ABSTRACT

Selected Topics in High-Dimensional Statistical Learning

John A. Ramey II, Ph.D.

Chairperson: Dean M. Young, Ph.D.

Advances in microarray technology have equipped researchers to measure gene expression levels simultaneously from thousands of genes, yielding increasingly large and complex data sets. However, due to the cost and time required to obtain individual observations, the sample sizes of the resulting data sets are often much smaller than the number of gene expressions measured. Hence, due to the curse of dimensionality [Bellman, 1961], the analysis of these data sets with classic multivariate statistical methods is challenging and, at times, impossible. Consequently, numerous supervised and unsupervised learning methods have been proposed to improve upon classic methods.

In Chapter 2 we formulate a clustering stability evaluation method based on decision-theoretic principles to assess the quality of clusters proposed by a clustering algorithm used to identify subtypes of cancer for diagnosis. We demonstrate that our proposed clustering-evaluation method is better suited to comparing clustering algorithms and to providing superior interpretability compared to the figure of merit (FOM) method from Yeung, Haynor, and Ruzzo [2001] and the cluster stability evaluation method from Hennig [2007] using three artificial data sets and a well-known microarray data set from Khan et al. [2001].

In Chapter 3 we investigate model selection of the regularized discriminant analysis (RDA) classifier proposed by Friedman [1989]. Using four small-sample, high-dimensional data sets, we compare the classification performance of RDA models selected with five conditional error-rate estimators to models selected with the leave-one-out (LOO) error-rate estimator, which has been recommended for RDA model selection by Friedman [1989]. We recommend the 10-fold cross-validation (CV) estimator and the bootstrap CV estimator from Fu, Carroll, and Wang [2005] for model selection with the RDA classifier.

In Chapters 4 and 5 we consider the diagonal linear discriminant analysis ($DLDA$) classifier, the shrinkage-based $DLDA$ ($SDLDA$) classifier from Pang, Tong, and Zhao [2009], and the shrinkage-mean-based $DLDA$ ($SmDLDA$) classifier from Tong, Chen, and Zhao [2012]. We propose four alternative classifiers and demonstrate that they are often superior to the diagonal classifiers using six well-known microarray data sets because they preserve off-diagonal classificatory information by nearly simultaneously diagonalizing the sample covariance matrix of each class.

Selected Topics in High-Dimensional Statistical Learning

by

John A. Ramey II, B.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

_____

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____

Dean M. Young, Ph.D., Chairperson

_____

Greg J. Hamerly, Ph.D.

_____

Dennis A. Johnston, Ph.D.

_____

James D. Stamey, Ph.D.

_____

Jack D. Tubbs, Ph.D.

Accepted by the Graduate School
August 2012

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

Copyright © 2012 by John A. Ramey II
All rights reserved

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to Dr. Dean Young, who has guided me and believed in me. This would not have been possible without you. Thank you to Mrs. Joy Young, who has contributed much to my improvement as a writer and to the overall writing quality of this dissertation. Thank you to the entire statistics department at Baylor University for maintaining an open-door policy. Seeing you hard at work motivated me to do the same. Your enthusiasm for statistics ignited my own. Thank you to my mentor and friend, Dr. Landon Sego, who has greatly contributed to my transition from student to researcher. I will always appreciate the advice and the wisdom that you have shared with me, both in honing my technical skills and in improving as a person. Thank you to Drs. Ryan Hafen and Luke Gosink for our helpful and productive conversations. Thank you to Anthony, Bonnie, Phil, Johnny, and the rest of my friends and family who kept me laughing these last four years.

