multivariate situations takes a number of forms. We shall concentrate on the one that is widely used in applied statistics (Cornish 1954; Dunnett and Sobel 1954; Little and Rubin, 1987; Little, 1988; Lang, Little and Taylor, 1989), which is defined as follows.

Definition 4. *(The multivariate $\boldsymbol{t}$ distribution) Let $\boldsymbol{Z} = (z_1, \ldots, z_p)^T \sim \mathcal{N}_p(\boldsymbol{\mu}_Z, \Sigma)(|\Sigma| > 0), \tau \sim \Gamma(\nu/2, \nu/2)$, let $\boldsymbol{Z}$ and τ be independent, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ be a $p-$dimensional vector , and*

$$\boldsymbol{X} = \tau^{-1/2}\boldsymbol{Z} + \boldsymbol{\mu},$$

then the (marginal) distribution of $\boldsymbol{X}$ *is called multivariate* $\boldsymbol{t}$ *distribution with* ν *degrees of freedom, denoted by*

$$\boldsymbol{t} \sim \boldsymbol{t}_p(\mu, \mu_Z, \Sigma, \nu). \tag{2.26}$$

If $\boldsymbol{\mu}_Z = \mathbf{0}$, (2.26) is called the **central** multivariate **t** distribution or simply the multivariate **t** distribution, denoted by

$$\boldsymbol{t} \sim \boldsymbol{t}_p(\mu, \Sigma, \nu); \tag{2.27}$$

otherwise (2.26) is called **noncentral** multivariate **t** distribution.

Note that if $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}_Y, \Sigma)$ and $S \sim \sqrt{\chi^2_{n_1+n_0-2}/(n_1+n_0-2)}$ then $\boldsymbol{X} = S^{-1}\boldsymbol{Y}$ follows a multivariate t-distribution. If $\boldsymbol{\mu}_Y = \mathbf{0}$, it is central otherwise it is non-central.

2.4 Feature Selection: two-sample t-test

Dimension reduction or feature selection is an effective strategy to deal with high dimensionality. With dimensionality reduced from high to low, the computational burden can be reduced drastically. Meanwhile, accurate estimation can be obtained by using some well-developed lower dimensional method.

In the previous sections we showed that Fisher is optimal when there is no estimation and Naive Bayes can perform well at the population level for broader conditions. Bickel and Levina (2004) showed that estimation accumulates noise and Fisher breaks down when using all the features. Fan and Fan (2008) also showed that using all the features for Naive Bayes increases the misclassification error rate and suggested using the subset of features. To get these features we appeal to the independence feature selection methods.

A popular method for independence feature selection is the two-sample t-test (Tibshirani et al., 2002, Fan and Fan, 2008), which is a specific case of marginal screening in Fan and Lv (2008). Other componentwise tests such as the rank sum test are also popular.

We divide our data set into three groups: training, testing and validation data sets. We rank the absolute values of the t-statistics from large to small and then we choose the first m features. We choose the optimal m which minimizes the upper bound on the misclassification error rate based on the validation data set.

Suppose that the noise vectors $\boldsymbol{\epsilon}_{ik}$ are i.i.d within class $\mathcal{C}_i$ with mean $\mathbf{0}$ and covariance matrix Σ_i.

Fan and Fan (2008) gave a condition under which the two-sample t-test picks up all the important s features with probability 1. They stated it as follows:

Condition 1.

(a) Assume that the vector $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is sparse and without loss of generality, only the first s entries are nonzero.

(b) Suppose that ϵ_{ikj} and $\epsilon_{ikj}^2 - 1$ satisfy the Cramer's condition, that is, there exist constants ν_1, ν_2, M_1 and M_2, such that $E|\epsilon_{ikj}|^m \leq m! M_1^{m-2}\nu_1/2$ and $E|\epsilon_{ikj}^2 - \sigma_{ij}^2| \leq m! M_2^{m-2}\nu_2/2$ for all $m = 1, 2, \ldots$.

(c) Assume that the diagonal elements of both Σ_1 and Σ_0 are bounded away from 0.

For unequal sample sizes, unequal variance, the absolute value of the two-sample t-statistic for feature j is defined as

$$T_j = \frac{|\bar{X}_{1j} - \bar{X}_{0j}|}{\sqrt{S_{1j}^2/n_1 + S_{0j}^2/n_0}}, \quad j = 1, \ldots, p. \tag{2.28}$$

Theorem 2.5. *Let s be a sequence such that $\log(p-s) = o(n^\gamma)$ and $\log s = o(n^{1/2-\gamma}\beta_n)$ for some $\beta_n \to \infty$ and $0 < \gamma < 1/3$. Suppose that $\min_{1\leq j\leq s} \frac{|\mu_{d,j}|}{\sqrt{\sigma_{1j}^2 + \sigma_{0j}^2}} = n^{-\gamma}\beta_n$ where $\mu_{d,j}$ is the j^{th} feature mean difference. Then under Condition 1, for $x \sim cn^{\gamma/2}$ with c some positive*

constant, we have

$$P\left(\min_{j\leq s} T_j \geq x \ \textit{and} \ \max_{j>s} T_j < x\right) \to 1.$$

Proof of Theorem 2.5: See Fan and Fan (2008).

Note that asymptotically the two-sample t-test can pick up all the important features. However we are interested in the probability of selecting all the important features (i.e the probability of getting all the s features in the first s ordered t-statistics) in the short run.

Consider the following p t-statistics, $T_1, T_2, \ldots, T_p$, defined in equation (2.28). We call the s t-statistics corresponding to the s non zero mean differences, the non-centrals (which are the important features in our case) and the remaining $p-s$ as centrals. Suppose that $T_{(1)}, T_{(2)}, \ldots, T_{(p)}$ are the reverse order t-statistics. Let m' =the largest rank assigned to the non-centrals. For example, if $m' = s$, it means the first s highest t-statistics contain the s non-centrals, if $m' = s+1$, it means the first $s+1$ highest t-statistics contain the s non-centrals and etc.

Suppose, further that, $T^*_{(1)} \geq T^*_{(2)} \geq \cdots \geq T^*_{(s)}$ are the reverse order t-statistics for the non-centrals and $T_{(1)} \geq T_{(2)} \geq \cdots \geq T_{(p-s)}$ are the reverse order t-statistics for the centrals.

We now give the formulas for probability of getting all the important s features in the first $m' = s, m' = s+1, \ldots, m' = p-1$ ordered t-statistics:

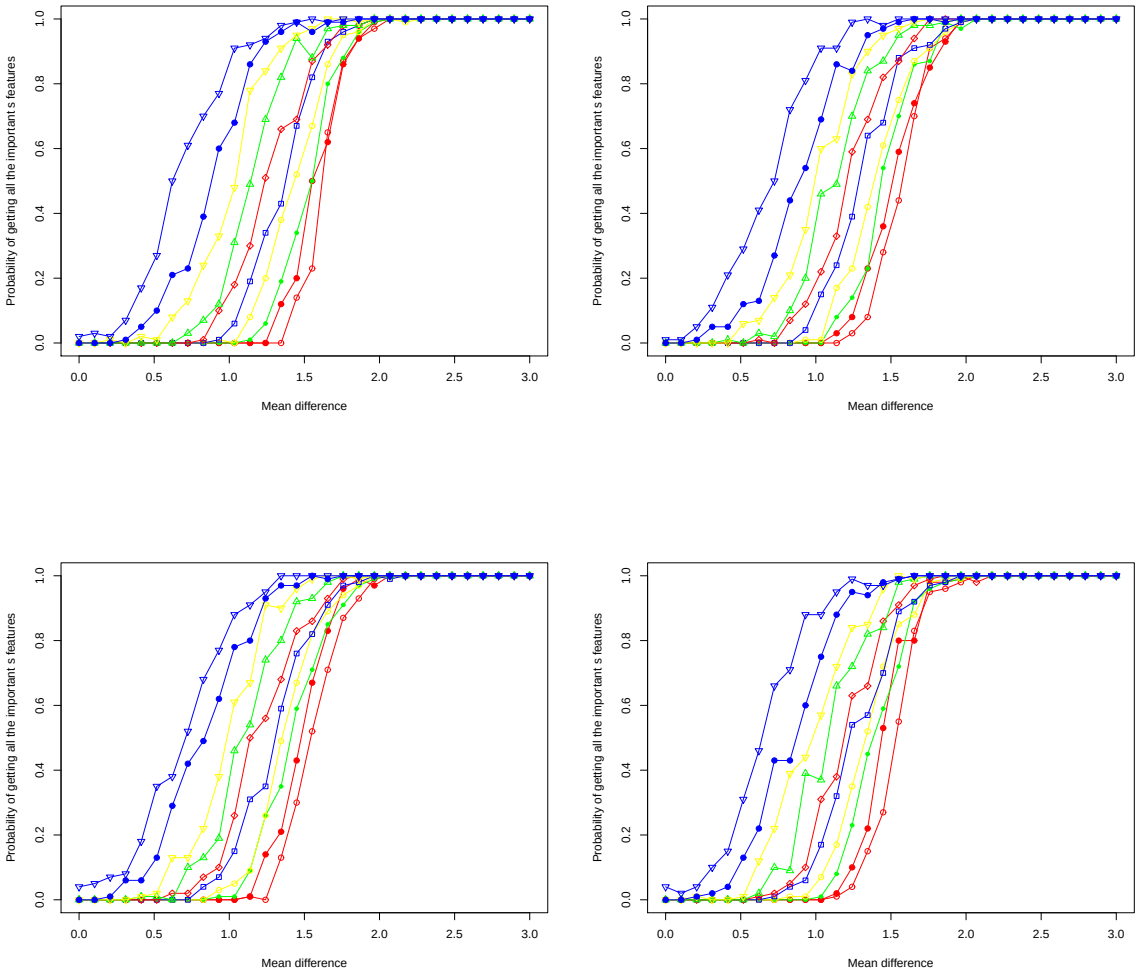
$$\begin{aligned} P(m' = s) &= P(T^*_{(s)} > T_{(1)}) \\ P(m' = s+1) &= P(T^*_{(s)} > T_{(2)}) \\ &\vdots \\ P(m' = p-1) &= P(T^*_{(s)} > T_{(p-s)}) \end{aligned}$$

Generally, for $i = s, \ldots, p-1$

$$P(m' = i) = P(T^*_{(s)} > T_{(i-s+1)}).$$

We use simulation to calculate the above probabilities. The simulation results are based on generating the p dimensional multivariate t-distribution of which dimension of $s = 90$ multivariate non-central t and dimension of $p - s = 4500 - 90 = 4410$ multivariate central t with $n_1 = n_0 = 30$ with the covariance matrix for the underlying distribution Σ assumed to be the $p \times p$ equicorrelation matrix with off diagonals $\rho \in [0, 1)$. We use the definition of multivariate t distributions whose marginal are also t. The definition is given in the 1994 PhD Dissertation (by Chuanhai Liu). The number of simulation is 100 for each case.

The following are the graphs of probability of getting all the important s features against the standardized mean differences of the underlying population.



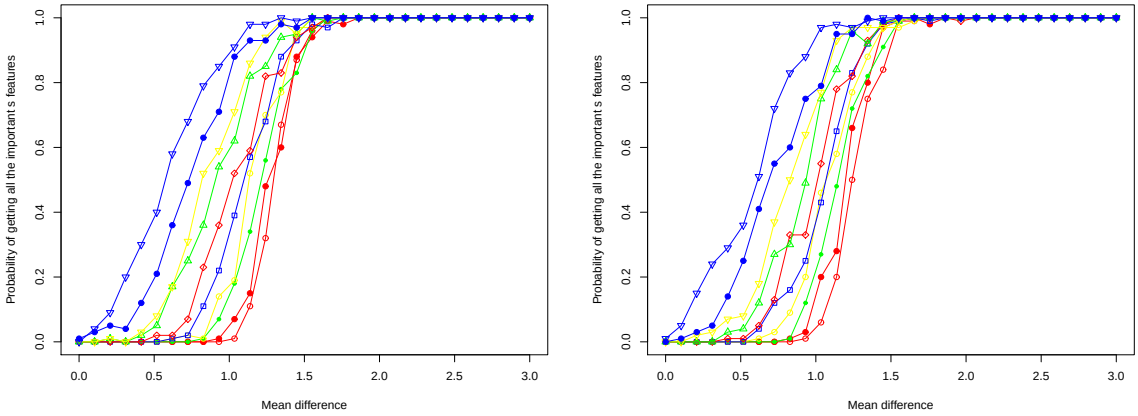


Figure 2.3: *Horizontal axis is the mean difference (α) and vertical is probability of getting all the important s features in the first $s, s+1, s+2, s+3, 3s/2, 2s$ t-statistics respectively. The results are for $\rho = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ from bottom to top respectively.*

As we can see from the first and last figures, when $\rho = 0.6$ and $\alpha = 1$ standard deviation difference, the probability of getting all the important features in the first s t-statistics is below 45% and the probability of getting all the important features in the first $2s$ t-statistics is around 80% respectively. We can see also that at around $\alpha = 1.5$, we have 100% chance of getting all the s features in the first $2s$ t-statistics with any ρ but we don't have 100% chance of getting all the s features in the first s t-statistic unless the correlation is very high.

