Nonparametric Methods for Classification and Related Feature Selection Procedures

by

Shuxin Yin

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 6, 2010

Keywords: Gene expression, Principal component, Misclassification, Rank-bassed

Approved by

Asheber Abebe, Chair, Associate Professor of Mathematics and Statistics
Peng Zeng, Associate Professor of Mathematics and Statistics
Ming Liao, Professor of Mathematics and Statistics

Abstract

One important application of gene expression microarray data is classification of samples into categories, such as types of tumor. Gene selection procedures become crucial since gene expression data from DNA microarrays are characterized by thousands measured genes on only a few subjects. Not all these genes are thought to determine a specific genetic trait. In this dissertation, I develop a novel nonparametric procedure for selecting such genes. This rank-based forward selection procedure rewards genes for their contribution towards determining the trait but penalizes them for their similarity to genes that are already selected. I will show that my method gives lower misclassification error rates than the dimension reduction methods such as principal component analysis and partial least square analysis. I also explore more properties of Wilcoxon-Mann-Whitney (WMW) statistic and propose a new classifier based on WMW to reduce the misclassification error rate. Real data analysis and Monte Carlo simulation demonstrate the superiority of the proposed methods to the classical methods in several situations.

Acknowledgments

My greatest thanks extend to Dr. Asheber Abebe for his patience and guidance throughout this endeavor. Dr. Abebe took considerable time and energy to further the progress in my studies of Statistics. I also wish to express my appreciation for my parents and other family members Man Peng and Shuyu Yin who have given of their love and constant support throughout my life. I would also like to thank Dr. Peng Zeng and Dr. Ming Liao on the committee for their contribution to this dissertation and serving as the committee members.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Nonparametric Methods

Originally nonparametric methods were introduced in the mid-1930. The rank correlation without normality assumption was discussed in Hotelling and Past (1936) which was considered as the the true beginning of topic of nonparametric statistics. Friedman (1940) developed the Fried test which was a nonparametric statistical test to detect differences in treatments across multiple test attempts in the complete block design. Durbin (1951) proposed the nonparametric test for the incomplete block design that reduces to the Friedman test in the case of a complete block design. Benard and Elteren (1953) generalized Durbin test to the case in which several observations are taken on some experimental units. During the same period, Wilcoxon (1945) proposed the signed-rank statistic named Wicoxon statistic to test the significance of the location differences of two samples. Later Mann and Whitney (1947) introduced the Mann-Whitney test statistic which is equivalent to Wilcoxon rank test statistic. Wilcoxon statistic played a central role in many nonparametric approaches in 1950s and 1960s.

Nonparametric methods have emerged as the preferred methodology in many scientific areas due to their outstanding advantages:

- Nonparametric procedures require fewer assumptions than the traditional methods so that they can be used more widely than the corresponding parametric methods. In particular, nonparametric procedures are applicable and more efficient in many situations where normality assumption is violated.

- Nonparametric methods are distribution-free methods, which do not rely on assumptions that the data are drawn from a certain probability distribution. So they can be used in many complicated situations where the distribution theory is not achievable.

- Nonparametric methods are resistant to outliers. When the data contain some outliers or the longer tail than the normal distribution, some traditional statistical procedures are inefficient, even though they can perform well when the error in the model follow a normal distribution.

- Another advantage for the use of non-parametric methods is simplicity. In some cases, even when the use of parametric methods is justified, non-parametric methods may be easier to use. Due both to this simplicity and to their greater robustness, non-parametric methods are preferred by some statisticians.

In recent years, nonparametric analysis has gained its popularity in the analysis of linear model (Sievers and Abebe, 2004), non-linear model (Abebe and McKean, 2007), classification (Nudurupati and Abebe, 2009; Montanari, 2004), generalized estimating equations (Abebe *et al.*, 2011), etc because it leaves less room for the improper use and misunderstanding.

1.2 Classification

Some of the basic ideas and history in classification is discussed in the following. Consider we have two populations, and the main goal of classification is to determine the membership of a new observation based on the training data set. A discriminant function is needed to find the criterion in order to assign the new observations and it generally projects the multidimensional real space into one dimension real line such that a clear cutting value of discriminant function can be applied to determine which class the new observation probably belongs. Fisher (1936) gave the linear discriminant classifier. He found the optimal projection direction by maximizing the two-sample t-statistic and allocated the new observations based on the Euclidian distance between the new observations and the centers of populations. If the covariance matrices of populations are not equal, quadratic discriminant classifier is preferred. Bickel and Levina (2004) proposed the independence classifier by setting the non-diagonal entries of the common covariance matrix to be zero.

However those methods above are adversely inefficient in many circumstances where the normality assumption is not proper because of their sensitivity to the skewness and outliers. Their limitations call for some robust rank-based classifiers which are highly related to the idea of transvariation probability given in Gini (1916). Transvariation probability is originally defined for the univariate case and can be extended to the multivariate case by following a certain projection pursuit. Montanari (2004) proposed transvariation-distance classifier to allocate a new observation according to the Euclidian distance to population centers in the projected space. He found the optimal projection direction that maximizes the two-sample Mann-Whitney-Wilcoxon (WMW) statistic (Lehmann, 2006). He also proposed the point-group classifier to determine the likelihood of the new observation belonging to which population based on the same projection pursuit. The results by using point-group classifier, however can be biased when two sample sizes are too different. An improved allocation method was proposed by Nudurupati and Abebe (2009). They put the new observation in two samples separately to smooth the data depth. We can also use some depth functions to measure the group separation in order to classify a new observation as shown in Liu *et al* (1999). A few popular depth functions are Mahalanobis depth (Mahalanobis, 1936; Liu and Singh, 1993), halfspace depth (Tukey, 1974), simplicial depth (Liu, 1990), majority depth (Singh, 1991), projection depth (Donoho, 1982), and spatial or L_1 depth (Vardi and Zhang, 2000).

1.3 Dimension Reduction

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, such as the ones involving the projection pursuit. Typically, the calculation working load can be increased dramatically by even adding one more gene if a projection pursuit is needed. 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 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.

Because of the sensitivity of some classifiers to the dimension, the dimension is needed to be reduced further even though the initial variable screening is applied. Based on the genes selected by the variables screening procedures, some dimension reduction methods can be introduced to reduce the dimension by performing a linear mapping of data to a lower dimensional space, such as principle component analysis (PCA) and partial least square analysis (PLS). In PCA (Massey, 1965; Jolliffe, 1986), a small set of orthogonal linear combinations of the original predictor variables can be found by maximizing the variance of these linear combinations matrix. In practice, the correlation matrix of the data is constructed and the eigenvectors on this matrix are computed. The eigenvectors that correspond to the largest eigenvalues can now be used to reconstruct a large fraction of the variance of the original data. Then the first few eigenvectors are selected and can often be considered as the optimal linear combinations and used in classification. Thus the original space is reduced to the lower dimensional space spanned by a few eigenvectors without loss of the information carried by the original variables.

In PCA, however, the correlation between the predictor variables and the response variable specifying the class of the observations is not considered, which may be inefficient. Efficient one must not treat the predictors separately from the response. Nguyen and Rocke (2002) proposed a new approach to obtain the optimal combinations by maximizing the covariance between those linear combinations and response vector, which is referred to as partial least square analysis. PLS surpasses PCA by taking the relation to response variable into account. A numerical algorithm to obtain the components is also included in that paper.

1.4 Organization

In Chapter 2, I discuss seven different classification methods as well as several dimension reduction methods and gene screening procedures. In Chapter 3, I prove that the WMW can pick up all the important variables with the probability tending to 1 followed by a comparison of the performances of WMW statistic on variable screening and classification with the procedure given in Fan and Fan (2008) using two real data sets and a large simulation study. Besides, a smoother is recommended when two sample sizes are too different. In Chapter 4, I propose a new forward variable selection method and demonstrate its superiority using some real data analysis and simulation.

1.5 Notations

Here are some notations used throughout my dissertation:

- Consider two populations $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$ with underlying distributions F and G with common support $\mathbb{R}^p$.
- $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$ are the mean vectors of $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$
- $\boldsymbol{\Sigma_x}$ and $\boldsymbol{\Sigma_y}$ are the covariance matrices of $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$
- $\mathbb{X}$ and $\mathbb{Y}$ are two samples from $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$ with sample sizes n_x and n_y respectively.
- $\mathbf{X}$ and $\mathbf{Y}$ are the predictor matrices of two samples, and $\mathbf{R}$ is the response vector (indictor of tumor versus normal tissue).

Chapter 2
Background

During the past decade and a half, classification and clustering methods have gained popularity for cancer classification based on gene expression profiles obtained via DNA microarray technology. But the dimension of microarray data is usually ultrahigh which makes some classifier inapplicable. Moreover we believe that a subset of genes whose expression suffices for accurately predicting the response. Dimension reduction allows us to replace a very large number of predictors with a small number of linear combinations, and perform a better prediction based on those optimal linear combinations. For the sake of simplicity, before I use some dimension reduction approaches, some variable screening methods are applied to get rid of those irrelevant variables which have smaller variation than noise measurement.

2.1 Variable Screening Procedures

Two sample t-statistic can be considered as a measurement of group separation since it can evaluate the differences in means between two groups. Larger t-statistic value implies the better group separation. Thus we firstly rank all the variables based on their two sample t-statistic. The two-sample t-statistic for variable k is defined as

$$T_k = \frac{\overline{x}_k - \overline{y}_k}{\sqrt{s_{xk}^2/n_x + s_{yk}^2/n_y}}, \quad k = 1, \ldots, p \tag{2.1}$$

where $\overline{x}_k$ and $\overline{y}_k$ are the means of two samples respectively for variable k. s_{xk}^2 and s_{yk}^2 are variances of of two samples respectively for variable k. I select the ones with the larger values of t-statistic to be the most informative variables.

However, t-statistic is sensitive to the outliers and skewness, so the nonparametric alternatives are recommended to be applied to measure the differences between two groups when the normality assumption is violated. The most commonly used one is WMW statistic which is robustly measuring the differences of two groups.

Rank the variables based on the two sample WMW statistic

$$W_k = 1 - \frac{2}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \phi \left\{ (x_{ki} - y_{kj})\, m_{xy}^k \right\} \tag{2.2}$$

where m_{xy}^k is the median of $\{x_{ki} - y_{kj}\}$ for $k = 1, \ldots, p$. The function ϕ is defined as

$$\phi(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x \geq 0 \ . \end{cases} \tag{2.3}$$

The most informative variables are the ones with the larger WMW statistic which indicates the less overlapped area under the density curves of two populations. More details are discussed in Chapter 3.

2.2 Dimension Reduction Methods

After initial variable screening procedure, most noninformative variables are eliminated and only few important ones left. But the dimension is still too high for some classifiers, especially for the ones requiring the projection pursuit. I consider to apply the dimension reduction approaches to reduce the dimension further. The purpose of dimension reduction is to create a smaller set of linear combinations of original variables without loss of too much information carried by the original ones. This is achieved by optimizing a defined objective criterion. PCA and PLS are two well-known dimension reduction methods.

2.2.1 Principal Component Analysis

In PCA (Massey, 1965; Jolliffe, 1986), orthogonal linear combinations are constructed to maximize the variance of the linear combinations of the predictor variables sequentially

$$\mathbf{c}_k = \underset{\mathbf{c}'\mathbf{c}=1}{\text{Argmax}}\, Var(\mathbf{Pc}), \text{ where } \mathbf{P} = \mathbf{X} \cup \mathbf{Y}$$

subject to the orthogonality constraint

$$\mathbf{c}'\mathbf{S}\mathbf{c}_j = 0 \text{ for all } 1 \leq j \leq k, \text{ where } \mathbf{S} = \mathbf{P}'\mathbf{P}$$

In fact, let $n = n_x + n_y$ and $\mathbf{V}$ be the $n \times p$ matrix found by stacking $\mathbf{X}$ and $\mathbf{Y}$; that is

$$\mathbf{V} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_{n_x} \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{n_y} \end{bmatrix}.$$

where $\mathbf{x}_i$ and $\mathbf{y}_j$ are the rows in $\mathbf{X}$ and $\mathbf{Y}$ respectively. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$ be the eigenvalues of $\mathbf{V}'\mathbf{V}$ and $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r$ be the corresponding eigenvectors, where r is the rank of $\mathbf{V}'\mathbf{V}$. The ith principal component is then $\mathbf{V}\boldsymbol{\alpha}_i$, which is a linear combination of the original columns of $\mathbf{V}$. The first principal component accounts the direction of maximum variability in the data, and each succeeding component accounts for increasingly smaller amounts of variability. Several optimal linear combinations can be obtained by using the eigenvectors corresponding to the larger eigenvalues.

PCA, however, only measures the variability in the predictor data $\mathbf{V}$ without any consideration to the contribution of variables towards the classification problem. Ignoring the relation to the response variable may make the components selected by PCA short of the information in terms of the group separation and consequently results in inaccuracy in classification. A dimension reduction approach considering the correlation between predictors and response variables is needed.

2.2.2 Partial Least Square Analysis

Nguyen and Rocke (2002) proposed the method of PLS that sequentially maximizes the covariance between the response variable and a linear combinations of predictor variables. In our case, the response variable $\mathbf{R}$ is made up of n_x zeros and n_y ones as $\mathbf{R} = (0, \ldots, 0, 1, \ldots, 1)'$. Thus in place of the eigenvectors used in PCA, PLS uses

$$\boldsymbol{\alpha}_i = \underset{\|\boldsymbol{\alpha}\|=1}{\text{Argmax}} \ \text{cov}(\mathbf{V}\boldsymbol{\alpha}, \mathbf{R})$$

subject to the constraint $(\mathbf{V}\boldsymbol{\alpha})'(\mathbf{V}\boldsymbol{\alpha}_i) = 0$ for $i = 1, \ldots, r$. This object criterion for the dimension reduction may be more appropriate for the prediction since the relation between predictors and response variable is considered. A basic algorithm implementing PLS is given in Nguyen and Rocke (2002). For the details, see also Helland (1988), Garthwaite (1994), Höskuldsson (1988), and Martens and Naes (1989).

2.3 Classification Methods

In general, classification is to solve the problem of classifying a new observation $\mathbf{z} \in \Pi_{\mathbf{X}} \cup \Pi_{\mathbf{Y}}$ in either $\Pi_{\mathbf{X}}$ or $\Pi_{\mathbf{Y}}$. We need to define a discriminant function to project the multidimensional data space into one dimensional real line such that we can make the decision that where the new observation belongs:

Definition 2.1. *A discriminant function* $\mathscr{D}(\mathbf{z}; F, G) : \mathbb{R}^p \to \mathbb{R}$ *is such that* $\mathbf{z}$ *is classified in* $\Pi_{\mathbf{X}}$ *if* $\mathscr{D}(\mathbf{z}; F, G) > 0$ *and in* $\Pi_{\mathbf{Y}}$ *otherwise.*

Discriminant function gives a linear combination of the predictor variables, whose values are as close as possible within populations and as far apart as possible between populations. Several popular classifiers are discussed in the following.

2.3.1 Non-Robust Classification

Linear Discriminant Analysis (LDA)

Fisher (1936) looked at a linear combination of the p-covariates that maximizes the separation between the two populations $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$. This gives rise to the linear discriminant function

$$\mathscr{L}(\mathbf{z}; F, G) \equiv (\boldsymbol{\mu_x} - \boldsymbol{\mu_y})' \, \boldsymbol{\Sigma}^{-1} \left[\mathbf{z} - \frac{1}{2} (\boldsymbol{\mu_x} + \boldsymbol{\mu_y}) \right] \, ,$$

where $\boldsymbol{\mu_y}$ and $\boldsymbol{\mu_y}$ are as defined in Section 1.5, and $\boldsymbol{\Sigma}$ is the pooled covariance matrix of F and G. A new observation $\mathbf{z}$ is classified in $\Pi_{\mathbf{X}}$ if $\mathscr{L}(\mathbf{z}; F, G) > 0$ and in $\Pi_{\mathbf{Y}}$ otherwise. Such classification is referred to as Linear Discriminant Analysis (LDA).

Given samples $\mathbb{X}$ and $\mathbb{Y}$ from $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$, respectively, the discriminant function of LDA is estimated by $\mathscr{L}(\mathbf{z}; F_{n_x}, G_{n_y})$, where F_{n_x} is the empirical distribution function of $\mathbf{X}$ obtained by putting mass $1/n_x$ on each $\mathbf{x}$ sample point and G_{n_y} is the empirical distribution function of $\mathbf{Y}$. Henceforth, it will be assumed that estimates of discriminant functions are obtained by replacing distribution functions by the corresponding empirical distribution function.

Quadratic Discriminant Analysis (QDA)

In LDA, we assume two populations share the same covariance matrix. If this assumption is not held, another commonly used classification method named Quadratic

Discriminant Analysis (QDA) can be applied, which is based on the classical multivariate normal model for each class. Assume $P_{\mathbf{X}} = P_{\mathbf{Y}}$ be the equiprobable priors of two populations. The quadratic discriminant function is defined as

$$Q(\mathbf{z}; F, G) = Q(\mathbf{z}; F) - Q(\mathbf{z}; G) \ ,$$

where

$$Q(\mathbf{z}; F) = -\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu_x})'\boldsymbol{\Sigma_x}^{-1}(\mathbf{z} - \boldsymbol{\mu_x}) - \frac{1}{2}\log\left(\left|\boldsymbol{\Sigma_x}^{-1}\right|\right)$$

$$Q(\mathbf{z}; G) = -\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu_y})'\boldsymbol{\Sigma_y}^{-1}(\mathbf{z} - \boldsymbol{\mu_y}) - \frac{1}{2}\log\left(\left|\boldsymbol{\Sigma_y}^{-1}\right|\right)$$

Here $\boldsymbol{\Sigma_x}$ and $\boldsymbol{\Sigma_y}$ are as defined in 1.5. A new observation $\mathbf{z}$ is classified in $\Pi_{\mathbf{X}}$ if $Q(\mathbf{z}; F, G) > 0$ and in $\Pi_{\mathbf{Y}}$ otherwise.

Independence Classifier

Both LDA and QDA require the projection pursuit, which is not an option for the large dimensional data. In particular, in order to evaluate some variable selection methods, we usually need to add more and more variables into the classification model to find the optimal number of variables for the prediction. Independence classifier can be applied for this purpose:

$$\mathscr{I}(\mathbf{z}; F, G) \equiv (\boldsymbol{\mu_x} - \boldsymbol{\mu_y})' \, D^{-1}\left[\mathbf{z} - \frac{1}{2}\left(\boldsymbol{\mu_x} + \boldsymbol{\mu_x}\right)\right] \ ,$$

where $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$ are as defined in Section 1.5, and $D = diag(\boldsymbol{\Sigma})$. A new observation $\mathbf{z}$ is classified in $\Pi_{\mathbf{X}}$ if $\mathscr{I}(\mathbf{z}; F, G) > 0$ and in $\Pi_{\mathbf{Y}}$ otherwise. As matter of fact, linear discriminant function can be reduced to independence discriminant function by setting the non-diagonal entries of common covariance matrix to be zero. This classifier is discussed in Bickel and Levina (2004). They also show that it is superior to LDA when the number of variables is large.

2.3.2 Rank Based Classification using Projections

LDA amounts to finding $\mathbf{u} \in \mathbb{R}^p$, say $\hat{\mathbf{u}}_0$, that maximizes the square of the two-sample t-statistic between the two projected samples $\mathbf{u}'\mathbb{X} = \{\mathbf{u}'\mathbf{x}_1, \ldots, \mathbf{u}'\mathbf{x}_{n_x}\}$ and $\mathbf{u}'\mathbb{Y} = \{\mathbf{u}'\mathbf{y}_1, \ldots, \mathbf{u}'\mathbf{y}_{n_y}\}$; that is

$$\hat{\mathbf{u}}_0 = \underset{\|\mathbf{u}\|=1}{\text{Argmax}} \frac{[\mathbf{u}'(\bar{\mathbf{x}} - \bar{\mathbf{y}})]^2}{\mathbf{u}'\mathbf{S}_p\mathbf{u}\left(\frac{n_x+n_y}{n_x n_y}\right)} \ ,$$

where $\mathbf{S}_p$ is the pooled covariance matrix, and $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are means of two samples. The data are then reduced to one dimension by projecting them in the direction given by $\hat{\mathbf{u}}_0$ and one would classify a new observation $\mathbf{z}$ into $\Pi_{\mathbf{X}}$ if $|z_0 - \bar{x}_0| < |z_0 - \bar{y}_0|$, where $x_{0i} = \hat{\mathbf{u}}_0'\mathbf{x}_i$, $y_{0j} = \hat{\mathbf{u}}_0'\mathbf{y}_j$, and $z_0 = \hat{\mathbf{u}}_0'\mathbf{z}$, $i = 1, \ldots, n_x$ and $j = 1, \ldots, n_y$. Otherwise, one classifies $\mathbf{z}$ into $\Pi_{\mathbf{Y}}$. Here $\bar{x}_0 = n_x^{-1} \sum_{i=1}^{n_x} x_{0i}$ and $\bar{y}_0 = n_y^{-1} \sum_{j=1}^{n_y} y_{0j}$.

When the underlying distributions are spherically symmetric, the direction of maximum separation is along the line that connects the centers of the distributions. LDA is equivalent to classifying $\mathbf{z}$ based on its Euclidean distance from the means. In the case of the normal distribution, the projection direction can be obtained easily as $\mathbf{u}_0 = \mathbf{S}_p^{-1/2}(\bar{\mathbf{x}} - \bar{\mathbf{y}})/\|\mathbf{S}_p^{-1/2}(\bar{\mathbf{x}} - \bar{\mathbf{y}})\|$. In other situations, the projection direction is not obvious and has to be determined numerically. This search for "interesting" low dimensional projection of high dimensional data is known as *projection pursuit* (Friedman and Tukey, 1974). "Interestingness" is measured through a suitable function known as the *projection index.* For LDA, this index is the two-sample t-statistic.

Montanari (2004) and Chen *et al.* (1994) used a two-sample WMW type statistic as a projection index to measure group separation. They showed that their projection pursuit method is not sensitive to deviations from the homoscedasticity and normality assumptions. Their method is related to the idea of transvariation probability given in Gini (1916):

Definition 2.2. *For univariate distributions F and G, the Transvariation Probability is defined as*

$$\tau(F, G) = \int_R \int_R \phi((x-y)(\mu(F) - \mu(G)))d(F(x))d(G(y))$$

where $\mu(F)$ and $\mu(G)$ are the medians of F and G.

This transvariation probability is the measure of common area under underlying distribution curves of two populations. Of course, the smaller transvariation probability indicates the better group separation.

In practice, instead of using the theoretical underlying distributions of two populations, we can use empirical distributions of $\mathbb{X}$ and $\mathbb{Y}$ to estimate τ

$$\tau^* = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \phi\{(\mathbf{x}_{ki} - \mathbf{y}_{kj})(m_x - m_y)\}$$

where m_x and m_y are medians of two samples.

However, this transvariation probability is only defined under univariate space, so for the multidimensional data, we need to redefine a general transvariation probability. This general transvariation probability can be redefined on a vector $\mathbf{u}$ through a certain projection pursuit.

The direction of minimum overlap measured by the general transvariation probability is given by

$$\hat{\mathbf{u}}_1 = \underset{\|\mathbf{u}\|=1}{\text{Argmin}} \left\{ \sum \phi \left\{ [\mathbf{u}'\mathbf{x} - \mathbf{u}'\mathbf{y}][m_x(\mathbf{u}) - m_y(\mathbf{u})] \right\} \right\} , \tag{2.4}$$

where the sum is over the set $\{\mathbf{x} \in \mathbb{X},\ \mathbf{y} \in \mathbb{Y}\}$ and $m_x(\mathbf{u})$ and $m_y(\mathbf{u})$ are medians of two projected samples $\mathbf{u}'\mathbb{X}$ and $\mathbf{u}'\mathbb{Y}$. The function ϕ is defined as in (2.3)

Once the direction of maximum separation is found, the next step is to project all the data (including the new sample point) onto that direction and allocate the new point to one of the two populations. Three allocation schemes are discussed in the following.