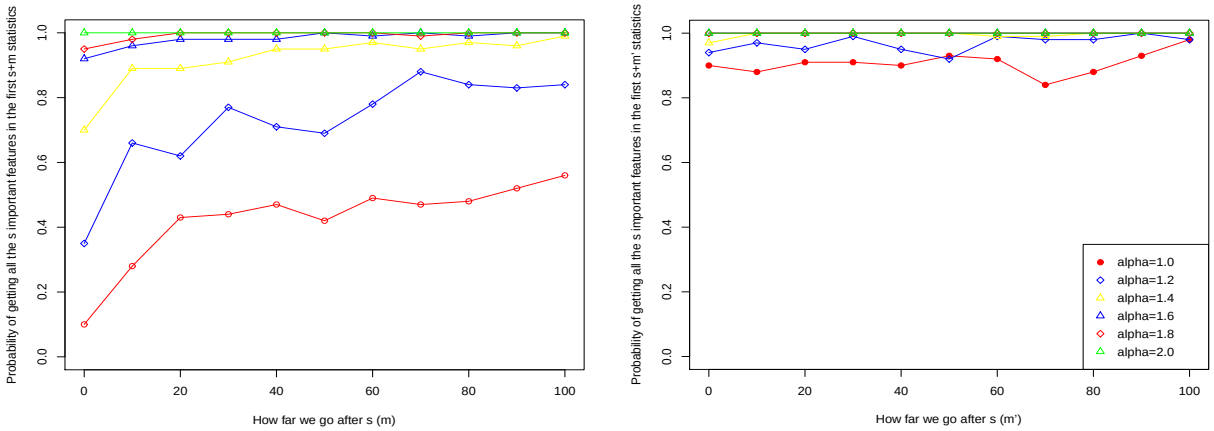


Figure 2.4: *Horizontal axis is the number we go after s (call it, m') and the vertical axis is probability of getting all the important s features in the first $s + m'$ t-statistics for $\rho = 0.5, 0.9$ respectively. The different lines are for $\alpha = 1, 1.2, 1.4, 1.6, 1.8, 2$ from bottom to top respectively.*

As we can see from the above figure, when $\rho = 0.9$, we have around 90% chance of getting all the important features for any α bigger or equal to 1. When we have large α and higher correlation, we have 100% chance of getting all the important s features in the first s t-statistics.

The following results are under the assumption that $T_1, \ldots, T_s$ are noncentral multivariate t-distribution (NCT) and $T_{s+1}, \ldots, T_p$ are central multivariate t-distribution (CT). We also assume that the centrals are independent of the noncentrals.

Let $T_1, \ldots, T_s$ have joint distribution function given by $F_{NCT}(t_1, \ldots, t_s) := P(T_1 \leq t_1, \ldots, T_s \leq t_s)$ which is multivariate NCT and $T_{s+1}, \ldots, T_p$ have joint distribution function given by $F_{CT}(t_{s+1}, \ldots, t_p) := P(T_{s+1} \leq t_{s+1}, \ldots, T_p \leq t_p)$ which is multivariate CT.

$$P(m' = s) = P(\min_{j \leq s} |T_j| > \max_{j > s} |T_j|) = P(\min_{j \leq s} |T_j| - \max_{j > s} |T_j| > 0) \tag{2.29}$$

We have the following densities when $x > 0$:

$$f_{\min_{j\leq s}|T_j|}(x) = -\frac{d}{dx}P(|T_1| > x, \ldots, |T_s| > x) \tag{2.30}$$

$$f_{\max_{j>s}|T_j|}(x) = \frac{d}{dx}P(|T_{s+1}| < x, \ldots, |T_p| < x) \tag{2.31}$$

Then, the probability of getting all the important features in the first s t-statistics is given by

$$\int_0^\infty \int_0^x f_{\min_{j\leq s}|T_j|}(x) f_{\max_{j>s}|T_j|}(y) dy dx \tag{2.32}$$

For unequal sample sizes, equal variance the two-sample t-test takes the following form

$$T_j = \frac{\bar{X}_{1j} - \bar{X}_{0j}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}, \text{ for } j = 1, 2, \ldots, p,$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)S_{1j}^2 + (n_0 - 1)S_{0j}^2}{n_1 + n_0 - 2}}.$$

Note that under normality assumption on the underlying distribution we have

$$T_j \sim \frac{\mathcal{N}(0, 1) + (\boldsymbol{\mu}_d)_j}{\sqrt{\chi^2_{n_1+n_0-2}/(n_1 + n_0 - 2)}}$$

which is a t-distribution with n_1+n_0−2 degrees of freedom and $(\boldsymbol{\mu}_d)_j/\sqrt{\chi^2_{n_1+n_0-2}/(n_1 + n_0 - 2)}$ as noncentrality parameter.

But using the order statistic and noting that

$$F_{|X|}(x) = P(|X| \leq x) = P(-x \leq X \leq x) = F_X(x) - F_X(-x) \tag{2.33}$$

for any random variable X we have the following density of $|X|$

$$f_{|X|}(x) = f_X(x) + f_X(-x) \tag{2.34}$$

For independent CT and independent NCT we have the following densities.

$$f_{\min\limits_{j \leq s} |T_j|}(x) = s f_{|T_{NC}|}(x)[1 - F_{|T_{NC}|}(x)]^{s-1}, \quad 0 \leq x < \infty \tag{2.35}$$

and

$$f_{\max\limits_{j > s} |T_j|}(x) = (p-s) f_{|T_C|}(x)[F_{|T_C|}(x)]^{p-s-1}, \quad 0 \leq x < \infty \tag{2.36}$$

based on the assumption that $T_1, \ldots, T_s$ are iid t-noncentral (T_{NC}) with $n_1 + n_0 - 2$ degrees of freedom and with

$$ncp = \sqrt{(n_1 + n_0 - 2)/\chi^2_{n_1+n_0-2}} \ .\alpha$$

noncentralilty parameter and $T_{s+1}, \ldots, T_p$ are iid t-central (T_C) with $n_1 + n_0 - 2$ degrees of freedom. Without loss of generality we assume the standard deviations for each feature is 1. In addition, we assume that the centrals are independent of the noncentrals. Assuming that $\alpha := (\mu_d)_1 = \cdots = (\mu_d)_s \neq 0$ and $(\mu_d)_{s+1} = \cdots = (\mu_d)_p = 0$ where $\boldsymbol{\mu}_d = ((\mu_d)_1, \ldots, (\mu_d)_p)^T$.

Equation (2.29) becomes

$$\int_0^\infty \int_0^x f_{\min\limits_{j\leq s}|T_j|}(x) f_{\max\limits_{j>s}|T_j|}(y) dy dx \tag{2.37}$$

Motivated by this, we give a formula for the probability of getting all the important features in the first $s, s+1, s+2, \ldots, p$ t-statistics. Let m' be how far we go to get all the important features after the first s t-Statistics. Here m' takes the values $m' = 0, 1, \ldots, p-s$.

Let M be the random variable that denotes the number of the absolute value of the central t which are bigger than the minimum of the absolute value of the noncentrals, then M follows a binomial distribution with $p-s$ trials and probability of success $Pr = P(|T| - \min\limits_{j\leq s}|T_j| > 0 \big| T_1, \ldots, T_s)$ parameters where T is any central t-distribution with $n_1 + n_0 - 2$ degrees of freedom. Therefore,

$$P(M = m' | p-s, Pr) = \binom{p-s}{m'} (Pr)^{m'} (1-Pr)^{p-s-m'}, \quad m' = 0, 1, \ldots, p-s \tag{2.38}$$

is the probability of getting all the important s features in the first $m' + s$ t-statistics.

The probability Pr can be calculated as follows:

$$Pr = P(|T| - \min_{j\leq s}|T_j| > 0) = \int_0^\infty \int_0^x f_{|T|}(x) f_{\min\limits_{j\leq s}|T_j|}(y) dy dx \tag{2.39}$$

Let $M^{'}$ be the number of the absolute value of the noncentral t which are bigger than the maximum of the absolute values of the centrals), $M^{'}$ follows a binomial distribution with s number of trials and probability of success $Pr^{'}$. Then

$$P(M^{'} = m^{'} | s, Pr^{'}) = \binom{s}{m^{'}} (Pr^{'})^{m^{'}} (1-Pr^{'})^{s-m^{'}}, \quad m^{'} = 0, 1, \ldots, s \tag{2.40}$$

where $Pr^{'} = P(|T_{NC}| - \max_{j>s} |T_j| > 0 \big| T_{s+1}, \ldots, T_p)$. T_{NC} is any noncentral t-statistics with $n_1 + n_0 - 2$ degrees of freedom and ncp noncentrality parameter given above. The probability $Pr^{'}$ can be calculated as follows:

$$Pr^{'} = P(|T_{NC}| - \max_{j>s} |T_j| > 0) = \int_0^{\infty} \int_0^{x} f_{|T_{NC}|}(x) f_{\max_{j>s} |T_j|}(y) dy dx \tag{2.41}$$

The probability that exactly half of the important features appear in the first s (assume it is even) T-statistics is given by

$$P(M^{'} = \frac{s}{2} | s, Pr^{'}) = \binom{s}{\frac{s}{2}} (Pr^{'} - (Pr^{'})^2)^{\frac{s}{2}} \tag{2.42}$$

2.5 Simulation Results

Model 1: Equal sparse mean difference vector and equicorrelation matrix for balanced group

We use simulation to compare the performance of Fisher and Naive Bayes. The simulation is done as follows: we generate training samples $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ and $\boldsymbol{X}_{01}, \ldots, \boldsymbol{X}_{0n_0}$ from multivariate normal distribution with $\boldsymbol{\mu}_1 = (\boldsymbol{\alpha}_s, \boldsymbol{0}_{p-s})^T$ and $\boldsymbol{\mu}_0 = (\boldsymbol{0}_p)^T$ with training sample size $n_1 = n_0 = 30$ and $p = 4500, s = 90$. We then construct the discriminant functions assuming Σ equicorrelation matrix with off diagonals $\rho \in [0, 1)$ and we then calculated the testing errors. The testing data sets are generated from multivariate normal distribution from each class with the above means and covariance matrix. The sample size for testing data is $n_1 = n_0 = 50$. We repeat the experiment 100 times. We report the average testing errors over the 100 simulations.

The following figures compare the misclassification error rates of Naive Bayes and Fisher when $\rho = 0.5$ and $\rho = 0.9$ respectively using the first 10, 30, 45 selected features selected based on the two-sample t-test.

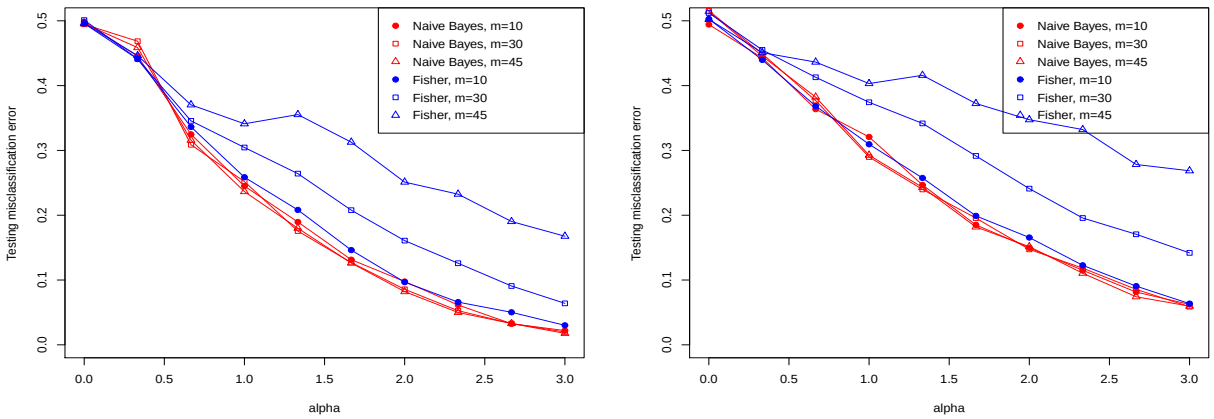


Figure 2.5: *Horizontal axis is the mean difference (α). The vertical axis is testing misclassification error rate using the first* $10, 30, 45$ *features for Naive Bayes vs Fisher. The first and second figures are when* $\rho = 0.5, 0.9$ *respectively.*

The above figures show us how Naive Bayes dominates Fisher on head to head comparison using the same number of features. This result is consistent with our theoretical result in theorem 2.1. Theorem 2.1 shows that under equal mean difference and equicorrelation matrix, the Naive Bayes and Fisher are both optimal. But because of the noise accumulation in estimating the covariance matrix (Bickel and Levina(2004), Fan and Fan (2008)), we expect Naive Bayes to do better than Fisher at the sample level. When the number of feature increases, Fisher will have an estimation error for the covariance matrix.

Model 2: Equal sparse mean difference vector and equicorrelation matrix for unbalanced group

We do the simulations as above except we take $n_1 = 30$ and $n_0 = 60$ for the training data with $\rho = 0.5$.

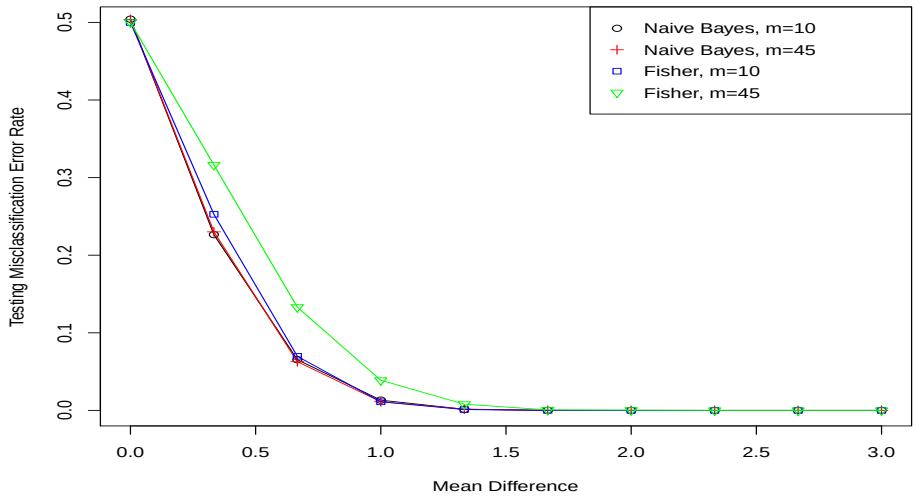


Figure 2.6: *Horizontal axis is the mean difference* (α) *and the vertical axis is testing misclassification error rate.*

As we can see from the above figure Naive Bayes still dominates Fisher for unbalanced group on head to head comparison using the same number of features. Naive Bayes is not affected much by the number of features we use but Fisher does. The reason for this is that the error rate increases for estimating the inverse of the covariance matrix for Fisher when the number of features increases.

Model 3: Equal sparse mean difference vector and random correlation matrix

We do the same simulation as above except we generate a positive definite $p \times p$ covariance matrix Σ whose eigenvalues are randomly generated in the interval $[0.5, 45.5]$ and we report the average testing errors over 100 simulations for Naive Bayes and Fisher in the following figures. Note that $\lambda_{\max}(\Sigma) = 45.49829$, and $\lambda_{\max}(\boldsymbol{\rho}) = 4.989593$ using the first 10 and 45 features respectively.

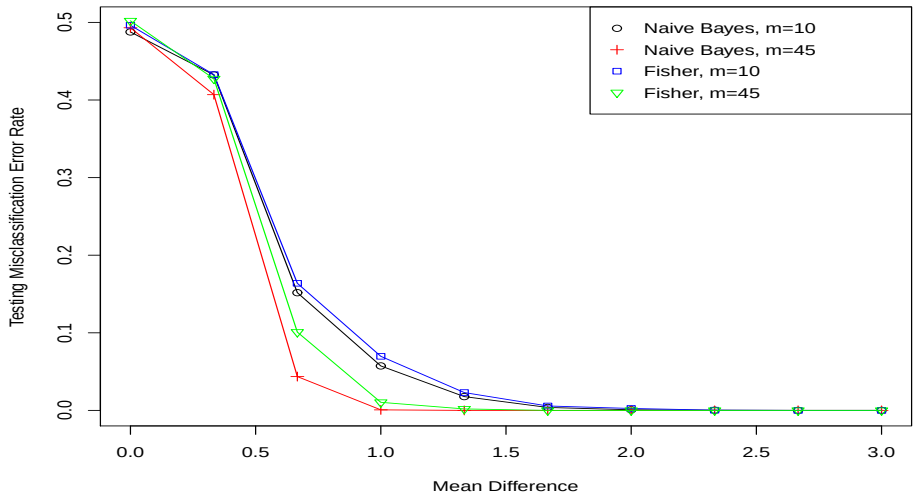


Figure 2.7: *Horizontal axis is the mean difference (α). The vertical axis is testing misclassification error rate using the first 10 and 45 features for Naive Bayes vs Fisher.*

The above figures show that Naive Bayes still dominates Fisher under random correlation on head to head comparison of using the same number of features. We compare them using 10 and 45 features. We know that for large number of features, larger than $(n_1+n_0-2=58)$ number of features, Fisher breaks down as the sample covariance matrix is singular.

Model 4: Equal sparse mean difference vector and equicorrelation matrix for Naive Bayes

We generate training samples $\boldsymbol{X}_{11},\ldots,\boldsymbol{X}_{1n_1}$ and $\boldsymbol{X}_{01},\ldots,\boldsymbol{X}_{0n_0}$ from multivariate normal distribution with $\boldsymbol{\mu}_1 = (\boldsymbol{\alpha}_s, \mathbf{0}_{p-s})^T$ and $\boldsymbol{\mu}_0 = (\mathbf{0}_p)^T$ with training sample sizes $n_1 = n_0 = 30$ and $p = 4500, s = 90$. We then construct the Naive Bayes discriminant assuming Σ equicorrrelation matrix with off diagonals $\rho \in [0,1)$ and we then calculated the testing errors. The testing data sets are generated from multivariate normal distribution from each class with the above means and covariance matrix with sample sizes $n_1 = n_0 = 50$. We repeat the experiment 100 times. We report the average testing errors over the 100 simulations. The following are how Naive Bayes performs for $\rho = 0.5$.

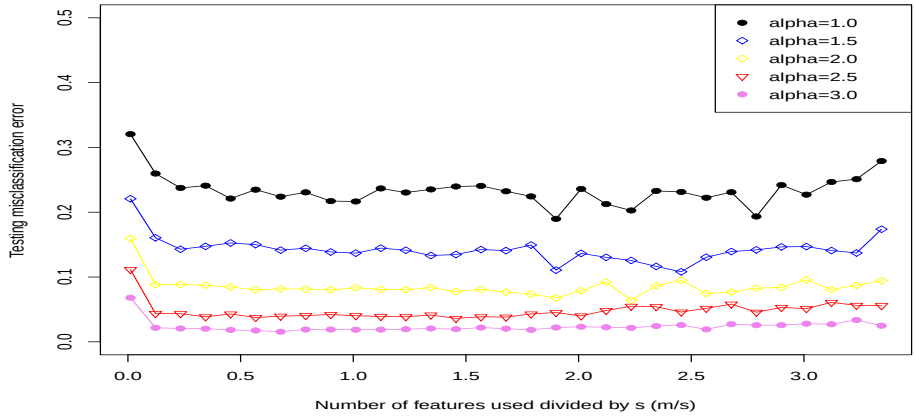


Figure 2.8: *Horizontal axis is the number of features used divided by* $s = 90$ *and the vertical axis is testing misclassification error rate. The different lines (starting from up) are for* $\alpha = 1.0, 1.5, 2.0, 2.5, 3.0$ *respectively.*

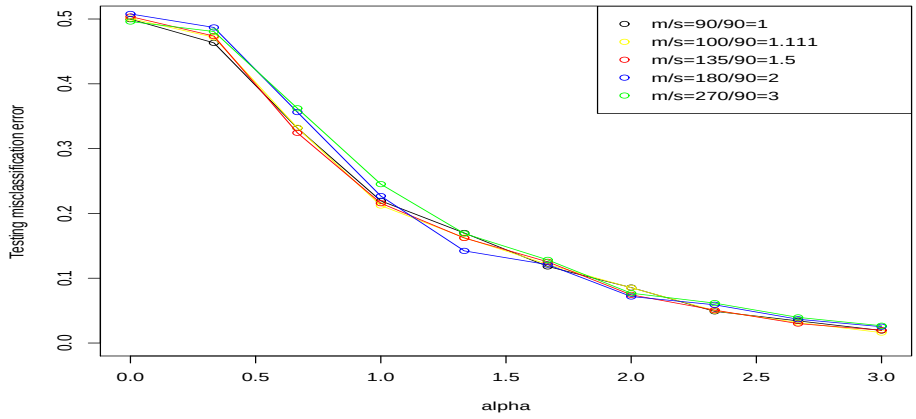


Figure 2.9: *Horizontal axis is the mean difference* (α) *and the vertical axis is testing misclassification error rate using the first* $90, 100, 135, 180$ *and* 270 *features.*

As we can see from the above figures, the Naive Bayes is not affected much by the number of features used. For example, for $\alpha = 1$ standard deviation difference the testing

error rate is a little above 20% using even below 45 features. As far as the signal is not too small, even with high correlation ($\rho = 0.5$) we get a decent classification.

2.6 Real Data Analysis: Leukemia Data

Leukemia data from high-density Affymetrix oligonucleotide arrays are available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. There are 7129 genes and 72 samples coming from two classes: 47 in class ALL (acute lymphocytic leukemia) and 25 in class AML (acute mylogenous leukemia). Among these 72 samples, 38 (27 in class ALL and 11 in class AML) are set to be training samples and 34 (20 in class ALL and 14 in class AML) are set as test samples.

Before analysis, we standardize each sample to zero mean and unit variance as follows: we subtract the grand mean from both classes and divide each class by their class standard deviation.

But in our analysis we used training sample sizes of $n_1 = 24$ from class ALL and $n_0 = 13$ from class AML. The validation sample sizes are $n_1 = 23$ from class ALL and $n_0 = 12$ from class AML. We then compare Naive Bayes and Fisher on head to had comparison using the same number of genes. As we can see from the figure, Naive Bayes dominates Fisher. Fisher breaks down using more than 35 genes.

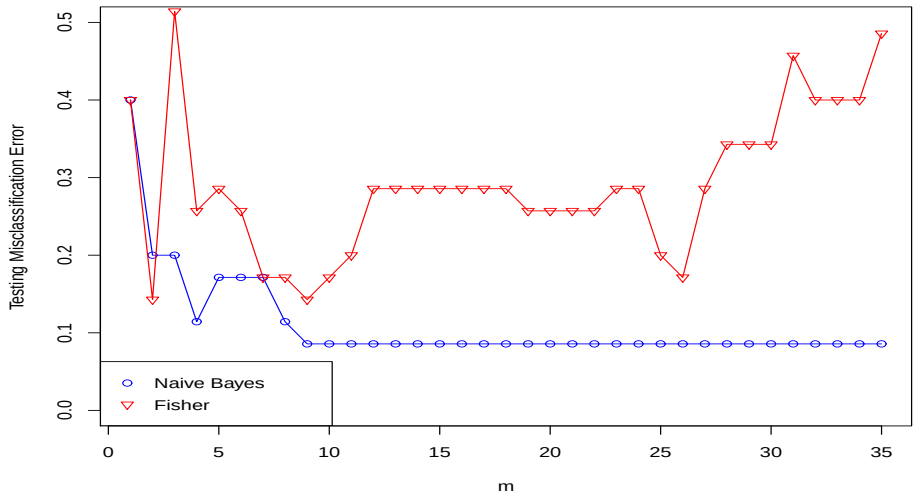


Figure 2.10: *Horizontal axis is the number of genes used and the vertical axis is testing misclassification error rate for Fisher vs Naive Bayes.*

The following figure shows how Naive Bayes performs using the first 100 genes. The minimum error is 0.05714286 which is 2 misclassified genes out of the 35 genes in the validation sample. The optimal number of genes selected is 43 using the validation data.