Transvariation-Distance (TD) Classifier

Montanari (2004) proposed classifying a new observation $\mathbf{z}$ in $\Pi_{\mathbf{X}}$ if

$$|\hat{\mathbf{u}}_1'\mathbf{z} - m_x(\hat{\mathbf{u}}_1)| < |\hat{\mathbf{u}}_1'\mathbf{z} - m_y(\hat{\mathbf{u}}_1)|$$

and in $\Pi_{\mathbf{Y}}$ otherwise, where $m_x(\hat{\mathbf{u}}_1)$ and $m_y(\hat{\mathbf{u}}_1)$ are medians of two projected groups. Hereafter the classifier obtained by using this allocation method will be referred to as Transvariation-Distance (TD) classifier. As shown in Nudurupati and Abebe (2009), this method can be adversely affected by skewness and outliers.

Point-Group Transvariation (PGT) Classifier

Another allocation method suggested by Montanari (2004) is based on a comparison of the ranking of the new observation $\mathbf{z}$ among $\mathbb{X}$ and among $\mathbb{Y}$. This utilizes the point-group transvariation. The observation $\mathbf{z}$ is classified in $\Pi_{\mathbf{X}}$ if $\mathscr{T}(\mathbf{z}; F, G) \equiv T(\mathbf{z}; F) - T(\mathbf{z}; G) > 0$, where

$$T(\mathbf{z}; F) = \frac{1}{n_x} \sum_{\mathbf{x} \in \mathbb{X}} \phi \left\{ [\hat{\mathbf{u}}_1'\mathbf{x} - \hat{\mathbf{u}}_1'\mathbf{z}][m_x(\hat{\mathbf{u}}_1) - \hat{\mathbf{u}}_1'\mathbf{z}] \right\}$$

and

$$T(\mathbf{z};G) = \frac{1}{n_y}\sum_{\mathbf{y}\in\mathbb{Y}} \phi\left\{[\hat{\mathbf{u}}_1'\mathbf{y} - \hat{\mathbf{u}}_1'\mathbf{z}][m_x(\hat{\mathbf{u}}_1) - \hat{\mathbf{u}}_1'\mathbf{z}]\right\} \ .$$

This allocation scheme is robust against skewness and outliers. However, it does not perform well for data with unequal sample sizes. This is because the vote of each member of $\mathbb{X}$ is either 0 or $1/n_x$ whereas the vote of each member of $\mathbb{Y}$ is 0 or $1/n_y$. This allocation scheme has also a problem of ties between $T(\cdot;F)$ and $T(\cdot;G)$. The likelihood of ties is the greatest in the case of equal sample sizes, which happens to be the only situation where this scheme works efficiently. We will use random tie breaking where a *coin* is flipped to decide allocation in the case of a tie. The classifier obtained by using this allocation scheme will be referred to as Point-Group Transvariation (PGT) classifier.

Group-Group Transvariation (GGT) Classifier

An improved allocation method was proposed by Nudurupati and Abebe (2009) to eliminate the problem caused by unequal sample sizes. Define two augmented samples $\mathbb{X}^*$ and $\mathbb{Y}^*$ by including the new point $\mathbf{z}$ in the two samples; that is $\mathbb{X}^* = \mathbb{X}\cup\{\mathbf{z}\}$ and $\mathbb{Y}^* = \mathbb{Y}\cup\{\mathbf{z}\}$. The point $\mathbf{z}$ is then classified in $\Pi_{\mathbf{X}}$ if $\mathscr{T}^*(\mathbf{z};F,G) \equiv \mathscr{T}_1^*(\mathbf{z};F,G) - \mathscr{T}_2^*(\mathbf{z};F,G) > 0$ where

$$\mathscr{T}_1^*(\mathbf{z};F,G) = \frac{1}{(1+n_x)n_y}\sum \phi\left\{[\hat{\mathbf{u}}_1'\mathbf{x}^* - \hat{\mathbf{u}}_1'\mathbf{y}][m_x(\hat{\mathbf{u}}_1) - \hat{\mathbf{u}}_1'\mathbf{z}]\right\}$$

and

$$\mathscr{T}_2^*(\mathbf{z};F,G) = \frac{1}{n_x(1+n_y)}\sum \phi\left\{[\hat{\mathbf{u}}_1'\mathbf{x} - \hat{\mathbf{u}}_1'\mathbf{y}^*][m_y(\hat{\mathbf{u}}_1) - \hat{\mathbf{u}}_1'\mathbf{z}]\right\}$$

The two sums are over the sets $\{\mathbf{x}^*\in\mathbb{X}^*,\ \mathbf{y}\in\mathbb{Y}\}$ and $\{\mathbf{x}\in\mathbb{X},\ \mathbf{y}^*\in\mathbb{Y}^*\}$, respectively.

The classifier obtained by using this allocation scheme will be referred to as Group-Group Transvariation (GGT) classifier. Note that we do not have the unequal voting problem here. The vote of all observations is either 0 or approximately $(n_x n_y)^{-1}$. In essence, here we are smoothing one sample using the empirical distribution of the other before applying the PGT rule.

2.3.3 Maximum Depth Classifiers

In the univariate setting, statistical methods that use rank-based nonparametric techniques do not depend on restrictive distributional assumptions and hence are robust to deviations from these assumptions. For higher dimensions, *statistical depth functions* give a multivariate version of ranks (Liu, 1992). Depth functions give a measure of the "centrality" of a given multivariate sample point with respect to its underlying distribution (Liu *et al*, 1999). In particular, a depth function assigns higher values to points that are more central with respect to a data cloud. This naturally gives a center-outward ranking of the sample points. A number of depth functions are available in the literature. A few popular depth functions are Mahalanobis depth (Mahalanobis, 1936; Liu and Singh, 1993), halfspace depth (Tukey, 1974), simplicial depth (Liu, 1990), majority depth (Singh, 1991), projection depth (Donoho, 1982), and spatial or L_1 depth (Vardi and Zhang, 2000).

In this paper, I use the maximum depth (MaxD) classification method (Ghosh and Chaudhuri, 2005) based on spatial (L_1) depth function defined as

$$\mathscr{S}(\mathbf{x}; F) = 1 - \left\| E_F \left\{ \frac{\mathbf{x} - \mathbf{X}}{\|\mathbf{x} - \mathbf{X}\|} \right\} \right\| \,, \tag{2.5}$$

where $\mathbf{X} \sim F$ and $\| \cdot \|$ is the Euclidean norm on $\mathbb{R}^p$.

The classifier MaxD uses the discriminant function

$$\mathscr{S}^*(\mathbf{z}; F, G) = \mathscr{S}(\mathbf{z}; F) - \mathscr{S}(\mathbf{z}; G)$$

and classifies $\mathbf{z}$ in $\Pi_{\mathbf{X}}$ if $\mathscr{S}^*(\mathbf{z}; F, G) > 0$. A major drawback of this classifier is that it lacks affine invariance. Thus it is necessary to transform the data so that all the variables are similarly scaled before using the spatial depth function. Vardi and Zhang (2000) suggest to make the spatial depth function affine invariant by taking ${\mathbf{\Sigma}_{\boldsymbol{x}}}^{-1/2}(\mathbf{z} - \mathbf{X})$ and ${\mathbf{\Sigma}_{\boldsymbol{y}}}^{-1/2}(\mathbf{z} - \mathbf{Y})$ in place of $\mathbf{z} - \mathbf{X}$ and $\mathbf{z} - \mathbf{Y}$ before computing $\mathscr{S}(\mathbf{z}; F)$ and $\mathscr{S}(\mathbf{z}; G)$, respectively, using equation (2.5). Note that one can use any affine equivariant estimators of $\mathbf{\Sigma}_{\boldsymbol{x}}$ and $\mathbf{\Sigma}_{\boldsymbol{y}}$ when computing the discriminant function. If the scatter estimator of Tyler (1987) is used, then the resulting maximum spatial depth classifier resembles the classifier given by Crimin *et al.* (2007). An alternative method of obtaining affine invariance is to scale the data along its principal component directions (PCA-scaling) as given in Hugg *et al* (2006).

An estimate of the MaxD discriminant function $\mathscr{S}^*(\mathbf{z}; F, G)$ is given by

$$\mathscr{S}^*(\mathbf{z}; F_{n_x}, G_{n_y}) = \left\| \frac{1}{n_y} \sum_{j=1}^{n_y} \frac{\mathbf{z} - \mathbf{y}_j}{\|\mathbf{z} - \mathbf{y}_j\|} \right\| - \left\| \frac{1}{n_x} \sum_{i=1}^{n_x} \frac{\mathbf{z} - \mathbf{x}_i}{\|\mathbf{z} - \mathbf{x}_i\|} \right\| .$$

Chapter 3

Applications of Wilcoxon-Mann-Whitney Statistic (WMW)

3.1 Feature Annealed Independence Rules (Fan and Fan, 2008)

Two sample t-statistic can be applied as an initial variable screening index because of its contribution to the group separation. Fan and Fan (2008) proved that theoretically t-statistic can pick up all the informative variables with probability approaching to 1.

They consider the p-dimensional classification problem between two populations $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$. $\mathbb{X}$ and $\mathbb{Y}$ are two samples from $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$ with sample sizes n_x and n_y respectively. Write ith observation in $\Pi_{\mathbf{X}}$ as

$$\mathbf{x}_i = \boldsymbol{\mu_x} + \boldsymbol{\epsilon_{xi}},$$

and ith observation in $\Pi_{\mathbf{Y}}$ as

$$\mathbf{y}_i = \boldsymbol{\mu_y} + \boldsymbol{\epsilon_{yi}},$$

where $\boldsymbol{\epsilon_{xi}} = (\epsilon_{xij})$ and $\boldsymbol{\epsilon_{yi}} = (\epsilon_{yij})$ are iid with mean 0 and covariance matrix $\boldsymbol{\Sigma_x}$ and $\boldsymbol{\Sigma_y}$ respectively. They assume that all the observations are independent across samples and in addition, within one population, observations are identically distributed. They also assume that the two classes have compatible sample size.

They first proved that without variable selection, discrimination based on linear projections to almost all directions performs nearly the same as random guessing under some assumptions. They then claim that using t-statistic defined in 2.1, all the informative variables can be selected in the below Theorem (3.1). In order to prove their theorem, they need the following conditions.

Condition1

- Assume that the vector $\boldsymbol{\alpha} = \boldsymbol{\mu_x} - \boldsymbol{\mu_y}$ is sparse and without loss of generality only first s entries are nonzero.

- Suppose that ϵ_{xij} and $\epsilon_{xij}^2 - 1$ satisfy the Cramér's condition, that is there exist constants ν_1, ν_2, M_1 and M_2, such that $E|\epsilon_{xij}|^m \leq m!M_1^{m-2}\nu_1/2$ and $E|\epsilon_{xij}^2 - \sigma_{xj}^2|^m \leq m!M_1^{m-2}\nu_1/2$ for all $m = 1, 2, \ldots$ where σ_{xj} is the diagonal entries of $\boldsymbol{\Sigma_x}$. Assumptions on ϵ_{yij} and $\epsilon_{yij}^2 - 1$ are the same as ϵ_{xij} and $\epsilon_{xij}^2 - 1$ respectively.

- Assume that the diagonal elements of both $\boldsymbol{\Sigma_x}$ and $\boldsymbol{\Sigma_y}$ are bounded away from 0.

Under Condition 1, they have the following theorem:

Theorem 3.1. *Let s be a sequence such that $log(p-s) = o(n^\gamma)$ and $logs = o(n^{1/2-\gamma}\beta^n)$ for some $\beta^n \longrightarrow \infty$ and $0 < \gamma < \frac{1}{3}$. Suppose that $\min_{1<j<s} \frac{|\alpha_j|}{\sqrt{\sigma_{xj}^2 + \sigma_{yj}^2}} = n^{-\gamma}\beta_n$.Then for $t \sim cn^{\gamma/2}$ with c some positive constant, we have*

$$P(\min_{j \leq s} |T_j| \geq t \ and \ \max_{j>s} |T_j| < t) \longrightarrow 1$$

where $n = n_x + n_y$.

Theorem 3.1 indicates that two sample t-statistic can potentially pick all the important variables as long as the rate of decay is not too fast and the sample size is not too small.

In order to demonstrate the performance of two sample t-statistic on the variable screening, they then apply the independence classifier which is mentioned in the Section 2.3 to calculate the misclassification error rate based on the most informative variables selected by two sample t-statistic.

They assume both populations are from Gaussian distributions and common variance matrix is identity, that is $\mathbf{\Sigma}_{\boldsymbol{x}} = \mathbf{\Sigma}_{\boldsymbol{y}} = I$. Rank all the variables according to two sample t-statistic and pick up the most informative m variables assuming that those m variables are all the important variables in terms of classification. The theoretical misclassification error rate calculated by this truncated independence classifier is given in the below theorem:

Theorem 3.2. *Consider $\mathbf{\Sigma}_{\boldsymbol{x}} = \mathbf{\Sigma}_{\boldsymbol{y}} = I$ and use a truncated independence classifier*

$$\widehat{\delta}^{m_n}(\mathbf{z}) = (\mathbf{z}^{m_n} - \widehat{\mu}^{m_n})(\widehat{\mu}_1^{m_n} - \widehat{\mu}_2^{m_n})$$

for a given sequence m_n. Suppose that $\frac{n}{\sqrt{m_n}} \Sigma_{j=1}^{m_n} \alpha_j^2 \longrightarrow \infty$ as $m_n \longrightarrow \infty$. Then the classification error of $\widehat{\delta}^{m_n}$ is

$$= 1 - \Phi\left(\frac{(1 + o_P(1))\Sigma_{j=1}^{m_n} \alpha_j^2 + m_n(n_1 - n_2)/(n_1 n_2)}{2\{(1 + o_P)\Sigma_{j=1}^{m_n} \alpha_j^2 + n m_n/(n_1 n_2)\}^{1/2}} \right)$$

where $\Phi(\cdot)$ is the standard Gaussian distribution function. They call this truncated classifier as feature annealed independence rule (FAIR). We can have the precise value of m by minimizing this theoretical misclassification error rate . In practice, however, this equation is unsolvable and it can only be done numerically. Fan and Fan (2008) used a simulation study and three real data analyses to demonstrate their theoretical results and show the superiority of their method over the nearest shrunken centroid method (Tibshirani *et al*, 2002).

3.2 WMW-Based Feature Annealed Classifier

As defined in Section 2.1, two sample Wilcoxon-Mann-Whitney statistic provides more useful information than two sample t-statistic in terms of group separation under some certain circumstances where the normality assumption is not achievable.

Inspired by Fan and Fan (2008), I expect that using WMW statistic can also pick up all the important variables. If so, then WMW will be used more widely than t-statistic because most gene expression data present heavier tail than normal distribution (Salas-Gonzalez *et al*, 2009).

3.2.1 Variable Selection Based on WMW Statistic

Using the similar strategy given in Fan and Fan (2008), I also prove that theoretically in the infinitely multidimensional data space, Wilcoxon-Mann-Whitney statistic can pick up all the informative variables with probability approaching to 1. The result is given in the following theorem:

Theorem 3.3. *Assume that the vector $\boldsymbol{\alpha} = \boldsymbol{\mu_x} - \boldsymbol{\mu_y}$ 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 $\log s = o(n^{1/2-\gamma}\beta_n)$ for some $\beta_n \to \infty$ and $0 < \gamma < \frac{1}{3}$. For $w \sim cn^{\gamma/2}$ with some constant $c > 0$, we have*

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

Proof. I divide the proof into two parts.

(a) Let us first look at the probability $P(\max_{j>s} |W_j| > w)$. Clearly,

$$P(\max_{j>s} |W_j| > w) \le \sum_{j=s+1}^{p} P(|W_j| \ge w)$$

By the Corollary 3.2 proved in Froda and Eeden (2000), there exist a $\alpha > 0$, such that , for $M_0 < x < \alpha n^{1/6}$,

$$P(|W_j| \ge w) = (1 - \Phi(w))(1 + O(w^3/n^{1/2}))$$

where $M_0 > 1$. For the normal distribution, we have the following tail probability inequality

$$1 - \Phi(x) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{w} e^{-w^2/2}$$

Combining with the symmetry of W_j, if we let $w \sim cn^{\gamma/2}$, then we have

$$\sum_{j=s+1}^{p} P(|W_j| \geq w) \leq (p-s) \frac{2}{\sqrt{2\pi}} \frac{1}{w} e^{-w^2/2} (1 + O(w^3/n^{1/2}) \to 0.$$

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

$$P(\max_{j>s} |W_j| > w) \to 0$$

(b) Next, we consider $P(\min_{j \leq s} |W_j| \leq w)$. Let $\eta_j = \frac{\alpha_j}{\sqrt{n_1 n_2 (n_1 + n_2 - 1)/12}}$ and define $\widetilde{W}_j = W_j - \eta_j$.

Then clone the lines in (a), we have

$$\sum_{j \leq s} P(|\widetilde{W}| \geq w) \leq s \frac{2}{\sqrt{2\pi}} \frac{1}{w} e^{-w^2/2} (1 + O(w^3/n^{1/2}) \to 0$$

Let $\alpha_0 = \min_{j \leq s} \eta_j$. Then it follows that

$$\begin{aligned} P(\min_{j \leq s} |W_j| \leq w) &\leq P(\max_{j \leq s} |\widetilde{W}_j| \geq \min_{j \leq s} |\eta_j| - w) \\ &\leq P(\max_{j \leq s} |\widetilde{W}_j| \geq \alpha_0 - w) \end{aligned}$$

If $w \sim cn^{\gamma/2}$ and $\alpha_0 \sim n^{-\gamma} \beta_n$ for some $\beta_n \to \infty$, then similarly to part (a), we have

$$P(\min_{j \leq s} |W_j| \leq w) \to 0.$$

Combination of Part (a) and (b) completes the proof. □

Compare to Fan and Fan (2008), Theorem 3.3 requires fewer assumptions. For example, there are no assumption on random errors and no assumptions on covariance matrices of two population, which makes WMW statistic more efficiently to identify the minimal subset of variables that succinctly predict the categories of the new observations.

3.2.2 Projection-free WMW-based Classifier

Fan and Fan (2008) used minimal misclassification error rate to explicitly the performance of two sample t-statistic. They named their procedures as Feature Anneal Independence Rule (FAIR), that is, allocate the new observations using independence classifier based on the variables chosen by t-statistic. However independence classifier is strongly related to the t-statistic which is not appropriate where normality assumption isn't held. That is the reason I propose this new nonparametric classifier which is simply based on WMW statistic.

Even though WMW statistic measures the group separation very well, itself can't be used as a classifier directly. Here I borrow the idea of Group-Group Transviation classifier (Nudurupati and Abebe, 2009) to translate the measure of group separation to a discriminant function.

To classify a new observation $\mathbf{z}$, define $\mathbb{X}^* = \mathbb{X} \cup \{\mathbf{z}\}$ and $\mathbb{Y}^* = \mathbb{Y} \cup \{\mathbf{z}\}$. $\mathbf{z} \overset{\in}{\sim} \Pi_{\mathbf{X}}$ if

$$\sum_{k=1}^{p} w_k(\mathbb{X}^*, \mathbb{Y}) > \sum_{k=1}^{p} w_k(\mathbb{X}, \mathbb{Y}^*)$$

where $w_k(\mathbb{X}^*, \mathbb{Y})$ is WMW statistic of variable k based on $\mathbb{X}^*$ and $\mathbb{Y}$ while $w_k(\mathbb{X}, \mathbb{Y}^*)$ is WMW statistic of variable k based on $\mathbb{X}$ and $\mathbb{Y}^*$.

This classifier indicates that I classify this new observation $\mathbf{z}$ into $\Pi_{\mathbf{X}}$ if a better group separation can be achieved by adding this new observation $\mathbf{z}$ into $\Pi_{\mathbf{X}}$, otherwise into $\Pi_{\mathbf{Y}}$.

The advantage of this classifier is that it is robust to deviations from the usual assumptions. The other hand it's projection free classifier such that it can be used to numerically find the minimal misclassification error rate by picking the proper number of informative variables. The combination of WMW statistic variable screening and WMW-based classifier results in the WMW-based feature annealed classifier (WFAC).

3.3 Results

Two real data analyses and a Monte Carlo simulation are provided to demonstrate the superiority of WFAC compared to FAIR.

3.3.1 Lung Cancer

I first use Lung Cancer data to compare the performances of WFAC and FAIR by classifying between two types of lung cancer: malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA). In total, the data set contains 181 sample, with 31 from MPM and 150 from ADCA. The training set contains 32 samples, with 16 from MPM and 16 from ADCA while the testing set contains 149 samples, with 15 from MPM and 134 from ADCA. Each sample is described by 12533 genes. This data is available at http://www.chestsurg.org.

Fan and Fan (2008) set the classification rule by using independence classifier with t-statistic variable screening based on the training data and predict each sample in the testing data to be MPM or ADCA by following this rule.

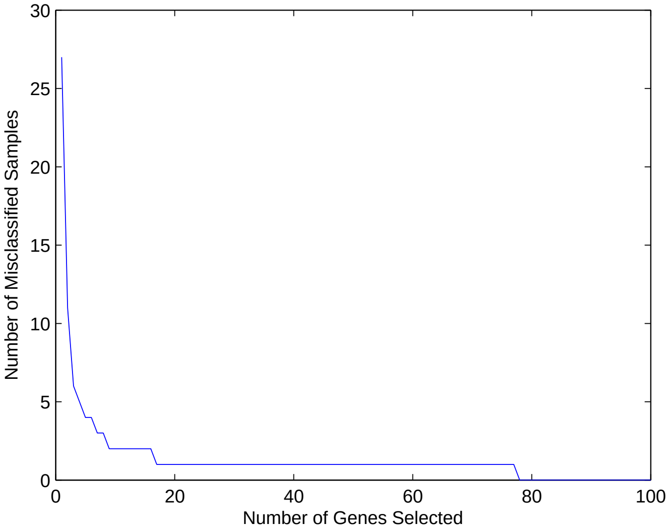
I, instead set the classification rule by using WMW-based classifier with WMW statistic variable screening based on the training data and predict each sample in the testing data to be MPM or ADCA by following this new classifier proposed above.

Numbers of incorrect classification out of 149 testing samples are given in the Table 3.1.

Table 3.1: Lung Cancer data. The minimum number of incorrectly classified samples out of 149 testing data

Method	**Test Error**	**No. of Selected Genes**
FAIR	11/149	26
WFAC	0/149	78

Figure 3.1: Scree Plot of WFAC for Lung Cancer Data



Apparently, FAIR reaches the minimal misclassification error rate (11) by selecting 26 genes while WFAC achieves zero misclassification error by picking 78 genes. In the sense of misclassification error rate, our method is superior to FAIR. However it seems WFAC needs to select much more genes to have this desirable result, which makes our method inefficient. For this reason, a scree plot is drawn to show how the misclassification error rate changes when more and more genes are added.

Plot 3.1 shows that 1 testing sample is misclassified based on top 17 selected genes. As shown in Table 3.1, using FAIR the minimum number of misclassified samples is 11 by selecting 26 variables, which indicates that FAIR use more genes but achieves large number of misclassified samples than WFAC.

Table 3.2: Prostate Cancer data. The minimum number of incorrectly classified samples out of 102 samples using leave-one-out cross validation

Method	**Training Error**	**No. of Selected Genes**
Fan and Fan	10/102	2
Our	5/102	4

3.3.2 Prostate Cancer

I then use Prostate Cancer data, which is available at http://www.broad.mit.edu/cgi-bi/cancer/datasets.cgi. The prostate cancer data contains 102 patient samples, 52 of which are prostate tumor samples and 50 of which are normal prostate samples. Each sample contains 12600 genes. The minimum number of misclassified samples is calculated by FAIR and WFAC respectively using leave-one-out cross validation.