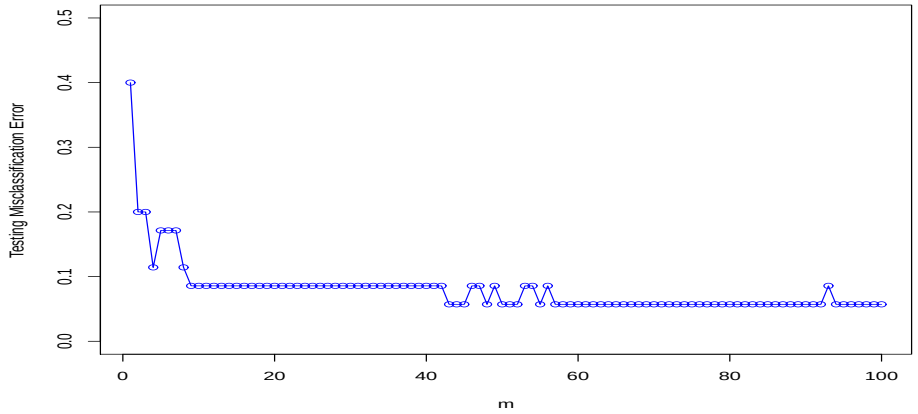


Figure 2.11: *Horizontal axis is the number of genes used and the vertical axis is testing misclassification error rate for Naive Bayes.*

2.7 Conclusion

In this chapter we considered a binary classification problem for sparse signals. We show that Naive Bayes is viable alternative to Fisher and good for experimental design. We give the conditions under which Naive Bayes performs as good as Fisher for the population model. Because of the accumulation of estimation error for Fisher (see Bickel and Levina (2004)), under these conditions, Naive Bayes performs better than Fisher for the sample model and we see this on our simulation results based on fair comparison of the two methods. The simulation results and the real data analysis support our theory. We give tight equal lower and upper bounds on the misclassification error rates of the two methods for the population model. These lower and upper bounds are based on the equicorrelation matrices formed from the mean of the correlation coefficients and the maximum of the absolute values of the correlation coefficients respectively. We show that our bounds are equal for equicorrelation covariance matrices and in this sense Naive Bayes is a minimax estimator for Fisher. These bounds make the Naive Bayes method practical method to use for experimental design.

Fan and et al. (2012) considered the class of linear discriminant functions of the form $\delta_{\boldsymbol{w}}(\boldsymbol{X}) = \mathbf{1}\{\boldsymbol{w}^T(\boldsymbol{X} - \boldsymbol{\mu}_a) > 0\}$ and their mission was to find the good data projection $\boldsymbol{w}$. In this chapter, we considered the class of linear discriminant functions of the form $\delta(\boldsymbol{X}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, M) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T M^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_a) > 0\right\}$ and our aim is to find the matrix M which gives better result for the sample model. We show that, under certain conditions given in lemma 2, these functions lead to better results. Specifically, we give the conditions under which Naive Bayes is optimal. But in our work we considered smaller sample sizes ($n_1 = n_0 = 30$) for training, ($n_1 = n_0 = 50$) for testing and large number of features ($p = 4500$) with $s = 90$. Fan and et al.(2012) considered large sample sizes ($n_1 = n_0 = 300$) for both training and testing and smaller dimension ($p = 1000$) with $s = 10$. They also compare Fisher and Naive Bayes on different number of features. We compare Fisher and Naive Bayes using the same number of features selected.

Chapter 3

Generalized Feature Selection

3.1 Introduction

However the gene expression data are usually ultrahigh dimensional such that sample size n is far smaller than the data dimension p which can make some classifiers not applicable. As we know high dimension can easily cause overflow in the calculation of inverse matrices that is required by some classifier. Typically, the calculation working load can be increased dramatically by even adding one more gene.

Besides, it is well known that only few genes carry the useful information which can determine a specific genetic trait, such as susceptibility to cancer while most of genes carry nothing useful but the noises. Taking all the genes instead of the most informative ones in to account in the process of classification can't provide a better accuracy but result in the widely inefficiency. Usually, a smaller set of genes are selected based the amount of the information in terms of the group separation to be considered as the most important genes in the process of classification. Basically, there are two ways to reduce the dimension of data:

• Select a subset of the original variables (genes) based on the power of class determination,

• Create new variables by combining the information of all the variables (genes) without loss much information from the original variables.

Many statisticians prefer that firstly a smaller set of variables are selected by following a certain variable screening method and then some optimal linear combinations of the selected variables are finally created to proceed the classification while some directly perform the classification after the variable screening.

Dudoit et al (2002) performed gene screening based on the ratio of between-group and within-group sums of squares. Many statisticians (Fan and Fan, 2008; Nguyen and Rocke, 2002; Ding and Gentleman, 2005) applied two-sample t-statistic which measures the distance between two populations and can be used as the criterion to preliminarily select the most important genes while other people (Liao et al , 2007) picked up the variables based on Wilcoxon-Mann-Whitney (WMW) statistic which is also good measurement in terms of group separation. Usually the variable screening method using WMW statistic is only slightly less efficient than the one using t-statistic when the underlying populations are normal, and it can be mildly or wildly more efficient than its competitors when the underlying populations are not normal.

3.2 The Generalized Feature Selection

Fan and Fan (2008) gave a condition under which the two-sample t-test pick up all the important s features with probability 1. Asheber Abebe and Shuxin Yin (PhD dissertation, 2010) gave a condition under which the Wilcoxon-Mann Whitney test can pick up all the important features with probability 1. Two-sample t-test heavily depends on (approximately) normal distribution. So we propose a new generalized feature selection method. We give a generalized condition under which any two-sample componentwise test T_j defined below can pick up all the important features with probability 1. Our T_j for feature j is defined as follows:

$$T_j = \frac{\sum_{k=1}^{n_1} w_{1kj} - \sum_{k=1}^{n_0} w_{0kj}}{SE(\sum_{k=1}^{n_1} w_{1kj} - \sum_{k=1}^{n_0} w_{0kj})} \tag{3.1}$$

where $w_{ikj}, i = 0, 1$, is the statistic for feature j in class i and assume that the standard error for any statistic T satisfies $SE(T) \xrightarrow{P} SD(T)$ and we assume that for some interval on $x > 0$ we have

$$P(|T_j - \eta_j| \geq x) = 2(1 - \Phi(x))(1 + f(x, n)), \tag{3.2}$$

where $f(x, n) = f_1(x, n) + f_2(-x, n) = o(x)$,

and we define

$$\eta_j = \frac{E(\sum_{k=1}^{n_1} w_{1kj}) - E(\sum_{k=1}^{n_0} w_{0kj})}{SE(\sum_{k=1}^{n_1} w_{1kj} - \sum_{k=1}^{n_0} w_{0kj})} \tag{3.3}$$

Note:

1. If we take $w_{1kj} = X_{1kj}/n_1$ and $w_{0kj} = X_{0kj}/n_0$ we get

$$T_j = \frac{\bar{X}_{1j} - \bar{X}_{0j}}{\sqrt{S_{1j}^2/n_1 + S_{0j}^2/n_0}}$$

which is the two sample t-test defined in Fan and Fan (2008).

2. If we take $w_{1kj} = r_{1kj}$ and $w_{0kj} = k$ for $1 \leq k \leq n_1$, or 0 otherwise. Assuming $n_1 \leq n_0$. Here r_{1kj} is the rank of X_{1kj} in the combined ranking of feature j in the two classes then we get

$$T_j = U_j$$

where $U_j = \frac{\sum_{k_1=1}^{n_1} \sum_{k_0=1}^{n_0} \phi(X_{1k_1j}, X_{0k_0j})}{SD(\sum_{k_1=1}^{n_1} \sum_{k_0=1}^{n_0} \phi(X_{1k_1j}, X_{0k_0j}))}$ is the Wilcoxon-Mann Whitney statistic where $\phi(x, y) = 1$ if $x - y < 0$, 0 otherwise.

3. If X_{ikj} is a binary data and if we take $w_{1kj} = X_{1kj}/n_1$ and $w_{0kj} = X_{0kj}/n_0$ in 1 then we get the following two-sample proportion test statistic

$$T_j = \frac{p_{1j} - p_{0j}}{\sqrt{p_{1j}(1 - p_{1j})/n_1 + p_{0j}(1 - p_{0j})/n_0}}$$

where $p_{1j} = \bar{X}_{1j}$ and $p_{0j} = \bar{X}_{0j}$.

The most important features will be the ones with large value of $|T_j|$ for $j = 1, \ldots, p$.

We state the theorem and give the proof for our generalized theorem as follows:

Theorem 3.1. *Assume that the vector* $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ *is sparse and without loss of generality only first* s *entries are nonzero. Let* s *be a sequence such that* $\log(p-s) = o(n^\gamma)$ *and* $\log\ s = o(n^\gamma)$ *for some* $0 < \gamma < 1/3$. *Suppose* $\min_{1\le j\le s} |\eta_j| = n^{-\gamma} C_n$ *such that* $C_n/n^{\frac{3\gamma}{2}} \to c^*$. *For* $t \sim cn^{\frac{\gamma}{2}}$ *with some constant* $0 < c < c^*/2$ *we have*

$$P(\min_{j\le s} |T_j| \ge t,\ \textit{and}\ \max_{j>s} |T_j| < t) \to 1.$$

Proof of Theorem 3.1: Without loss of generality we assume $n_1 \le n_0$.

Note that

$$\begin{aligned} P(\min_{j\le s} |T_j| \ge t,\ \text{and}\ \max_{j>s} |T_j| < t) &= 1 - P(\min_{j\le s} |T_j| \le t,\ \text{or}\ \max_{j>s} |T_j| > t) \\ &\ge 1 - P(\min_{j\le s} |T_j| \le t) - P(\max_{j>s} |T_j| > t). \end{aligned}$$

To prove that $P(\min_{j\le s} |T_j| \ge t,\ \text{and}\ \max_{j>s} |T_j| < t) \to 1$, it is enough to prove that $P(\min_{j\le s} |T_j| \le t) \to 0$ and $P(\max_{j>s} |T_j| > t) \to 0$.

We divide the proof in two parts.

(a) Let us first look at the probability $P(\max_{j>s} |T_j| > t)$. Using Boole's inequality,

$$P(\max_{j>s} |T_j| > t) \le \sum_{j=s+1}^{p} P(|T_j| > t)$$

Since for $j > s,\ \ \eta_j = 0$, so by our assumption (3.2)

$$P(|T_j| \ge t) = 2(1 - \Phi(t))(1 + f(t,n)). \tag{3.4}$$

Since for the normal distribution, it is easy to show the following tail probability inequality

$$1 - \Phi(t) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{t} e^{\frac{-t^2}{2}},$$

equation (3.4) becomes

$$\begin{aligned} P(|T_j| \geq t) &= 2(1 - \Phi(t))(1 + f(t, n)) \\ &\leq \frac{2}{\sqrt{2\pi}} \frac{1}{t} e^{\frac{-t^2}{2}} (1 + f(t, n)). \end{aligned} \tag{3.5}$$

Then

$$\sum_{j=s+1}^{p} P(|T_j| > t) \leq (p - s) \frac{2}{\sqrt{2\pi}} \frac{1}{t} e^{\frac{-t^2}{2}} (1 + f(t, n)).$$

If we let $t \sim cn^{\frac{\gamma}{2}}$, then we have

$$(p - s) \frac{2}{\sqrt{2\pi}} \frac{1}{t} e^{\frac{-t^2}{2}} (1 + f(t, n)) \to 0$$

since $\log(p - s) = o(n^{\gamma})$ with $0 < \gamma < 1/3$. Thus, we have

$$P(\max_{j>s} |T_j| > t) \to 0.$$

(b) Now, we consider $P(\min_{j \leq s} |T_j| \leq t)$. Define $\tilde{T}_j := T_j - \eta_j$.

Then similar with the lines in (a), we have

$$\sum_{j \leq s} P(|\tilde{T}_j| \geq t) \leq s \frac{2}{\sqrt{2\pi}} \frac{1}{t} e^{\frac{-t^2}{2}} (1 + f(t, n)) \to 0 \tag{3.6}$$

Let $\alpha_0 := \min_{j \leq s} |\eta_j|$ and it follows that

$$P(\min_{j \leq s} |T_j| \leq t) = P(\min_{j \leq s} |\tilde{T}_j + \eta_j| \leq t) \leq P(\min_{j \leq s} |\eta_j| - \max_{j \leq s} |\tilde{T}_j| \leq t) = P(\max_{j \leq s} |\tilde{T}_j| \geq \alpha_0 - t).$$

The inequality is from reverse triangle inequality.

Then

$$P(\min_{j\leq s}|T_j| \leq t) \leq P(\max_{j\leq s}|\tilde{T}_j| \geq \alpha_0 - t) \tag{3.7}$$

If $t \sim cn^{\frac{\gamma}{2}}$ and $\alpha_0 \sim n^{-\gamma}C_n$ for some $C_n/n^{\frac{3\gamma}{2}} \to c^*$, since $\alpha_0 - t \geq t$,

$$P(\max_{j\leq s}|\tilde{T}_j| \geq \alpha_0 - t) \leq P(\max_{j\leq s}|\tilde{T}_j| \geq t).$$

Therefore, using equations (3.6) and (3.7) we have

$$P(\min_{j\leq s}|T_j| \leq t) \to 0.$$

Combination of parts (a) and (b) complete the proof. □

3.3 Conclusion

We know that the two-sample t-test heavily depends on the assumption that the population are normally distributed. To over come the assumption, in this chapter we proposed a generalized feature selection method which does not need normality assumption. We need only asymptotic normality. Our generalized feature selection statistic is a special case of two-sample t-test, Wilcoxon-Mann Whitney, and two-sample proportion test statistics. We have shown that our generalized feature selection test method can pick up all the important features with probability 1.