Numbers of incorrect classification out of 102 testing samples are given in the Table 3.2.

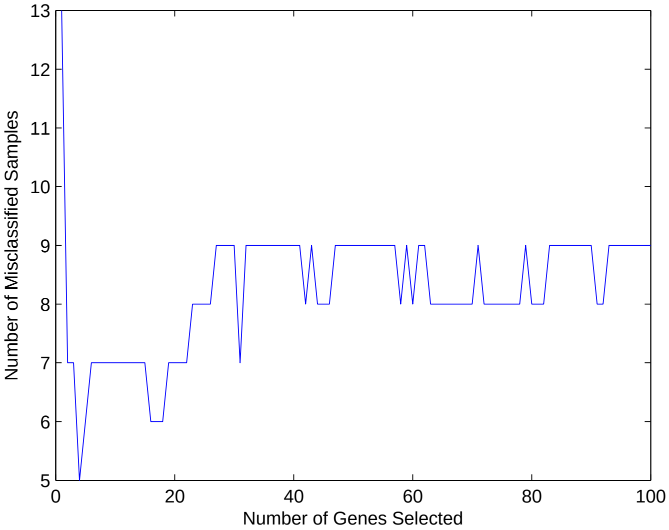
It shows that by selecting top 2 genes, FAIR gives 10 misclassified samples while by selecting top 4 genes, WFAC only gives 5 misclassified samples. It still seems WFAC sacrifices the efficiency to obtain the accuracy. For the same reason, a scree plot is drawn to illustrate the efficiency of WFAC.

As shown in Plot 3.2, WFAC gives 7 misclassified samples if only selecting top 2 genes, which implies that WFAC uses the same number of genes to obtain lower misclassification error rate.

3.3.3 Simulation

I perform a large Monte Carlo simulation to study the optimality (in terms of misclassification error) of FAIR and WAFC under a variety of distributional settings. To that end, I generated two classes of data from normal, Cauchy, and t with two degrees of freedom (t_2) distributions with dimension $p = 200$. I set the center of one class at the origin $(0, 0, \ldots, 0)$ and the center of second class at $(1/4, 1/2, 3/4, 1, 5/4, 3/2, 0, 0, \ldots, 0)$. I considered variance-covariance matrices

Figure 3.2: Scree Plot of WFAC for Prostate Cancer Data



$\Sigma_1 = I_{200}$ and

$$\Sigma_2 = \begin{pmatrix} 1 & -1/2 & -1/2 & \ldots & -1/2 \\ -1/2 & 1 & -1/2 & \ldots & -1/2 \\ -1/2 & -1/2 & 1 & \ldots & -1/2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -1/2 & -1/2 & -1/2 & \ldots & 1 \end{pmatrix}.$$

In the simulation, training samples of sizes 20 and 30 were generated as well as testing samples of size 1000 for each group are generated. Use FAIR and WFAC to set the classification rule based on the training data and calculate the minimum misclassification error rate by computing the proportion of misclassified testing samples in each group respectively. Use the Monte Carlo simulation to generate 50 different training and testing data having the same structure for each distributional setting and apply the same procedure to all those different samples separately. Comparison boxplots containing the misclassification error rates are given in Figure 3.3.

It is clear from the plots that WFAC provides lower misclassification error rates for the heavier tailed distributions (Cauchy, t_2). In particular, for Cauchy data, FAIR leads to misclassification error rates consistently around 50% which is nearly as same as guessing. This is somewhat improved for the t_2 distribution even though WFAC is still better than FAIR. As expected, for normal data FAIR gives better performance than WFAC.

3.4 Smoothed Projection-Free WMW-Based Classifier

3.4.1 Description

As discussed above, WFAC provides an improvement over FAIR in dealing with non-normal distributed data. As shown in (2.2), in the calculation of WFAC, sign function ϕ is used to count the number of transvariated observations. In order to simplify the following discussion, let us assume the median of the differences of all the observations from $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$ respectively is positive. Thus two observations $\mathbf{x}$ and $\mathbf{y}$ from $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$ respectively are treated as transvariation as long as $\mathbf{x} < \mathbf{y}$ regardless of whether $\mathbf{y}$ is barely greater than $\mathbf{x}$ or much greater than $\mathbf{x}$. A weight associated with the magnitude of difference is needed and I will assign such weight s by replacing ϕ with a $[0, 1]$-valued and non decreasing function that is continuously differentiable on an interval $(-\delta, \delta)$ for some $\delta > 0$. Inspired by Abebe and Nudurupati (2011), a continuous cumulative distribution functions defined on $\mathscr{R}$ are applied.

Using smoothed WFAC, I will classify $\mathbf{z}$ to $\Pi_{\mathbf{x}}$ if

$$\sum_{k=1}^{p} w_k(\mathbb{X}^*, \mathbb{Y}) > \sum_{k=1}^{p} w_k(\mathbb{X}, \mathbb{Y}^*)$$

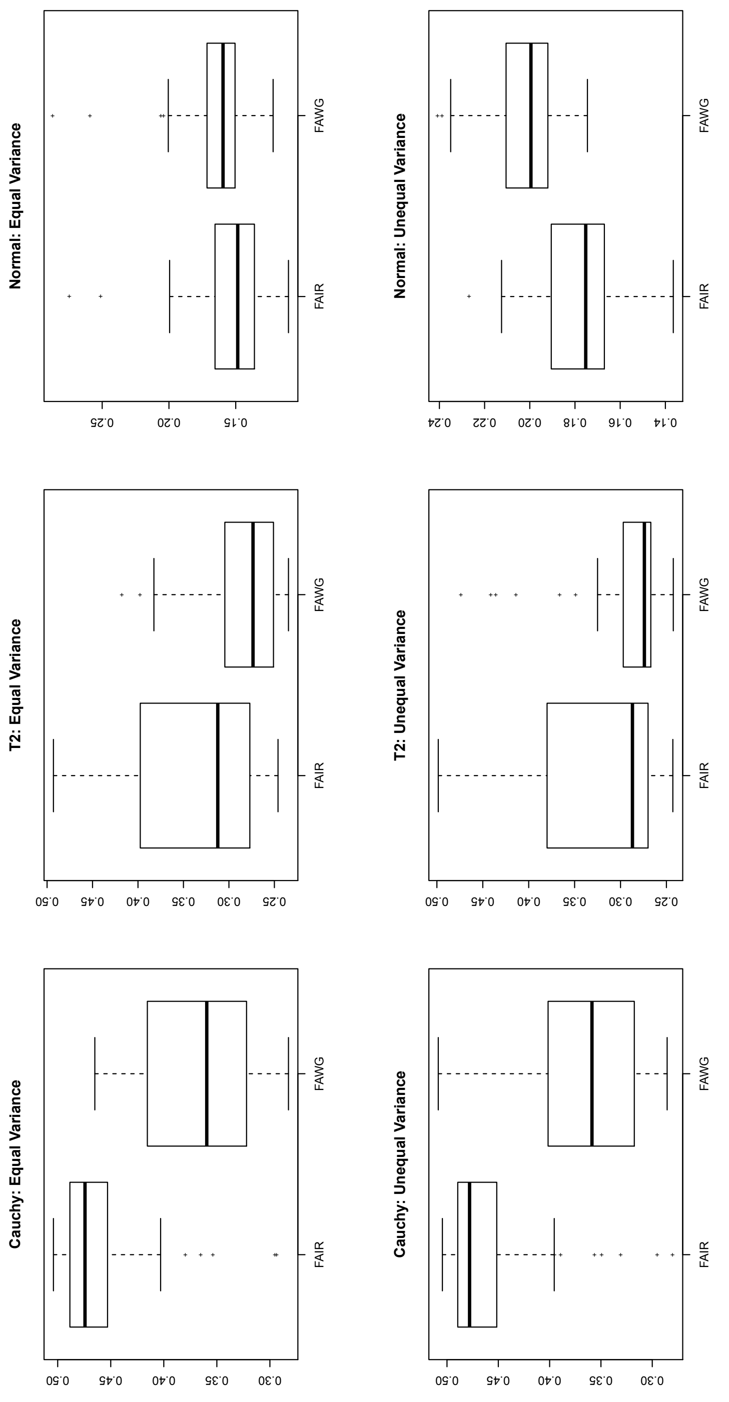
where

$$w_k(\mathbb{X}^*, \mathbb{Y}) = 1 - \frac{2}{n_x n_y} \sum_{\mathbf{x}^* \in \mathbb{X}^*} \sum_{\mathbf{y} \in \mathbb{Y}} \mathbf{K}_{\alpha} \left\{ (x_{ki} - y_{kj})\, m_{xy}^{k} \right\}$$

and

$$w_k(\mathbb{X}, \mathbb{Y}^*) = 1 - \frac{2}{n_x n_y} \sum_{\mathbf{x} \in \mathbb{X}} \sum_{\mathbf{y}^* \in \mathbb{Y}^*} \mathbf{K}_{\alpha} \left\{ (x_{ki} - y_{kj})\, m_{xy}^{k} \right\}$$

Figure 3.3: Misclassification Error Rate



Here $\mathbf{K}_\alpha$ is $[0,1]$-valued functions that are non decreasing on $\mathscr{R}$, where α is smoothing parameter. For example, one could take cumulative density function of $N(0, \alpha)$. Here I use cumulative density function of t-distribution with α degrees of freedom.

For the data with training and testing samples, the training samples are used to find the best smoother for each group. Similarly as mentioned in Abebe and Nudurupati (2011), I use a bivariate grid (α_1, α_2) and apply a leave-one-out cross validation to the training samples to find the misclassification error for each possible pairs of degrees of freedom. The combination with the least misclassification error is then selected as the pair of smoothing constants.

3.4.2 Monte Carlo Simulation

To demonstrate the optimality of smoother, several common simulation setting are used: normal distribution, t distribution (heavy tail distribution), and log-normal distribution (skewed distribution). I set the center of one class at the origin $(0, 0, \ldots, 0)$ and the center of second class at $(1/4, 1/2, 3/4, 1, 5/4, 3/2, 0, 0, \ldots, 0)$. I considered variance-covariance matrices $\Sigma_1 = I_{200}$.

I consider normal, t_2 and log-normal distribution to generate 200-dimensional data. In the simulation, training samples of sizes 30 and 30 as well as testing samples of size 100 for each group were generated by following the distributional settings mentioned above. We use WFAC to set the classification rule based on the training data and calculate the minimum misclassification error rate by computing the proportion of misclassified testing samples in each group respectively. I then calculate the minimum misclassification error rate for the testing sample by including the optimal smoother determined by training data. We use the Monte Carlo simulation to generate 10 different training and testing data having the same structure for each distributional setting and apply the same procedure to all those different samples separately. Comparison boxplots containing the misclassification error rates are given in Figure 3.4.

It shows that with the equal sample sizes, the smoother doesn't improve the performance of WFAC very well.

I then change the training samples sizes to 20 and 30 respectively and proceed the same simulation as described above to show how the smoother behaves for the different sample size case. Comparison boxplots containing the misclassification error rates are given in Figure 3.5.

Figure 3.5 shows that the one with smoother works much better for the skewed data (log-normal) while it is slightly more efficient for the heavy tailed data (t_2).

3.5 Conclusion

Both two real data analysis shows the better efficiency and accuracy of WFAC than FAIR. A large Monte Carlo simulation study further demonstrates the obvious advantage of WFAC for heavier tailed data and slight disadvantage of normally distributed data. Then the necessity of smoothed WFAC under some circumstance where two sample sizes are unequal, is discussed in a Monte Carlo simulation study.

Figure 3.4: Misclassification Error Rate (Smoothed with equal Size)

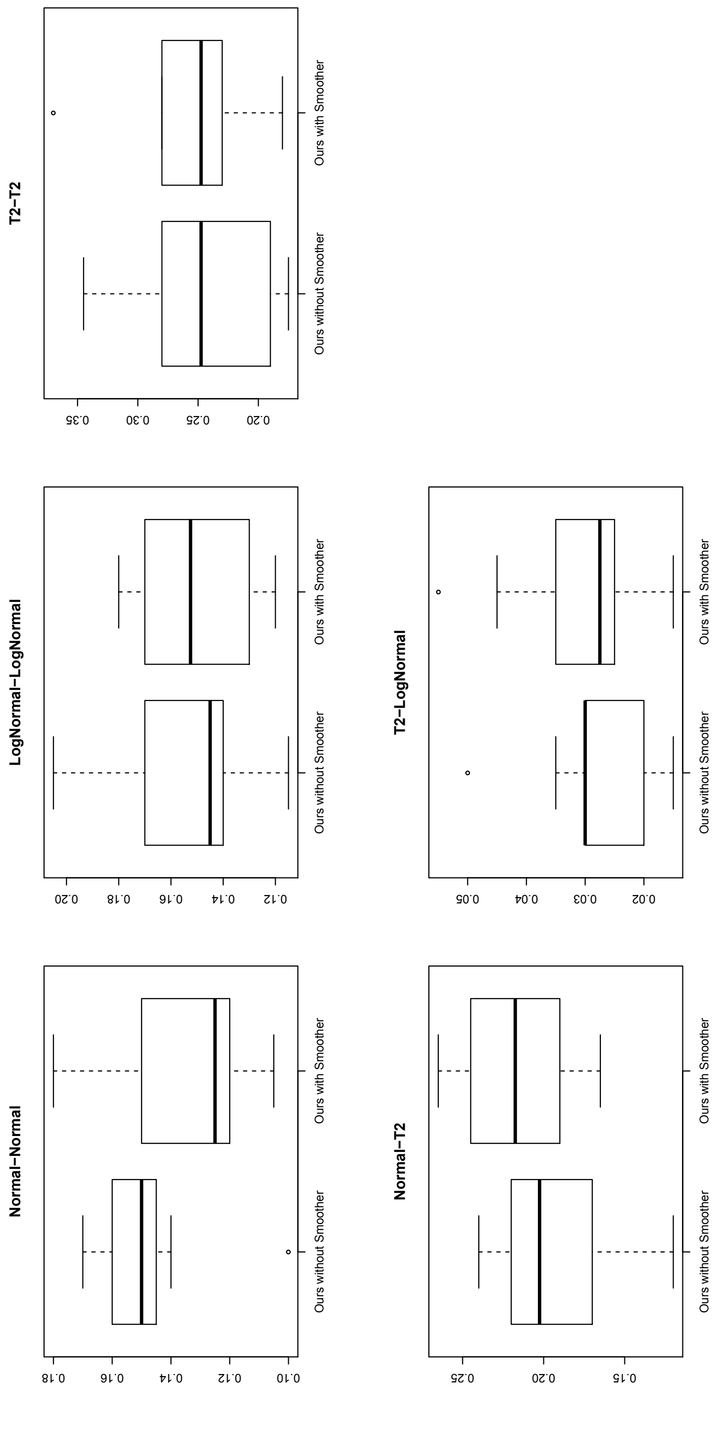
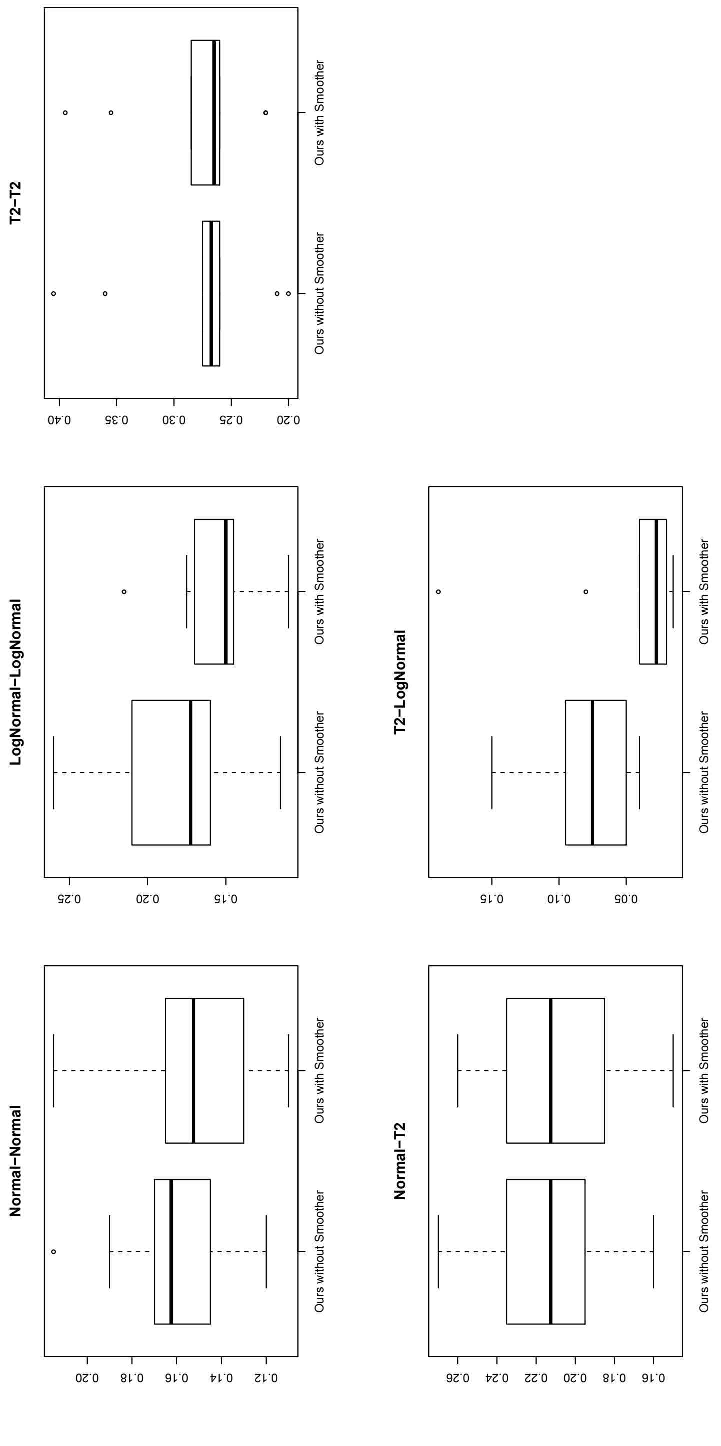


Figure 3.5: Misclassification Error Rate (Smoothed with Unequal Size)



Chapter 4

WMW Forward Variable Selection Method

Two sample t-statistic and WMW statistic can be used as variable screening index and their properties are discussed in Chapter 3. But both variable selection procedures fail to take the correlations among predictor variables into account. They are not recommended when there are many variables that are dependent on each other. As shown in Section 2.2, two popular dimension reduction methods, PCA and PLS can also be used to reduce the dimension by finding a smaller set of uncorrelated linear combinations of the original variables. However, the problem with both methods is that they use the linear combinations to combine all the predictor variables to create new variables. Those created variables contain the information of all predictor variables and they are very hard to be interpreted, especially in biological study. That is the reason I propose this new WMW forward variable selection procedure to improve those existing methods.

4.1 Descriptions

Consider two populations $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$. $\mathbb{X}$ and $\mathbb{Y}$ are two samples from $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$ with sample sizes n_x and n_y respectively. $\mathbf{X} = \{x_{ij}\}$ and $\mathbf{Y} = \{y_{ij}\}$ are the predictor matrices of two samples. Write $\mathbf{V}$ as a matrix of column vectors $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_p]$ where each $\mathbf{v}_i$ is an $n \times 1$ vector, where $n = n_x + n_y$. I would like to order the variables $\mathbf{v}_1, \ldots, \mathbf{v}_p$ in a decreasing order according to the amount of information they provide for class determination, $\mathbf{v}_{[1]} \geq \cdots \geq \mathbf{v}_{[p]}$ say. The most informative variable is the one that gives maximum separation between the two groups. As the second most informative variable, it seems reasonable to pick the variable that is the most dissimilar to $\mathbf{v}_{[1]}$ while at the same time giving the highest contribution to distinguishing between the two groups.

The approach I propose uses the WMW statistics to measure overlap and dissimilarity in the sense of classification. For $k = 1, \ldots, p$, define

$$t_{ij}^k = x_{ki} - y_{kj} \ , \qquad i = 1, \ldots, n_x; j = 1, \ldots, n_y$$

and, as a measure of overlap between $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{Y}}$, consider the two sample WMW statistic based on $\mathbf{v}_k$

$$w(\mathbf{v}_k) = 1 - \frac{2}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \phi(t_{ij}^k) m_k \ , \tag{4.1}$$

where $m_k = \text{median}\left\{t_{ij}^k : 1 \leq i \leq n_x, 1 \leq j \leq n_y\right\}$. It can be seen that $0 \leq w(\mathbf{v}_k) \leq 1$. Higher values of $w(\mathbf{v}_k)$ indicate smaller overlap between $\{x_{k1}, \ldots, x_{kn_x}\}$ and $\{y_{k1}, \ldots, y_{kn_y}\}$. The most informative variable is the one with the the least overlap and hence the highest $w(\cdot)$.

To measure the dissimilarity between two variables $\mathbf{v}_r$ and $\mathbf{v}_s$ I use

$$d(\mathbf{v}_r, \mathbf{v}_s) = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \phi\left\{\left(t_{ij}^r m_r\right)\left(t_{ij}^s m_s\right)\right\} \ , \tag{4.2}$$

where $r, s = 1, \ldots, p$. This quantity $d(\mathbf{v}_r, \mathbf{v}_s)$ resembles the measure of dissimilarity studied by Sokal and Michener (1958) and Rand (1971) (see discussion in Albatineh *et al*, 2006) that is given by $(n_x n_y)^{-1} \sum \sum t_{ij}^r t_{ij}^s$. The measure $d(\mathbf{v}_r, \mathbf{v}_s)$ counts how often observations i and j in $\mathbf{v}_r$ and $\mathbf{v}_s$ behave in an opposite direction with respect to their medians for $i = 1, \ldots, n_x;\ j = 1, \ldots, n_y$. It is clear that $0 \leq d(\mathbf{v}_r, \mathbf{v}_s) \leq 1$ where large values of $d(\mathbf{v}_r, \mathbf{v}_s)$ indicate large dissimilarity between $\mathbf{v}_s$ and $\mathbf{v}_r$. It is also easily observed that $d(\mathbf{v}, \mathbf{v}) = 0 = d(\mathbf{v}, -\mathbf{v})$. Moreover, $d(\mathbf{v}_r, \mathbf{v}_s) = 1$ if $t_{ij}^r m_r < 0$ whenever $t_{ij}^s m_s > 0$, for all i and j, and vice versa. In such cases, variables $\mathbf{v}_r$ and $\mathbf{v}_s$

are totally dissimilar in the sense that they provide opposing information about class membership.

4.2 Algorithm

An algorithm to implement WFS is as follows:

Algorithm 4.1.

Step 1: Let

$$\mathbf{v}_{[1]} = \operatorname*{Argmax}_{k=1,\ldots,p} \; w(\mathbf{v}_k) \;,$$

where w is defined in (4.1)*.*

Step 2: Use (4.2) *to compute $d(\mathbf{v}_s, \mathbf{v}_{[1]})$ for $s \neq [1]$. Let*

$$\mathbf{v}_{[2]} = \operatorname*{Argmax}_{\substack{k=1,\ldots,p \\ k\neq[1]}} \{w(\mathbf{v}_k) d(\mathbf{v}_k, \mathbf{v}_{[1]})\} \;.$$

Step 3: For $c = 2, \ldots, p$, find direction of maximum separation $\hat{\mathbf{u}}_1 \in \mathbb{R}^c$ given in (2.4) *using $\mathbf{v}_{[1]}, \ldots, \mathbf{v}_{[c]}$ and set*

$$\mathbf{v}_{[c+1]} = \operatorname*{Argmax}_{\substack{k=1,\ldots,p \\ k\notin\{[1],\ldots,[c]\}}} \; w(\mathbf{v}_k) d(\mathbf{v}_k, \hat{\mathbf{u}}_1'[\mathbf{v}_{[1]} \cdots \mathbf{v}_{[c]}]) \;.$$

One may use stability of the misclassification error rate as a stopping criterion. The downside of Algorithm 4.1 is that one needs to search higher and higher dimensional spaces as the value of c in Step 3 increases. This introduces a huge computational burden. The following modification avoids high dimensional projections by combining selected variables using the direction of maximum separation:

Algorithm 4.2.