Chapter 4

Applications of High-Dimensional Classification in Text Mining

4.1 Introduction

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

Text categorization/classification (TC) is the grouping of a text into two or more classes (Mahinovs & Tiwari, 2007). The goal of TC is to classify documents (academic articles, emails, etc) into categories. For example, news articles into "local" and "global", e-mails into "spam" and "ham", and customer feedbacks into "positive" and "negative" can be classified.

Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features which do not help separating the classes. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context for classification.

Weighting is a process consists of choosing terms which are important (contribute more than others) for a document and giving these terms more importance (weight) in the analysis. There are several methods to apply Weighting process; Boolean Retrieval, Term Frequency Weighting, and Term Frequency-Inverse Document Frequency (Tf-Idf) Weighting Methods (Adsiz, 2006).

Boolean Retrieval: if a word occurs in a document then the weight of the word is 1, otherwise it is 0.

$$a_{kj} = \begin{cases} 1 & \text{if } f_{kj} > 0, \\ 0 & otherwise \end{cases}$$

where a_{kj} is the j^{th} term of the k^{th} document, and f_{kj} is the number of occurrence of the term in the document.

Term Frequency Weighting: In this method, the number of occurrences of a word in a document is considered as weight for that word. That is, term frequency weight is equal to the number of occurrence of the word in document.

$$a_{kj} = f_{kj}.$$

Term Frequency-Inverse Document Frequency (Tf-Idf) Weighting: the tf-idf weights are often used to evaluate how a word is significant to a document. We know that tf measures how frequent a word in a document, while idf measures infrequency. We define the inverse document frequency of a term as follows:

$$idf_j = \log \frac{n}{df_j},$$

where n is the total number of documents and df_j is the number of documents in that the word occurs.

Therefore, the tf-idf weighting value for the j^{th} term of the k^{th} document is given by the following formula.

$$tf - idf_{kj} = f_{kj}(\log \frac{n}{df_j}).$$

Vector Space Models for Text: Although the vector space model ignores the context of each word in a document (commonly referred to as the "bag of words" approach), it is useful because it provides an efficient, quantitative representation of each document. In this

approach, documents are represented as vectors of length p, where p is the number of unique terms that are indexed in the collection. For any given document, the j^{th} entry of its vector representation is typically a function of the frequency of term j in that document multiplied by a weighting for the term. The vector for each document is generally very sparse (i.e., it contains a high proportion of zeroes) because few of the terms in the collection as a hole are contained in any one given document.

Text Preprocessing: In the text preprocessing, the key question is how to process unstructured textual information and extract meaningful numeric indices from the text. There are many special techniques for pre-processing text documents to make them suitable for mining. We first parse the text of documents into separate words, perform following preprocessing dimension reduction techniques and use the resulting information from the dimension reduction to significantly improve the classification accuracy of the documents.

Stopwords and punctuations removal: The stopword list used in this dissertation has the most frequent words that often do not carry much meaning. These words are the stopwords list from English language. Since including non-informative words will dilute our analysis, the data should be as clean and consistent as possible. After removing stopword list, a simple collection of low-information or extraneous words that you want to remove from the text such as a, an, the, be, with, by, etc., we can create a crucial start for obtaining valid and useful results. Moreover, synonym list can also be used to improve the quality of the text mining output, but creating the synonym list is a very labor intensive and time-consuming process. Normally, a change in the stopword list and synonym lists can dramatically alter the term weightings. Sometimes it is a bit hard to tune. The porter stemming algorithm treats words with the same stem as synonyms and you can use it as a substitute to synonym list. So by considering the costs and accuracy, we will not devote a large amount of effort to create good stopword and synonym lists and only throw away all punctuations and English language stopwords.

Excluding too short and long strings: The next factor we have to consider is a length of the character string. Short strings like us express a little useful information and meanwhile undesirably long and redundant strings are usually expected to have low frequencies. According to several experiments on given dataset, excluding too long and too short strings can highly reduce dimension and clean text. Thus, after removing all symbols but only letters and digits, the strings whose length are less than three and larger than sixteen will be excluded.

Porter stemming algorithm: English words like work can be inflected with a morphological suffix to produce works, working, worked which share the same stem work. The porter stemmer has five steps to progressively strip the suffixes as s, es, ed, ing, al, er, ic, able, ment, ive, etc. for short and long stems.

In addition, some special word like kept has kept as its lemma. However, using the porter stemmer, their link will be missed and the stemmed word for original word kept is still kept. Since the porter stemmer operates on a single word without knowledge of the context, and we cannot distinguish the words of different meanings solely depending on the part of speech. It requires lemmatization, a dictionary look-up process, which can essentially select the appropriate lemma depending on the context to solve this issue. But the porter stemmer is normally run faster and easier to implement.

If p is the number of distinct terms in a collection of $n = n_1 + n_0$ documents, then let A be the $n \times p$ matrix that represents this collection. This matrix is known as the document-term matrix, where the documents are rows (taken as the observations) and the terms are columns (taken as variables). The transpose of this matrix, where terms are rows and documents are columns, is known as the term-document frequency matrix. Let the document vectors be $\{\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}\}$ for class $\mathcal{C}_1$ and $\{\boldsymbol{X}_{01}, \boldsymbol{X}_{02}, \ldots, \boldsymbol{X}_{0n_0}\}$ for class $\mathcal{C}_0$. The document vectors for class $\mathcal{C}_i, i = 0, 1$ can be written as

$$\boldsymbol{X}_{ik} = (X_{ik1}\ X_{ik2}\ \ldots\ X_{ikp}),\ k = 1, \ldots, n_i;\ i = 0, 1,$$

where X_{ikj} is the weight frequency of term j in document k from class i. The document-term matrix is the $n \times p$ matrix given by

$$A = \begin{pmatrix} \boldsymbol{X}_{11} \\ \vdots \\ \boldsymbol{X}_{1n_1} \\ \boldsymbol{X}_{01} \\ \vdots \\ \boldsymbol{X}_{0n_0} \end{pmatrix}.$$

As an example, consider these collection of two of my gmail message subjects with the first is from class $\mathcal{C}_1$, which is in my inbox and the second one from class $\mathcal{C}_0$, which is in my spam where each message subject is considered to be a document:

1. Application Confirmation

2. Sale-Ink and toner sale up to 85% off

Then the corresponding document-term frequency matrix with term frequency weighting is displayed in the following table.

Documents	and	Application	Confirmation	Ink	off	Sale	to	toner	up
$\boldsymbol{X}_{11}$	0	1	1	0	0	0	0	0	0
$\boldsymbol{X}_{01}$	1	0	0	1	1	2	1	1	1

4.2 Singular Value Decomposition (SVD) and Principal Component Analysis (PCA)

Definition 5. *Singular Value Decomposition (SVD): Let $\boldsymbol{X}$ be a $n \times p$ matrix of real numbers. Then there exist an $n \times n$ orthogonal matrix U and a $p \times p$ orthogonal matrix V such that*

$$\boldsymbol{X} = U\Lambda V^T, \tag{4.1}$$

where the $n \times p$ *matrix* Λ *has* (j, j) *entry* $\lambda_j \geq 0$ *for* $j = 1, 2, \ldots, \min(p, n)$ *and the other entries are zero. The positive constants* λ_j *are called the singular values of* $\boldsymbol{X}$.

The columns of U are called the singular values for documents and the columns of V are the singular values for the terms.

The singular value decomposition can also be expressed as a matrix expansion that depends on the rank r of $\boldsymbol{X}$. Specifically, there exist r positive constants $\lambda_1, \lambda_2, \ldots, \lambda_r, r$ orthogonal $n \times 1$ unit vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r$, and r orthogonal $p \times 1$ unit vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r$, such that

$$\boldsymbol{X} = \sum_{j=1}^{r} \lambda_j \boldsymbol{u}_j \boldsymbol{v}_j^T = \boldsymbol{U}_r \Lambda_r \boldsymbol{V}_r^T,$$

where $\boldsymbol{U}_r = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r], \boldsymbol{V}_r = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_r]$, and Λ_r is an $r \times r$ diagonal matrix with diagonal entries λ_j.

Here $\boldsymbol{X}\boldsymbol{X}^T$ has eigenvalue-eigenvector pairs $(\lambda_j^2, \boldsymbol{u}_j)$, so

$$\boldsymbol{X}\boldsymbol{X}^T \boldsymbol{u}_j = \lambda_j^2 \boldsymbol{u}_j$$

with $\lambda_1^2, \lambda_2^2, \ldots, \lambda_r^2 > 0 = \lambda_{r+1}^2, \lambda_{r+2}^2, \ldots, \lambda_p^2$ for $(p > n)$. Then $\boldsymbol{v}_j = \lambda_j^{-1} \boldsymbol{X}^T \boldsymbol{u}_j$. Alternatively, the $\boldsymbol{v}_j$ are the eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ with the same nonzero eigenvalues λ_j^2.

Suppose that the random variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_0$ representing two classes $\mathcal{C}_1$ with mean vector $\boldsymbol{\mu}_1$ and $\mathcal{C}_0$ with mean vector $\boldsymbol{\mu}_0$ follow $p-$variate distributions with densities $f(\boldsymbol{Y}|\boldsymbol{\theta}_1)$ and $f(\boldsymbol{Y}|\boldsymbol{\theta}_0)$ respectively with Σ the common covariance matrix where $\boldsymbol{\theta}_i \in \boldsymbol{\Theta} = \{(\boldsymbol{\mu}_i, \Sigma) : \boldsymbol{\mu}_i \in \mathcal{R}^p, \det(\Sigma) > 0,\ i = 0, 1\}$ is the parameter space consisting of the mean vectors and the common covariance matrix. In other words,

$$\boldsymbol{Y}_i \sim f_i(\boldsymbol{Y}|\boldsymbol{\theta}_i) = f(\boldsymbol{Y}|\boldsymbol{\theta}_i), \quad i = 0, 1. \tag{4.2}$$

Definition 6. *Principal Component Analysis (PCA): Principal Component Analysis (PCA) specifies that the square covariance or correlation matrix* Σ *is formed and then the eigenvalue decomposition of* Σ *is calculated:*

$$\Sigma = Q\Lambda Q^T, \tag{4.3}$$

where Q *is the* $p \times p$ *orthogonal matrix whose columns are the eigenvectors of* Σ. Λ *is the diagonal matrix whose diagonal elements are the corresponding eigenvalues arranged in decreasing order.*

The columns of Q are called the principal components. We call Q^T the projection weight matrix W and the transformed data matrix S can be obtained from the original data matrix $\boldsymbol{X}$ by

$$S = \boldsymbol{X}W = \boldsymbol{X}Q. \tag{4.4}$$

Every data set has principle components, but PCA works best if data are Gaussian-distributed. For large sample size data the central limit theorem allows us to assume Gaussian distributions.

4.2.1 Comparing SVD and PCA

Although based on equivalent procedures, since PCA and SVD approach operate on different data, they do not produce the same results. Depending on whether the raw data is used or the covariance matrix is used, different vectors will be found as basis vectors for the reduced space. If we were to use the mean-adjusted document-term frequency data, rather than the raw data, the SVD approach and PCA, applied to the covariance matrix, would produce identical results (Albright, Russ (2004)).

4.2.2 Sparse vectors for SVDs

Suppose that

$$\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, \quad \boldsymbol{\mu}_a = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)/2, D = diag(\Sigma), \tag{4.5}$$

where $\boldsymbol{\mu}_1 = (S\bar{V}D_{11}, \ldots, S\bar{V}D_{1p})^T$ and $\boldsymbol{\mu}_0 = (S\bar{V}D_{01}, \ldots, S\bar{V}D_{0p})^T$. Let $t_{ikj}, k = 1, \ldots, n_i, i = 0, 1$ be the SVD term score for variable j from class i in the training data matrix. The sample mean vectors from classes $\mathcal{C}_1$ and $\mathcal{C}_0$, based on training data are given as,

$$S\bar{V}D_{1j} = \frac{1}{n_1}\sum_{k=1}^{n_1} t_{1kj} \text{ and } S\bar{V}D_{0j} = \frac{1}{n_0}\sum_{k=1}^{n_0} t_{0kj}, j = 1, \ldots, p. \tag{4.6}$$

Note that theoretically $D = \Sigma$, as the SVD's are uncorrelated. Note also that the SVD's are approximately normally distributed.

We define sparse vector and signal for the SVDs as follows:

Definition 7. *Suppose that* $\boldsymbol{\mu}_d = (\alpha_1, \alpha_2, \ldots, \alpha_s, 0, \ldots, 0)^T$ *is the* $p \times 1$ *mean difference vector where* $\alpha_j \in \mathbb{R}\backslash\{0\}, j = 1, 2, \ldots, s$. *We say that* $\boldsymbol{\mu}_d$ *is sparse if* $s = o(p)$. *Signal is defined as* $C_s = \boldsymbol{\mu}_d^T D^{-1} \boldsymbol{\mu}_d = \sum_{j=1}^{s} \frac{\alpha_j^2}{\sigma_j^2}$ *where* σ_j^2 *is the common variance for feature* j *in the two classes.*

4.2.3 Fisher and Naive Bayes Discriminant Functions for SVDs

Let π_0 and π_1 be the class prior probabilities for classes $\mathcal{C}_0$ and $\mathcal{C}_1$ respectively. A new observation $\boldsymbol{Y}$ is to be assigned to one of $\mathcal{C}_1$ or $\mathcal{C}_0$. The optimal classifier is the Bayes rule:

$$\delta(\boldsymbol{Y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_0) = \mathbf{1}\left\{\log \frac{f(\boldsymbol{Y}|\boldsymbol{\theta}_1)}{f(\boldsymbol{Y}|\boldsymbol{\theta}_0)} > \log \frac{\pi_0}{\pi_1}\right\}, \tag{4.7}$$

where $\mathbf{1}$ denotes the indicator function with value 1 corresponds to assigning $\boldsymbol{Y}$ to $\mathcal{C}_1$ and 0 to class $\mathcal{C}_0$.

Unless specified, throughout this section we let that $\boldsymbol{Y}_1 \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \Sigma)$ and $\boldsymbol{Y}_0 \sim \mathcal{N}_p(\boldsymbol{\mu}_0, \Sigma)$. Under these assumptions (4.7) becomes

$$\delta(\boldsymbol{Y}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, \Sigma) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_a) > \log\frac{\pi_0}{\pi_1}\right\}. \tag{4.8}$$

Note that if $\pi_1 = \pi_0 = 1/2$, then we have the Fisher discriminant rule:

$$\delta_F(\boldsymbol{Y}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, \Sigma) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_a) > 0\right\}, \tag{4.9}$$

with corresponding misclassification error rate

$$W(\delta_F, \boldsymbol{\theta}) = \bar{\Phi}\left(\frac{(\boldsymbol{\mu}_d^T \Sigma^{-1} \boldsymbol{\mu}_d)^{1/2}}{2}\right). \tag{4.10}$$

Alternatively, assuming independence of components and replacing off-diagonal elements of Σ with zeros leads to a new covariance matrix,

$$D = diag(\Sigma), \tag{4.11}$$

and a different discrimination rule, the Naive Bayes,

$$\delta_{NB}(\boldsymbol{Y}, \boldsymbol{\mu}_d, \boldsymbol{\mu}_a, D) = \mathbf{1}\left\{\boldsymbol{\mu}_d^T D^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_a) > 0\right\}, \tag{4.12}$$

whose misclassification error rate is

$$W(\delta_{NB}, \boldsymbol{\theta}) = \bar{\Phi}\left(\frac{\boldsymbol{\mu}_d^T D^{-1}\boldsymbol{\mu}_d}{2(\boldsymbol{\mu}_d^T D^{-1}\Sigma D^{-1}\boldsymbol{\mu}_d)^{1/2}}\right). \tag{4.13}$$

Note that when $D = \Sigma$, when the variables are uncorrelated, then Fisher and Naive Bayes would coincide.

For the raw data: generally $D \neq \Sigma$, as the terms may be correlated, then Fisher and Naive Bayes would produce different results.

For the SVD: since theoretically $D = \Sigma$, Fisher and Naive Bayes are equivalent. Therefore, Naive Bayes is optimal method.

It is important for us to distinguish between discriminative and signal sets. The definitions in (Mai etal., 2012) are given below.

Definition 8. *A discriminative set is* $A = \{j : \{(\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\}_j \neq 0\}$, *since the Bayes classification direction is* $\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. *Variables in A are called discriminative variables.*

Definition 9. *The signal set is defined as* $\tilde{A} = \{j : \boldsymbol{\mu}_{1j} \neq \boldsymbol{\mu}_{0j}\}$; *variables in* $\tilde{A}$ *are called signals.*

Ideally, $\tilde{A}$ is the variable selection outcome of an independence rule. Practically, independence rules pick the strongest signals indicated by the data. When Σ is diagonal, $A = \tilde{A}$. For a general covariance matrix, however, the discriminative and the signal sets can be very different.

In most text mining problems, we reduce the dimension of the feature space using the Singular Value Decomposition (SVD). Researchers take the first few SVDs for classification. They choose the first m important ones based on the ratio

$$\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{j=1}^{p}\lambda_j}$$

which is the amount of variation explained by the first m SVDs.

In this dissertation, we argue that the first few singular value decompositions (SVDs) which account for most of the variation are not necessarily the most important ones for classification. The variation comes from both noise and signal. If the variation of noise is more than the variation for signal, even though, we have large variation, the SVD will not carry much information for classification. Therefore, we further select the important SVDs based on two sample t-test which gets us the discriminative set. We are motivated by the following simple example:

Let $\boldsymbol{\mu}_1 = (\alpha, 0, 0)^T, \alpha \neq 0, \boldsymbol{\mu}_0 = (0, 0, 0)^T, \sigma_{jj} = 1,$ and $\sigma_{kj} = \rho \in [0, 1)$ when $k \neq j$ for $j = 1, 2, 3$. The eigen-values of Σ are $\lambda_1(\Sigma) = 1 + 2\rho, \lambda_2(\Sigma) = 1 - \rho,$ and $\lambda_3(\Sigma) = 1 - \rho$ with one possible choice corresponding eigen-vectors $e_1 = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^T, e_2 = (1/\sqrt{2}, -1/\sqrt{2}, 0)^T$, and $e_3 = (1/\sqrt{6}, 1/\sqrt{6}, -2/\sqrt{6})^T$ respectively. Let $\boldsymbol{X}_1 = (x_{11}, x_{12}, x_{13})^T$ and $\boldsymbol{X}_0 = (x_{01}, x_{02}, x_{03})^T$ be random vectors from classes $\mathcal{C}_1$ and $\mathcal{C}_0$ respectively. Then, $SVD_{ij} = e_j^T(\boldsymbol{X}_i - \boldsymbol{\mu}_a), i = 0, 1,$ and $j = 1, 2, 3$. Therefore, the SVD vectors are $\boldsymbol{Y}_1 = (SVD_{11}, SVD_{12}, SVD_{13})^T$ and $\boldsymbol{Y}_0 = (SVD_{01}, SVD_{02}, SVD_{03})^T$ for the two classes respectively.

Our aim is to choose one SVD out of the three SVDs. The absolute value of the expected two sample t-statistics are $T_1 = \frac{|\alpha|}{\sqrt{3(1+2\rho)}\sqrt{1/n_1+1/n_0}}, T_2 = \frac{|\alpha|}{\sqrt{2(1-\rho)}\sqrt{1/n_1+1/n_0}}, T_3 = \frac{|\alpha|}{\sqrt{6(1-\rho)}\sqrt{1/n_1+1/n_0}}$ respectively. It is easy to show that the second SVD has higher absolute value two sample t-statistic value than the first and third. Therefore, the important one for classification will be the second SVD not the first.

We generalize the above example into the following theorem for equal mean difference and equicorrelation matrix.

Theorem 4.1. *Let $\boldsymbol{\mu}_d$ be the $p \times 1$ equal sparse mean difference vector, each with $\alpha \neq 0$, with the number of non-zero mean differences being s and Σ is an equicorrelation matrix with off-diagonals $\rho \in [0, 1)$. Suppose that $(p - s)^2 \geq s^2(p - 1)$ (which works for all ρ) or $\rho \geq \frac{s^2(p-1)-(p-s)^2}{(p-s)(p-1)+s^2(p-1)}$. Then there is an SVD with index $j \neq 1$ such that the absolute value of its expected value of the t-statistic is the largest and has better classification performance.*

Proof of Theorem 4.1: Let us consider the mean difference vector $\boldsymbol{\mu}_d$ and let $j=$ the first location, from left to right, for which we get all the s non-zero mean differences. It easy to see that $j = s, s+1, \ldots, p$. We want to show that for all j in $\{s, s+1, \ldots, p\}$ it is true that the absolute value of the expected value of the j^{th} SVD is bigger or equal to the absolute value of the expected value of the 1^{st} SVD.

Note that the eigen-values for Σ are $\lambda_1(\Sigma) = 1 + (p-1)\rho, \lambda_2(\Sigma) = \lambda_3(\Sigma) = \cdots = \lambda_p(\Sigma) = 1 - \rho$. Let us take the following choice of eigen-vectors for Σ:

$$\begin{aligned}
\boldsymbol{e}_1^T &= \left(\tfrac{1}{\sqrt{p}}, \tfrac{1}{\sqrt{p}}, \ldots, \tfrac{1}{\sqrt{p}}\right) \\
\boldsymbol{e}_2^T &= \left(\tfrac{1}{\sqrt{1\times 2}}, \tfrac{-1}{\sqrt{1\times 2}}, 0, \ldots, 0\right) \\
\boldsymbol{e}_3^T &= \left(\tfrac{1}{\sqrt{2\times 3}}, \tfrac{1}{\sqrt{2\times 3}}, \tfrac{-2}{\sqrt{2\times 3}}, 0, \ldots, 0\right) \\
&\vdots \\
\boldsymbol{e}_i^T &= \left(\tfrac{1}{\sqrt{(i-1)\times i}}, \ldots, \tfrac{1}{\sqrt{(i-1)\times i}}, \tfrac{-(i-1)}{\sqrt{(i-1)\times i}}, 0, \ldots, 0\right) \\
&\vdots \\
\boldsymbol{e}_p^T &= \left(\tfrac{1}{\sqrt{(p-1)\times p}}, \ldots, \tfrac{1}{\sqrt{(p-1)\times p}}, \tfrac{-(p-1)}{\sqrt{(p-1)\times p}}\right).
\end{aligned}$$

Let us consider the following absolute values of the expected values for the t-statistics:

$$\begin{aligned}
E[T_1] &= \tfrac{|\boldsymbol{e}_1^T \boldsymbol{\mu}_d|}{\mathrm{SD}(\boldsymbol{e}_1^T \boldsymbol{\mu}_d)} &&= \tfrac{s|\alpha|/\sqrt{p}}{\sqrt{\frac{1+(p-1)\rho}{n_1} + \frac{1+(p-1)\rho}{n_0}}} \\
E[T_j] &= \tfrac{|\boldsymbol{e}_{s+1}^T \boldsymbol{\mu}_d|}{\mathrm{SD}(\boldsymbol{e}_{s+1}^T \boldsymbol{\mu}_d)} &&= \tfrac{s|\alpha|/\sqrt{(s+1)s}}{\sqrt{\frac{1-\rho}{n_1} + \frac{1-\rho}{n_0}}}, \text{ for } j = s \\
E[T_j] &= \tfrac{|\boldsymbol{e}_j^T \boldsymbol{\mu}_d|}{\mathrm{SD}(\boldsymbol{e}_j^T \boldsymbol{\mu}_d)} &&= \tfrac{(j-s)|\alpha|/\sqrt{(j-1)j}}{\sqrt{\frac{1-\rho}{n_1} + \frac{1-\rho}{n_0}}}, \text{ for } j = s+1, \ldots, p.
\end{aligned}$$

It is enough to show that $E[T_j] \geq E[T_1]$ for $j = s, s+1, \ldots, p$. It is obvious to see that $E[T_s] \geq E[T_1]$ for $p >> s$.

For $j = p$, $E[T_p] = \frac{(p-s)|\alpha|/\sqrt{(p-1)p}}{\sqrt{\frac{1-\rho}{n_1} + \frac{1-\rho}{n_0}}}$. It is easy to see that the statement $E[T_p] \geq E[T_1]$ is equivalent to $((p-s)(p-1) + s^2(p-1))\rho + (p-s)^2 - s^2(p-1) \geq 0$. If $(p-s)^2 \geq s^2(p-1)$ or $\rho \geq \frac{s^2(p-1)-(p-s)^2}{(p-s)(p-1)+s^2(p-1)}$, it is easy to see the inequality holds. Hence, $E[T_p] \geq E[T_1]$.

Since $E[T_j]$ is a decreasing function of j, we have $E[T_j] \geq E[T_1]$ for $j = s+1, s+2, \ldots, p-1$. This completes the first part of the proof.