Step 1: Set $c = 1$. *Let*

$$\mathbf{v}_{[1]} = \underset{k=1,\ldots,p}{\text{Argmax}} \; w(\mathbf{v}_k) \; ,$$

where w *is defined in* (4.1).

Step 2: If $c \geq p$, *then STOP. Otherwise use* (4.2) *to compute* $d(\mathbf{v}_s, \mathbf{v}_{[1]})$ *for* $s \in \{[c+1], \ldots, [p]\}$. *Let*

$$\mathbf{v}_{[2]} = \underset{k \in \{[c+1],\ldots,[p]\}}{\text{Argmax}} \; \{w(\mathbf{v}_k) d(\mathbf{v}_k, \mathbf{v}_{[1]})\} \; .$$

Step 3: Find direction of maximum separation $\hat{\mathbf{u}}_1 \in \mathbb{R}^2$ *given in equation* (2.4) *using* $[\mathbf{v}_{[1]} \; \mathbf{v}_{[2]}]$ *and set*

$$\mathbf{v}_{[1]} \leftarrow \hat{\mathbf{u}}_1' [\mathbf{v}_{[1]} \; \mathbf{v}_{[2]}]$$

$$c \leftarrow c + 1$$

Go back to **Step 2**.

This algorithm is convenient for selecting variables in high dimensional data since it only requires two dimensional projections. This is especially useful when performing classification based on gene expression data. Besides, forward selection allows one to start with fewer variables and proceed to higher dimensions if necessary. As a stopping rule, one may use predetermined dimensions (Nguyen and Rocke, 2002) or cross validation using the misclassification error rate.

4.3 Results

One real data analysis demonstrates the advantage of WFS over both simple t-statistic and WMW statistic variable screening. Then two real data analyses and a Monte Carlo simulation are used to compare the performances among PCA, PLS and WFS.

4.3.1 Caribbean Food Data

Caribbean food data is applied to compare the performance among t-statistic, WMW statistic and WFS. This data set contains information from Food and Drug Administration (FDA) and U.S. Department of Agriculture (USDA) on the number of rejections by country for certain Latin American and Caribbean (LAC) countries for the years 1992 to 2003. This data set was investigated in Jolly *et al.* (2007) using zero-inflated count data mixed models. The variables considered in the current study are

- t = year (1992 - 2003)
- FDI = Foreign direct investment, net inflows (Balance of Payments (BoP), current US $)
- Fertcons = Fertilizer consumption (metric tons)
- USImp = U.S. Imports by Country, (1985-03; Millions of Dollars)
- AgImp = Total Agricultural Import to the US (million $)
- GNI = Gross national income per capita, Atlas method (current US $)
- Y = Detention Status (Y=0 no detention; Y=1 detention)

Consider variable Y as response variable and the other six variables as predictor variables. I first select top 2 variables based on t-statistic, WMW statistic and WFS.

Table 4.1: Caribbean food data. Misclassification error rates using leave-one-out cross validation.

	t Selection	WMW Selection	WFS	Without Selection
LDA	0.4155	0.3873	0.3873	0.3803
MaxD	0.3873	0.3170	0.3169	0.3170
PGT	0.3944	0.3099	0.3028	0.3592
GGT	0.3803	0.3099	0.3028	0.3451
TD	0.3732	0.3310	0.3310	0.3310

Variables USImp and AgImp are selected by t-statistic while WMW statistic chooses variables FDI and USImp. WFS, instead, picks the variables FDI and Fertcons. Then classifiers LDA, TD, MaxD, PGT, and GGT discussed in Section 2.3 are applied to calculate the proportion of the misclassified samples by using leave-one-out cross validation based on the top 2 variables selected by t-statistic, WMW statistic and WFS respectively. I also calculate the corresponding misclassification error rate based on all those six predictor variables to show the necessity of variable selection procedure. Misclassification error rates are given in the Table 4.1

It is clear that using t-statistic to select the variables gets the worse classification than without any variable selection, which indicates that it fails to identify the most informative variables. The results can be improved when I apply PGT and GGT to determine the class membership based on the important variables (FDI and Fertcons) selected by WMW. Finally WFS followed by PGT and GGT gives the minimal misclassification error rate (0.3028). In the sense of classification, variables FDI and Fertcons chosed by WFS should be considered as the most informative ones in prediction of food detention.

4.3.2 Colon Data

A two way cluster study is conducted by Alon *et al.* (1999) using a data set composed of 40 colon tumor samples and 22 normal colon tissue samples, analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes.

Table 4.2: Colon data. The number of incorrectly classified samples out of 62 samples using leave-one-out cross validation. Screening for PCA and PLS uses WMW and t statistics. The numbers in parentheses are the number of genes kept after the screening.

	WMW Selection				t Selection				
	PCA	PLS	PCA	PLS	PCA	PLS	PCA	PLS	WFS
	(50)	(50)	(100)	(100)	(50)	(50)	(100)	(100)	
LDA	10	8	8	8	9	8	8	7	5
QDA	10	8	10	7	10	8	9	8	11
MaxD	9	12	10	11	10	12	12	11	10
PGT	8	8	8	8	8	8	8	8	5
GGT	9	9	9	9	9	10	9	9	5
TD	8	8	8	8	8	8	8	9	7

Alon *et al.* (1999) used an algorithm based on deterministic-annealing algorithm to cluster the data set into two clusters. One cluster consisted of 35 tumor and 3 normal samples while the other cluster contained 19 normal and 5 tumor samples.

A leave-one-out cross validation is used to determine the misclassification error rates based on 4 genes selected by WFS and 4 gene components selected by PCA and PLS. Prior to using PCA and PLS, the top 50 and 100 genes were selected based on the values of WMW and t statistics. Numbers of incorrect classification out of 62 samples are given in the Table 4.2.

The results show that WFS followed by PGT, LDA, or GGT gives the fewest (5) misclassified samples of any combination of dimension reduction/selection and classifier. MaxD results in between 9 and 12 misclassified samples. A misclassification of 12 samples is the highest in the study. The fewest number of misclassified samples by any classifier following PCA is 8 samples. The minimum number of samples misclassified following PLS is 7 samples. This is achieved by LDA when 100 genes were selected by the t-statistic and QDA when 100 genes were selected by the WMW statistic.

Table 4.3: Leukemia training data. The number of incorrectly classified samples out of 38 training samples using leave-one-out cross validation. Screening for PCA and PLS uses WMW and t statistics. The numbers in parentheses are the number of genes kept after the screening.

	WMW Selection				t Selection				
	PCA	PLS	PCA	PLS	PCA	PLS	PCA	PLS	WFS
	(50)	(50)	(100)	(100)	(50)	(50)	(100)	(100)	
LDA	1	1	1	2	1	2	1	2	4
QDA	2	2	2	0	3	2	2	3	5
MaxD	4	3	3	3	4	4	4	4	1
PGT	1	2	1	1	2	3	2	3	0
GGT	1	2	1	1	2	4	2	4	0
TD	1	2	2	1	1	1	2	2	3

4.3.3 Leukemia Data

Golub *et al* (1999) used a classification procedure to discover the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The original data set (training) used consisted of 38 bone marrow samples (27 ALL and 11 AML) obtained from acute leukemia patients at the time of diagnosis. The independent (testing) data set consisted of 24 bone marrow samples as well as 10 peripheral blood specimens from adults and children (20 ALL and 14 AML).

I first use the training data to apply a leave-one-out cross validation as described earlier. Numbers of incorrect classification out of 38 samples based on WFS as well as PCA and PLS based on four genes or gene components were calculated. The results are shown in the Table 4.3.

The results show that WFS followed by PGT or GGT gave no misclassified samples. The same result is attained by QDA using 100 genes selected by the t statistic followed by PLS. The minimum number of samples misclassified following PCA is 1.

I then calculated the number of misclassified samples out of testing samples by using the variables selected by WFS and top 4 principal components obtained by PCA and PLS based on training samples. The results are shown in the Table 4.4.

Table 4.4: **Leukemia training and testing data.** The number of incorrectly classified samples out of 24 testing samples based on training samples. Screening for PCA and PLS uses WMW and t statistics. The numbers in parentheses are the number of genes kept after the screening.

	WMW Selection				t Selection				
	PCA (50)	PLS (50)	PCA (100)	PLS (100)	PCA (50)	PLS (50)	PCA (100)	PLS (100)	WFS
LDA	1	3	1	1	1	3	2	3	5
QDA	2	2	3	2	2	2	1	2	3
MaxD	5	3	2	2	3	2	3	3	3
PGT	2	3	1	1	2	2	1	2	3
GGT	2	3	1	1	2	2	1	3	3
TD	2	1	1	1	2	2	2	3	3

We note that WFS gives inferior performance to PCA and PLS for this experiment involving training and testing data.

The results of the analyses on real data demonstrate the need to characterize cases where WFS performs better than PCA and PLS and vice versa. This is investigated in the following section using simulated data.

4.3.4 Monte Carlo Study

I perform a Monte Carlo simulation to study the optimality (in terms of misclassification error) of PCA, PLS and WFS under a variety of distributional settings. Two classes of data are generated from normal, Cauchy, and t with two degrees of freedom (t_2) distributions with dimension $p = 200$. Set the center of one class at the origin $(0, 0, \ldots, 0)$ and the center of second class at $(1/4, 1/2, 3/4, 1, 5/4, 3/2, 0, 0, \ldots, 0)$. I then consider variance-covariance matrices $\Sigma_1 = I_{200}$ and

$$\Sigma_2 = \begin{pmatrix} 1 & -1/2 & -1/2 & \ldots & -1/2 \\ -1/2 & 1 & -1/2 & \ldots & -1/2 \\ -1/2 & -1/2 & 1 & \ldots & -1/2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -1/2 & -1/2 & -1/2 & \ldots & 1 \end{pmatrix}.$$

In the simulation, training samples of sizes 20 and 30 were generated. After the initial screening of 50 variables using WMW and t-statistics, the samples were used to determine the PCA and PLS loadings and set the classification rules based on the top four components. Testing samples of size 1000 from each group were then generated and the loadings found from the training samples are applied. The misclassification error rate is calculated based on the top four components by computing the proportion of misclassified testing sample observations in each group. For WFS, I directly selected the top 4 variables without any screening. These same variables were retained for the testing samples. The entire process is replicated 50 times.

For the sake of brevity, I only report the results of QDA, MaxD, and GGT. The performance of LDA was similar to QDA and that of PGT and TD was similar to GGT. Comparison boxplots containing the misclassification error rates are given in Figure 4.1.