To prove the second part of theorem 4.1: note that for the Naive Bayes rule the misclassification error rate using the first SVD alone is $\bar{\Phi}\left(\frac{(\boldsymbol{e}_1^T\boldsymbol{\mu}_d)^T\boldsymbol{e}_1^T\boldsymbol{\mu}_d}{2\sqrt{(\boldsymbol{e}_1^T\boldsymbol{\mu}_d)^T\lambda_{\max}(\Sigma)\boldsymbol{e}_1^T\boldsymbol{\mu}_d}}\right) = \bar{\Phi}\left(\frac{|\boldsymbol{e}_1^T\boldsymbol{\mu}_d|}{2\sqrt{\lambda_{\max}(\Sigma)}}\right)$. Similarly, and the misclassification error rate using the j^{th} SVD alone is $\bar{\Phi}\left(\frac{|\boldsymbol{e}_j^T\boldsymbol{\mu}_d|}{2\sqrt{\lambda_j(\Sigma)}}\right)$. From the first part of theorem 4.1, we know that $\frac{|\boldsymbol{e}_1^T\boldsymbol{\mu}_d|}{2\sqrt{\lambda_{\max}(\Sigma)}} \leq \frac{|\boldsymbol{e}_j^T\boldsymbol{\mu}_d|}{2\sqrt{\lambda_j(\Sigma)}}$. Since, $\bar{\Phi}(x)$ is a decreasing function of x, the misclassification error rate using the first SVD alone is higher than the misclassification error rate using the j^{th} SVD alone. This completes the second part of the proof. □

Let us consider the following example to illustrate the conditions in theorem 4.1. Consider $p = 4500$. The following figure shows s versus $f(s) = \frac{s^2(p-1)-(p-s)^2}{(p-s)(p-1)+s^2(p-1)}$, the right hand side in the condition of theorem 4.1.

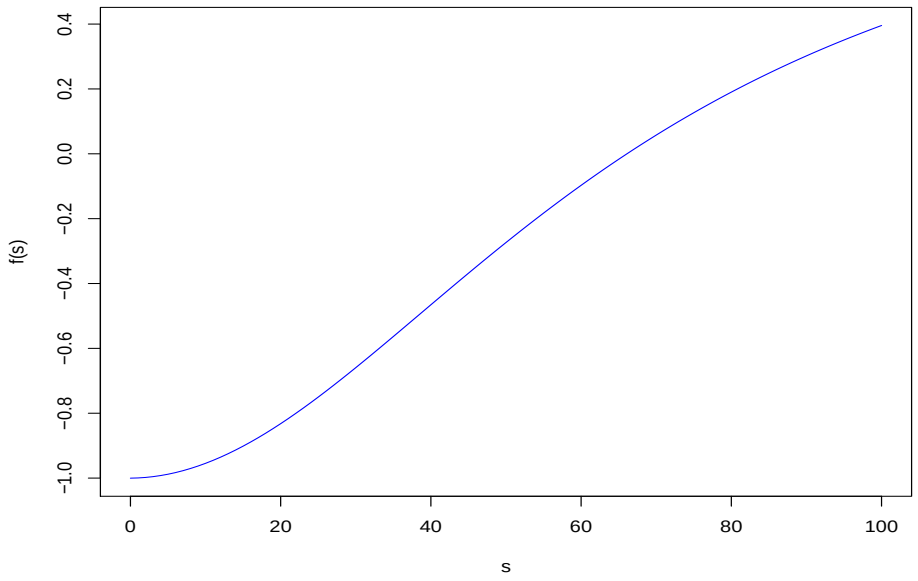


Figure 4.1: *The horizontal axis is s and vertical axis is $f(s)$.*

As we can see from the above figure, we need $s \leq 66$, before we make any restrictions on ρ. For example, if we take $s = 90$, we need $\rho \geq 0.3019376$ so that the conditions on the theorem 4.1 to hold.

Note that the conditions in theorem 4.1 can be relaxed to $s < p/2$ if the first s features have non-zero mean difference.

4.3 Partitioning of the Data Matrix into Training and Validation Data Matrices

We then take a random sample of size n_1^* and n_0^* ($n^* = n_1^* + n_0^*$) for validation and/or prediction performance, leaving n_1 and n_0 from each class to be used in the training data set, with $n = n_1 + n_0$ documents and p terms. Thus, $N = n^* + n = (n_1^* + n_0^*) + (n_1 + n_0) = (n_1^* + n_1) + (n_0^* + n_0)$. The rows of our original document-term matrix, A, are ordered so that it can be represented as the joining of two matrices, the training document-term matrix, A_T, and the validation document-term matrix, A_V,

$$A = \begin{pmatrix} A_T \\ A_V \end{pmatrix},$$

where the first n_1 rows of A_T represent the documents from class $\mathcal{C}_1$ training documents and the next n_0 rows of A_T represent the class $\mathcal{C}_0$ documents, and the $n_1^* + n_0^*$ rows of A_V are similarly ordered. It is conventional to choose $n^* \leq n$.

The prediction model is based on Singular Value Decomposition (SVD) of A_T and the SVD transformations derived from A_T are used on the validation/prediction data in A_V. So in the next section, we focus on the SVD of the training data set.

4.3.1 Singular Value Decomposition of the Training Data Matrix

Suppose our document-term matrix, A_T, is $n \times p$ where n represents the number of documents and p represents the number of rows. Let A_T be the centered document-term matrix where the mean frequencies of each term are subtracted from the frequencies of these

terms in each document. That is, the mean term frequency vector is given by,

$$\bar{\boldsymbol{a}} = \frac{1}{n}(A_T)^T.\mathbf{1} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{a}_i. \tag{4.14}$$

Then the centered training and validation document-term matrices are given subtracting the mean term frequency vector, given in (4.14),

$$A^* = A - \bar{\boldsymbol{a}},\ A_T^* = A_T - \bar{\boldsymbol{a}} \text{ and } A_V^* = A_V - \bar{\boldsymbol{a}}. \tag{4.15}$$

To keep the notion simple we just refer to the centered matrices in (4.15), as A_T and A_V.

The singular value decomposition (SVD) of A_T is given as

$$A_T = U\Lambda V^T, \tag{4.16}$$

where $U(n\times n)$ is the eigenvector matrix of $A_T(A_T)^T$ and $V(p\times p)$ is the eigenvector matrix of $(A_T)^T A_T$. Recall, $rank((A_T)^T A_T) = rank(A_T(A_T)^T) \leq \min(n, p)$. Note in our case, $p >> n$. To keep track of its origin, i.e. the SVD is on the training matrix, we may denote U as U_T.

SVD Scoring on variables (terms): V contains the weights for the SVD scoring on terms and U contains the weights for SVD scoring on documents. We want the scores on terms so the matrix containing the SVD scores is given as, transformed document-term matrix (rows are the documents and columns are SVDs)

$$\text{SVD}_T = A_T V. \tag{4.17}$$

SVD Scoring on Observations (documents): U contains the weights for the SVD scoring on document if we want the scores on documents for each term

$$\text{SVD} = U^T A_T. \tag{4.18}$$

SVD Term Scoring on the hold-out data (validation term-document matrix, A_V): Using the matrix V given in (4.16) and used in (4.17), the transformed SVD document-term matrix for A_V is given as

$$\text{SVD}_{A^T A_V} = A_V V. \tag{4.19}$$

Notice that the transformation in (4.17) and (4.19) could be completed using the centered document-term matrix, A, given in (4.15), so

$$SVD_A = \begin{pmatrix} A_T V \\ A_V V \end{pmatrix}.$$

So, the columns in the above matrix contain the sample SVD values $(\text{SVD}_1, \text{SVD}_2, \ldots, \text{SVD}_p)$ for each document from each dataset (training and prediction) and each class (class $\mathcal{C}_1$ and $\mathcal{C}_0$).

4.3.2 Sorting Features Based on T-statistics on the Training Data

In this process, for each of the N rows of SVD_A from the above matrix (SVD_A), we compute the absolute value of the two sample t-statistics using the first N_1 and N_0 rows of SVD_A. Then we sort the rows of SVD_A from the largest to smallest absolute t-statistics. Here we assume that the SVDs are approximately normally distributed and are sparse vectors.

4.4 Overall Prediction Modeling

4.4.1 Feature Selection Algorithm

The following are the steps for our new algorithm:

Step 1: Partition the rows of the document-term matrix into the training and validation document-term matrices, A_T and A_V, and produce the matrix A formed by joining A_T and A_V.

Step 2: Compute the mean term frequency vector from A_T and compute centered matrices, A and A_T.

Step 3: Get the SVD transform matrix $V = V_T$ based on A_T , then compute then transform the full document-term matrix $\text{SVD}_A = A_T V = AV$.

Step 4: Compute the absolute value of the two-sample t-statistics using the first N_1+N_0 rows of SVD_A, then sort from the largest to smallest.

Step 5: Compute discriminant functions for all observations using the statistics (mean vectors and variance-covariance matrix) from SVD_A, and compute the misclassification errors for the models that use, $m = 1, 2, \ldots, p$ features. The misclassification error rates are calculated based on the Naive Bayes or Fisher discriminant rule.

4.5 Real Data Analysis

4.5.1 NASA flight data set

The NASA flight data set can be found at https://c3.nasa.gov/dashlink/resources/138/.

This is the dataset used for the SIAM 2007 Text Mining competition. This competition focused on developing text mining algorithms for document classification. The documents in question were aviation safety reports that documented one or more problems that occurred during certain flights. The goal was to label the documents with respect to the types of

problems that were described. This is a subset of the Aviation Safety Reporting System (ASRS) dataset, which is publicly available. The data for this competition came from human generated reports on incidents that occurred during a flight.

In our analysis we combined the training and testing data to get 28596 documents. We then pick two flight problems (or two columns)-column 2 and column 19 which correspond to the faults "Operation in noncompliance" and "Aircraft Malfunction" respectively because these are the two largest faults out of the 22 faults. Each of the faults have uniquely 2081 and 2486 documents respectively. We divide our data into three groups: training (1081 from the first fault and 1486 from the second fault), validation (500 from each of the two faults), and testing (500 from each of the two faults). After cleaning we have $p = 26694$ terms. The following graph shows how Naive Bayes vs Fisher performs.

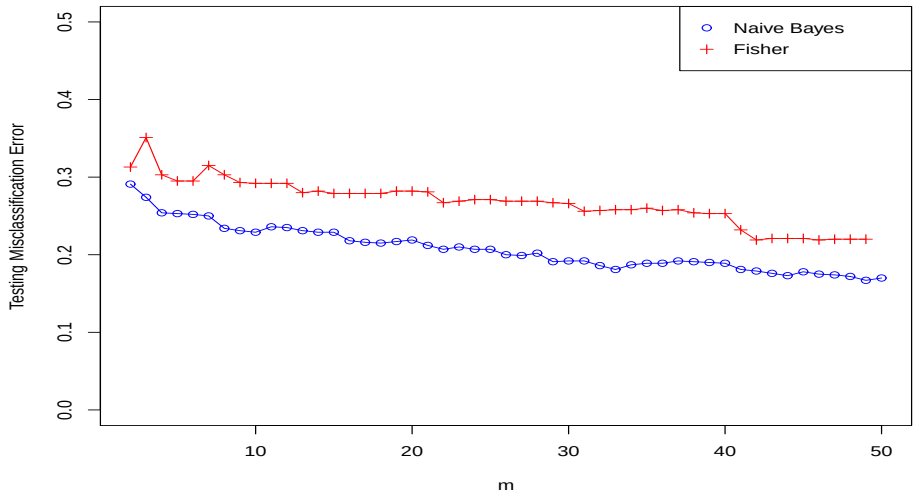


Figure 4.2: *Horizontal axis is the number of features used and the vertical axis is testing misclassification error rate for Fisher vs Naive Bayes.*

As we can see from the above figure, Naive Bayes dominates Fisher. For Naive Bayes classifier the optimal number of features (or terms) selected using the validation data set is 148 with minimum error rate 0.116.

After we apply SVD transformations for the combined document-term matrix, the first 10 indexes for the rankings of the SVDs by applying two-sample t-test is 2, 6, 9, 7, 5, 19, 8, 15, 36, and 13. As we can see the first SVD does not show up even in the first 10 ranks. The following figure shows how Naive Bayes and Fisher performs after we take the SVD transformation on the training data. As we can see from the figure the SVD after we rank them based on the two sample t-statistic performs better than the unranked ones for both Naive Bayes and Fisher. We can also see that Naive Bayes and Fisher perform close to each other as the SVDs are uncorrelated. This results supports our theoretical result given in this chapter.

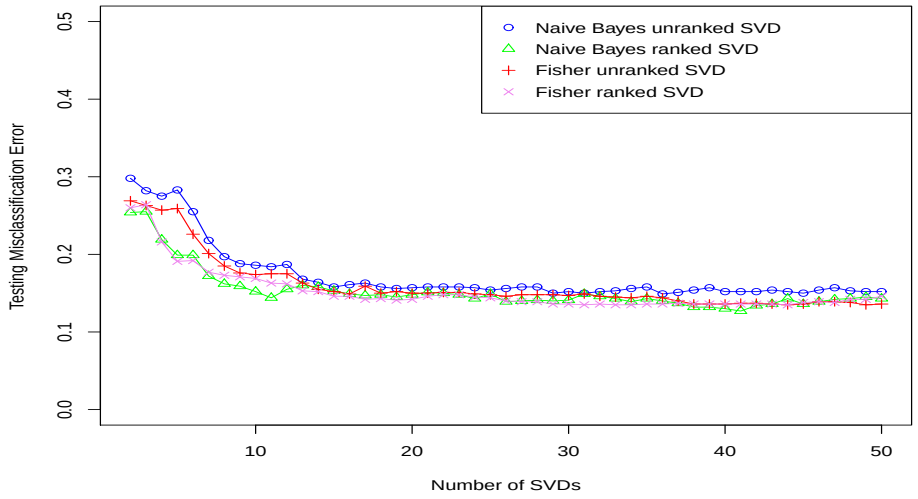


Figure 4.3: *Horizontal axis is the number of svds used and the vertical axis is testing misclassification error rate for Naive Bayes vs Fisher.*

For Naive Bayes classifier the optimal number of svds selected using the validation data set is 28 with minimum error rate 0.117. Comparing it with before the SVD transformation, we can see that we achieve almost the same error rate for fewer SVDs.

4.5.2 DBWorld Email Messages

DBWorld mailing list announces conferences, jobs, books, software and grants. Publishing new announcements does not require to provide their category. Some writers use to insert specific acronyms in the title (e.g. CFP, CFW, CFE), although it is not a widely shared practice.

Michele Falannino (2011) have manually collected the last 64 e-mails that he received and he has built two different data sets. The first one uses only the subjects, while the second one uses bodies. Both data sets have been represented by a term-document matrix using one of the most common data structure in Text mining: bag of words. Every e-mail is represented as a vector containing p binary values, where p is the size of the vocabulary extracted from the entire corpus. The binary value is 1 if the corresponding word belongs to the document, 0 otherwise. Features are unique words extracted from the entire corpus with some constraints: words that have more than 3 characters with a maximum length of 30 characters. Bag-of-words model produces a large number of features, also in the case in which there are few documents. In both data sets, he has also removed stop words. The data set of subjects has got 242 features while the second one has got 4702 features. Both have $64 = 29 + 35$ samples. Each data set contains also a further binary feature that indicates the class of each sample: 1 if the sample is an announcement of conference, 0 otherwise.

We then apply the Naive Bayes rule on the SVDs of the whole term-document matrix based on the subjects. We then calculated the training errors based on the SVDs and the SVDs ranked according to two-sample $t-$statistic. We call the SVDs ranked according to the two-sample t-test SVD then FAIR (which is the Feature Annealed Rule given in Fan and Fan (2008)). FAIR is applying the Naive Bayes rule on the selected features. As we can see the ranked SVDs do have lower error rates than the SVDs without ranking them by their corresponding two-sample $t-$test statistic. We do the same analysis for bodies too. The following figures are for the subjects and the bodies respectively.

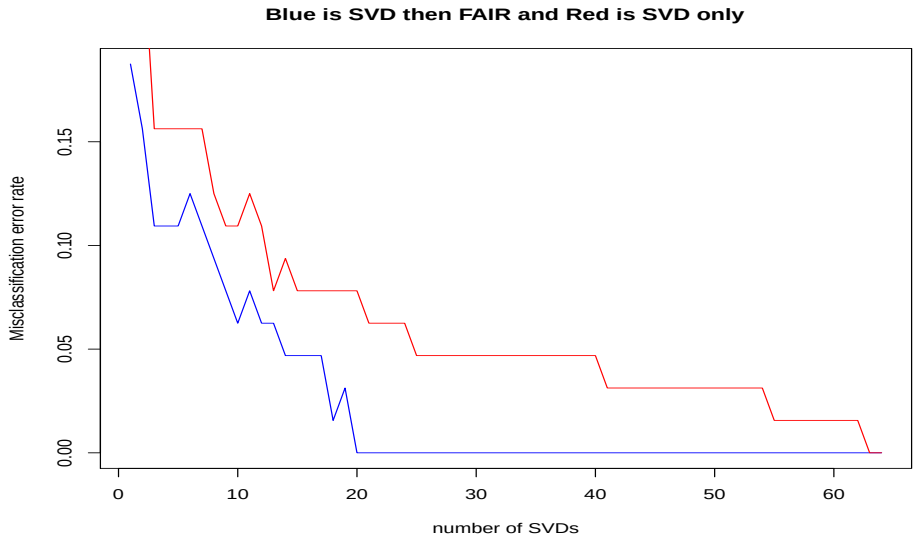


Figure 4.4: Error rate vs number of svds for subjects. We can see that ranking the svds based on their t-statistic improves the error rate. We need also fewer number of svds.

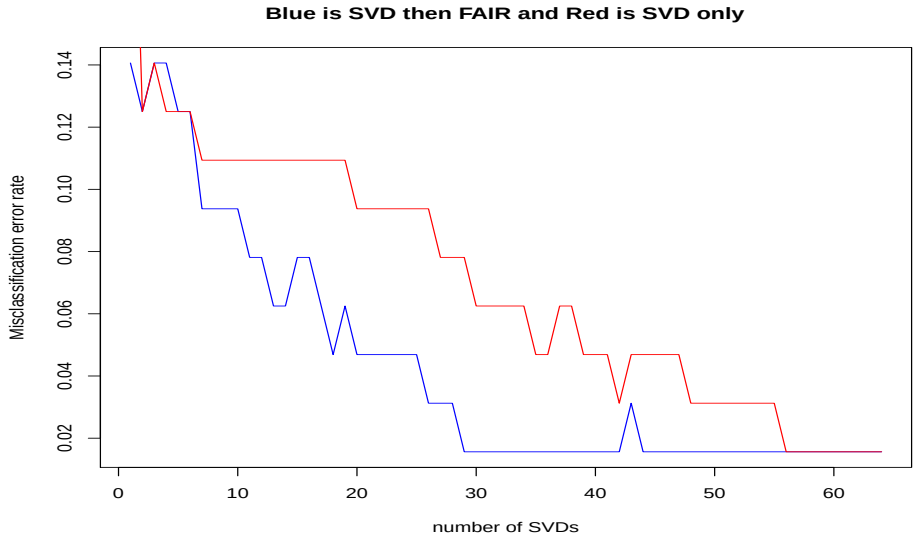


Figure 4.5: Error rate vs number of svds for bodies. We can see that ranking the svds based on their t-statistic improves the error rate. We need also fewer number of svds.

4.6 Conclusion

In this chapter we considered a binary classification text mining problem. We know that researchers take the first few SVDs which account for most of the variation. But we have shown that the first singular value decompositions (svds) which account for most of the variation are not necessarily the most important ones for classification. This is because the noise may be higher than the signal for the first few svds. We then select the important SVDs based on the two-sample t-test. We have given a new feature selection algorithm for text mining problems. Our flight and DBWorld data analyses support our theory.

Chapter 5

Summary and Future Work

5.1 Summary

In this dissertation we considered a binary classification problem when the feature dimension p is much larger than the sample size n. We know that Fisher is an optimal classifier and Naive Bayes is sub optimal for the population model. But in the second chapter we have given conditions under which Naive Bayes is optimal. Through theory, simulation and data analysis we have shown that Naive Bayes is the practical method to use than Fisher for high-dimensional data. In designing binary classification experiments, Fisher requires full correlation structure but using equicorelation structure we can design our experiment using Naive Bayes. Through simulation we characterized that the two-sample t-test can pick up all the important features as far the signal is not too low. In the third chapter we proposed a generalized independence feature selection method and we showed that our test statistic can pick up all the important features with probability converging to 1. Our generalized feature selection method includes the two-sample t-statistic, Wilcoxon Mann Whitney statistic, and the two-sample proportion test. In the fourth chapter we considered a two-class text mining classification problem. We showed that the first few singular value decompositions (svds) are not necessarily the most important ones for classification. When we first apply svd transformation in our training document-term matrix, and we then further select the important svds based on the two-sample t-statistic the misclassification error rates can be reduced. We select the optimal number of svds based on the performance on the validation data. Our data analyses examples have showed the improvement on the error rates.

5.2 Future Work

Our future work include extending the two-class classification problem into multi-class problem. We have theoretically tracked the sample error rates for Naive Bayes and we are working on the sample error rate for Fisher. We are also interested in extending the theory and simulation given for equicorrelation structure to any correlation structure. Extending the simulation and theoretical results given in chapter 2 will be also an interest. In this dissertation we considered only the two-sample t-test as our feature selection method. We are working on the general theory for heavy tailed distributions. Specifically, applying the generalized feature selection method for feature screening and comparing it with the features selected using two-sample t-test.

Bibliography

[1] Aigars Mahinovs and Ashutosh Tiwari. Text Classification Method Review. Decision Engineering Report Series, April 2007.

[2] Albright, Russ (2004). Taming Text with the SVD. SAS Institute Inc., cary, NC.

[3] Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. Bernoulli **10**, 989-1010.

[4] Cao, Hongyuan (2007). Moderate Deviations For Two Sample T-Statistics. ESAIM: Probability and Statistics, Vol. **11**, 264271.

[5] Cormack, G., Gomez, J., and Sanz, E. (2007). Spam Filtering for Short Messages. Proceedings of the sixteenth ACM conference on information and knowledge management, 313-320.

[6] Cornish, E.A. (1954). The multivariate small t-distribution associated with a set of normal sample deviates. Australian J. of Physics, **7**, 531-542.

[7] Ding, B. and Gentleman, R. (2005). Classification using generalized partial least squares. J. Comput. Graph. Stat., **14(2)**: 280-298.

[8] Dunnett, C.W. and Sobel, M. (1954). A bivariate generalization of Student's t distribution with tables for certain special cases, Biometrika **41**, 153-169.

[9] Deerwester, S., Dumais, G., Landauer, T. and Harshman, R.(1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, **41**, 391-407.

[10] Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Statist. Assoc., **97**, 7787.

[11] Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. Ann. Statist., **36**, 2605-2637.

[12] Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. J. R. Statist. Soc. B. **74**, 745-771.

[13] Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultra-High Dimensional Feature Space. J. R. Statist. Soc. B. **70**, 849-911.

[14] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space (invited review article). Statistica Sinica **20**, 101-148.

[15] Filannino, M. (2011). DBWorld e-mail classification using a very small corpus. The University of Manchester.

[16] Froda, Sorana and Eeden, Constance (2000). A uniform saddlepoint expansion for the null-distribution of the Wilcoxon-Mann-Whitney statistic. The Canadian Journal of Statistics. Vol. **28**, No. 1, 2000, 137-149.

[17] Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under $l1$ constraint. Ann. Statist. **34**, 2367-2386.

[18] Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli **10**, 971-988.

[19] Guo, Y., Hastie, T. and Tibshirani, R. (2005). Regularized discriminant analysis and its application in microarrays. Biostatistics, **1**, 1-18.

[20] Hastie, T., Tibshirani, R., and Friedman, J.(2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition). Springer-Verlag, New York.

[21] Jin, Rungao and Robinson, John (2003). Saddlepoint Approximations of the Two-sample Wilcoxon Statistic. IMS, 149-158.

[22] Joachims, Thorsten (1997). Text categorization with support vector machines. Technical report, LS VIII Number 23, University of Dortmund.

[23] Hajek, J. and Sidak, Z. (1967). Theory of Rank Tests. Academic Press, New York.

[24] Karim M. Abadir and Jan R. Magnus. Matrix Algebra. Cambridge University Press, 2005.

[25] Lange, K. L, Little, R. J. A and Taylor, J. M. G (1989). Robust statistical modeling using the t distribution. Journal of the American Statistical Association, **84**, 881-896.

[26] Liao, C., Li, S. and Luo Z. (2007). Gene selection using wilcoxon rank sum test and support vector machine for cancer classification. In Y. Wang, Y.-m. Cheung, and H. Liu, editors, Computational Intelligence and Security, volume **4456** of Lecture Notes in Computer Science, pages 57-66. Springer Berlin / Heidelberg.

[27] Little, R.J.A. (1988). A test of missing compeltely at random for multivariate data with missing values. Journal of the American Statistical Assciation. Dec. 1988, Vol. **83**, No. 404, Theory and Methods.

[28] Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley and Sons.

[29] Mai, Q., Zou, H., and Yau, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. Biometrika **99**, 29-42.

[30] Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics, Vol. **18**, No. 1 (Mar., 1947), 50-60.

[31] Meyer, C.D. (2000). Matrix Analysis and Applied Linear Algebra. SIAM, Philadelphia.

[32] Nguyen, D. V. and Rocken, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics, **18(1)**: 39-50.

[33] Richard A. Johnson and Dean W. Wichern (6th edition). Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.

[34] Roger A. Horn and Charles R. Johnson. Matrix Analysis. Cambridge University Press, 1985.

[35] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. **99**, 6567-6572.

[36] Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. Annals of Math. Stat., Vol. **11**, 147-162.