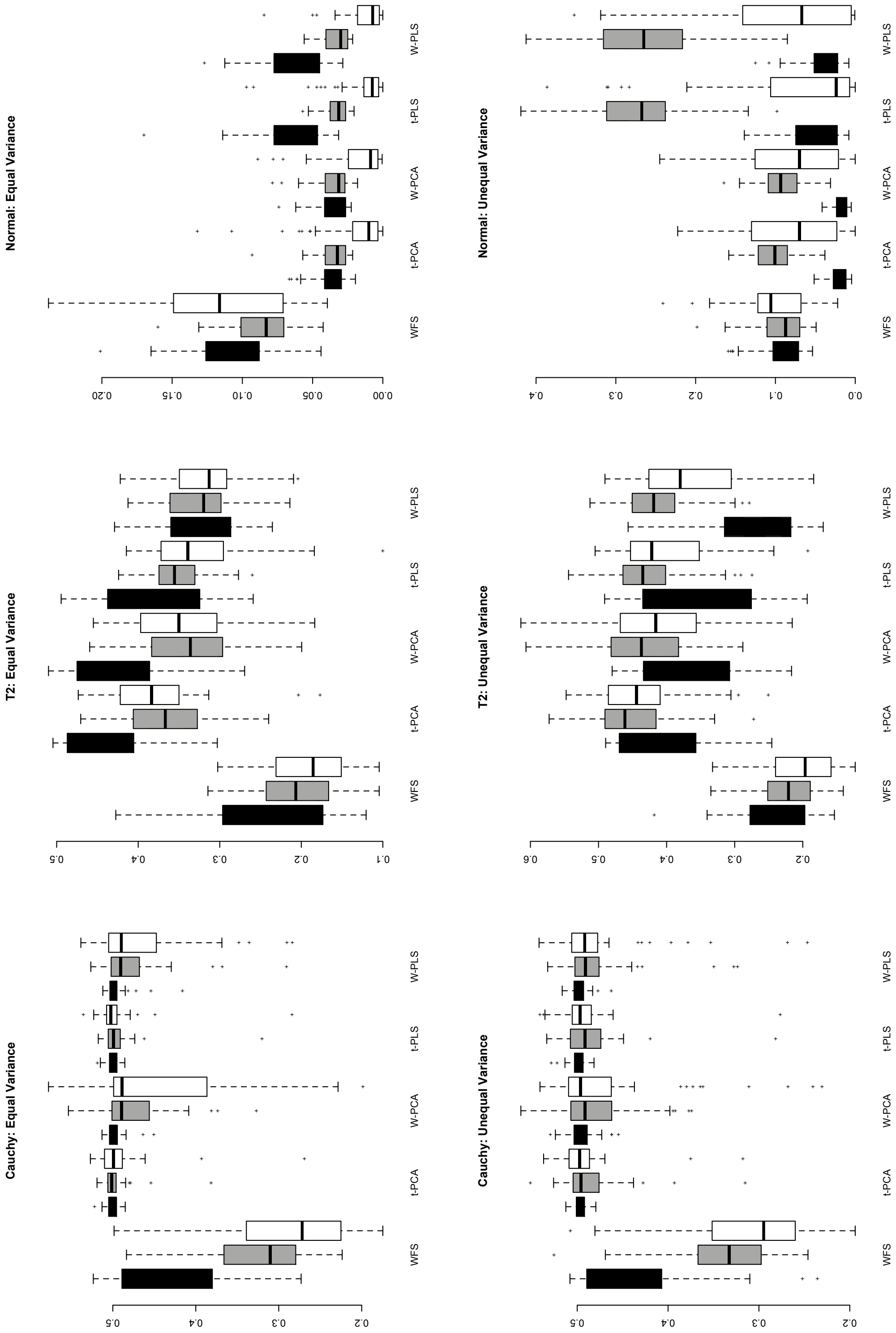
It is clear from the plots that WFS provides lower misclassification error rates for the heavier tailed distributions (Cauchy, t_2). For Cauchy data, PCA and PLS lead to misclassification error rates consistently around 50%. This is akin to flipping a coin to decide group membership without regard to the information contained in the variables. This is somewhat improved for the t_2 distribution even though WFS is still the best among the methods considered. As expected, for normal data PCA and PLS provide better performance than WFS. In the homoscedastic normal case, GGT is the classifier with the lowest misclassification error rate while for the heteroscedastic normal case the best method is QDA. The latter is expected under normality.

4.4 Conclusion

Using two real data, it is shown that transvariation-based classifiers following the rank-based forward variable selection procedure provide better class prediction than LDA and QDA following dimension reduction using PCA and PLS. The forward selection procedure also provides superior performance when the data come from heavy-tailed distributions.

Because it starts with low dimensions, the use of forward selection makes intuitive sense for variable selection in very high dimensional data. Even then the original formulation of the proposed forward selection procedure required projection pursuit in high dimensional spaces. This becomes computationally very expensive especially for gene expression data that are ultra-high dimensional. Complicated methods of mesh-generation and a large number of points are required to effectively cover high dimensional spaces. In this paper, an alternative algorithm that sequentially combines information in two variables using the most informative direction is given as a way to optimize the computation. This modified algorithm only requires projections in two dimensions which can be done by picking evenly spaced points on the unit circle.

Figure 4.1: Misclassification error rates (*Black=QDA, Gray=MaxD, White=GGT*)



Chapter 5

Conclusion and Future Work

Gene expression data usually contains a large number of genes, but a small number of samples. It is well known that not all these genes contribute to determining a specific genetic trait. Feature selection for gene expression data aims at finding a set of genes that best discriminate biological samples of different types.

In this dissertation, inspired by FAIR of Fan and Fan (2008), a new nonparametric classifier (WFAC) is proposed to classify new observations based on the most informative variables selected by Wilcoxon-Mann-Whitney statistic. Its similarity to and differences with FAIR are discussed theoretically and using real data analysis and a Monte Carlo simulation study. I also introduced a smoothed version of WFAC to improve its performance when there is a large sample size discrepancy in the two samples. I then developed a nonparametric forward selection procedure for selecting features to be used for classification. This rank-based forward selection procedure rewards genes for their contribution towards determining the trait but penalizes them for their similarity to genes that are already selected. Lower misclassification error rates are achieved by WFS compared to the dimension reduction methods such as PCA and PLS.

It is of interest to find a specific rule to determine the number of variables I need to select by using WFS. This requires a theoretical description of the misclassification error rate which can then be minimized with respect to the number of variables. So far there is no clear stopping rule and I may only use the predetermined dimensions or cross validation that uses the misclassification error rate. This is currently being studied by the author.

Bibliography

Abebe,A. and Mckean, J. W., (2007) Highly efficient nonlinear regression based on the Wilcoxon norm,

Abebe,A. and Nudurupati, S. V. (2011) Smooth Nonparametric Allocation of Classification, *J. Stat. Comp. Simul.*, 40: 1-16

Abebe, A., McKean, J.W., and Kloke, J.D., (2011) Iterated reweighted rank-based estimates of GEE models, summitted

Albatineh, A. N., Niewiadomska-Bugaj, M. and Mihalko, D. (2006). On similarity indices and correction for chance agreement. *J. Class*, 23:301–313.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999) Broad patterns of gen expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl Acad. Sci. USA*, 96, 6745-6750.

Benard, A. and Elteren, P. V. (1953) A gereralization of method of m rankings, *Nederl. Akad. Wetensch. Proc.* (Indag, Math. 15), Ser. A, 56, 358-369.

Bickel, P.L. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes," and some alternatives when there ae many more variables than observations. *Bernoulli* 10 989-1010. MR2108040

Chen, Z. Y. and Muirhead, R. J. (1994) Comparison of robust linear discriminant procedures using projection pursuit methods, *In Multivariate analysis and its applications (Hong Kong, 1992)* volume 24 of *IMS Lecture Notes Monogr. Ser.*, pages 163-176. Inst. Math. Stat., Hayward, CA.

Crimin, K., McKean, J. W., and Sheather, S. J. (2007) Discriminant procedures based on efficinet robust discriminant coordiantes, *J. Nonpara. Stat.* , 19(45): 199-213.

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

Donoho, D. (1982) Breakdown properties of multivariate location estimators, PhD Qualifying paper, Department of Statistics, Harvard University.

Dudoit, S., Fridly, J. and Speed, T. P. (2002) Classification efficiencies for robust linear discriminant analysis, *Stat. Sinica*, 18(2):581-599.

Durbin, J. (1951) Incomplete Blocks in Ranking Experiments, *Br. J. Psychol.* (Statistical Section), 4, 85-90.

Fan, J. and Fan, Y. (2008) High-dimensional classification using features annealed independence rules. *Ann. Stat.* 36, 2605–2637.

Fan, J. and Lv, J. (2008) Sure inpendence screening for ultrahigh dimensional feature space, *J. R. Stat.* 70, Part 5, pp. 849-911

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems, *Ann. Eugenics*, VII(II):179-188.

Friedman, M. (1940) A Comparison of Alternative Tests of Significance for the Problem of m Rankings, *Ann. Math. Stat*, 11 (1): 86-92.

Friedman, J. H. and Tukey, J. W. (1974) Projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on Computers*, C 23(9):881-890.

Froda, S. and Eeden, C. V. (2000) A uniform saddlepoint expansion for the null-distribution of the Wilcoxon-Mann-Whiteney statistic *J. Stat, Canadian*, Vol. 28, No 1. 137-149

Garthwaite, P.H. (1994) A interprtation of partial least squares, *Am. Stat. Assoc.*, 89, 122-127

Ghosh, A. K. and Chaudhuri, P. (2005) On maximum depth and related classifiers. Scand, *J. Stat.*, 32(2):327-350.

Gini, C. (1916) Il concetto di transvariazione e le sue prime applicazioni. *Giornale degli economisti Rivista di statistica.*

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Caasenbeek, M., Mesirov, P., Coller, H., Loh, M. I., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999) Molecular classification of caver: calss discovery and class prediction by gene expression monitoring, *Science*, 286, 531-537.

Helland, I.S. (1988) On the structure of partial least squares, *Commun. Stat. Simul. Comput.*, 17, 581-607.

Höskuldsson, A. (1988) PLS regression methods, *J. Chem.*, 2, 211-228.

Hotelling, M. and Pabst, M. R. (1936) Rank correlation and tests of significance involving no assumption of nomality. *Ann. Math. Stat*, 7:29-43.

Hugg, J., Rafalin, R., Seyboth, K., and Souvaine, D. (2006) An experimental study of old and new depth measures, Workshop on Alforithm Engineering and Experiments (ALENEX06), Lecture notes in Computer Sceience. New York: Springer-Verlag, pp: 51-64.

Jolliffe, I. T. (1986) *Principal component analysis.* Springer, New York.

Jolly, C. M., Namugabo, E., and Abebe, A. (2007). Food safety issues between latin american and caribbean countries. *Paper presented at the 27th West Indies Agricultural Economics Conference.*

Lehmann, E. L. (2006) *Nonparametrics. Statistical methods based on ranks.* Revised first edition. Springer, New York.

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.

Liu, R. Y. (1990) On a notion of data depth based on random simplices, *Ann. Stat.*, 18(1):405-414.

Liu, R. Y. (1992) Data depth and multivariate rank tests, L_1*-statistical analysis and related methods (Neuchâtel, 1992)*, 79-294, North-Holland, Amsterdam.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference, *Ann. Stat.*, 27(3):783-858, With discussion and a rejoinder by Liu and Singh.

Liu, R. Y. and Singh, K. (1993) A quality index based on data depth and multivariate rank tests, *J. Amer. Stat. Assoc.*, 88(421):252-260.

Mahalanobis, P. C. (1936) On the generalized distance in statistics, *Proc. Natl Acad. Sci. India*, 49-55.

Mann, H. B.; Whitney, D. R. (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". *Annals of Mathematical Statistics* 18 (1): 5060.

Martens, H. and Taes, T. (1989) *Multivariate Calibration.* Wiley, NewYork.

Massey, W.F. (1965) Principal component regression in exploratory statistical Research,*J. AM. Stat. Assoc.*, 60, 234-246.

Montanari, A. (2004) Linear discriminant analysis and transvariation *J. Classi.*, 21(1):71-88.

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

Nudurupati, S. V. and Abebe, A. (2009) A nonparametric allocation scheme for classification based on transvariation probabilities, *J. Stat. Comp. Simul.*, 79(8):977-987.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Stat. Assoc.*, 66(336):846–850.

Salas-Gonzalez, D., Kuruoglu, E. E., Ruiz, D. P. (2009) A heavy-tailed empirical Bayes method for replicated microarray data, *J. Comput. Stat. Data. Anal*, 53 (2009) 1535-1546

Sievers, G. L., and Abebe, A. (2004) Rank estimation of regression coeffcients using iterated reweighted least squares, *J. Stat. Comp. Simul.*, 74, 821-831

Singh, K. (1991) A notion of majority depth, *Technical Report*, Department of Statistics, Rutgers University.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409-1438.

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.

Tukey, J. (1974). *Address to international congress of mathematicians*, Vancouver.

Tyler, D. E. (1987) A distribution-free M-estimator of multivariante scatter, *Ann. Stat.*, 15(1):234–251.

Vardi, Y. and Zhang, C. H. (2000) The multivariate L1-median and associated data depth, *Proc. Natl Acad. Sci. USA*. 97, 1423C1426.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80-83.