DEDICATION

To my wife, Megan, and my son, A.J., for your support and your patience while you waited those extra five minutes every day so that I could write one more line of code or apply one more edit at the coffee shop

CHAPTER ONE

Introduction

Advances in microarray technology have equipped researchers to measure gene expression levels simultaneously from thousands of genes, yielding increasingly large and complex data sets. The diagnosis of such diseases as cancer has become heavily reliant on gene expression microarray data sets. However, due to the cost and time required to obtain individual observations, the sample sizes of the resulting data sets are often much smaller than the number of gene expressions measured. Hence, due to the curse of dimensionality [Bellman, 1961], the analysis of these data sets with classic multivariate analysis methods is challenging and, at times, impossible because the methods require that a large number of parameters be estimated. Furthermore, for small-sample, high-dimensional data sets, the statistics utilized with classic multivariate analysis methods often exhibit large variability and are frequently incalculable without a regularization or dimension-reduction technique. Therefore, numerous alternative supervised and unsupervised learning methods have been proposed to improve upon classic methods and ultimately to automate the disease diagnostics.

In Chapter 2 we discuss the evaluation of clustering algorithms that are often used to identify subtypes of cancer for diagnosis. As an initial exploratory step, one often employs unsupervised learning techniques, in particular clustering methods, to provide indirect evidence of functional relationships among genes. Additionally, the application of clustering algorithms facilitates insightful discovery, such as uncovering prognostic subclasses or tumor subtypes of cancer [McLachlan, Do, and Ambroise, 2004]. Consequently, the identification of useful and accurate group structures is essential. However, the numerous clustering algorithms available

to the researcher are not immune to discovering coincidental relationships among genes. Thus, the assessment of a clustering algorithm performed on a given data set is essential to reinforce the validity of the proposed clusters and to provide reasonable doubt regarding anomalous clusters. Here, we present a clustering stability evaluation method based on decision-theoretic principles from Fisher and Van Ness [1971] to assess the quality of the discovered clusters and to identify fallacious clusters. Using three artificial data configurations and the well-known microarray data set from Khan et al. [2001], we demonstrate that our proposed clustering-evaluation method is better suited to comparing clustering algorithms and to providing superior interpretability compared to the well-known figure of merit (FOM) method from Yeung, Haynor, and Ruzzo [2001]. Also, we show that our method yields similar results to the cluster stability evaluation method from Hennig [2007] but with reduced variability because we avoid the *ad hoc* cluster matching that is employed with Hennig's method to overcome the label switching problem.

In Chapter 3 we consider the regularized discriminant analysis (RDA) classifier, proposed by Friedman [1989], that is a widely used supervised-learning method with two tuning parameters that are typically selected by the minimization of an empirical loss function, such as a conditional error rate estimator, over a grid of possible parameter values. Friedman [1989] has suggested that the leave-one-out (LOO) cross-validation (CV) error-rate estimator be employed for model selection with the RDA classifier. We remark that the LOO estimator is well known to have large variance and to yield ties among multiple candidate RDA models. Here, we consider five alternative error-rate estimators and compare them with the LOO estimator in the model selection process for the RDA classifier using four well-known, small-sample, high-dimensional microarray data sets. We find that the .632 error rate estimator from Efron and Tibshirani [1994] and the .632+ estimator from Efron and Tibshirani [1997] yield models that degrade the classification performance of the

RDA classifier. We recommend the 10-fold *CV* estimator or the bootstrap *CV* estimator from Fu, Carroll, and Wang [2005] for model selection with the *RDA* classifier and have developed the `R` package `regdiscrim`, which implements the *RDA* classifier. Additionally, we have developed the `R` package `errorest`, which implements the competing conditional error-rate estimators.

In Chapters 4 and 5 we consider a family of naive diagonal classifiers for small-sample, high-dimensional microarray data, such as the diagonal linear discriminant analysis (*DLDA*) classifier, popularized by Dudoit, Fridlyand, and Speed [2002] and rigorously studied by Bickel and Levina [2004]. Although the *DLDA* classifier has been shown to have excellent classification performance, both in theory and in practice, the classifier omits relevant pairwise correlations and, thus, classificatory information present in the off-diagonal elements of the sample covariance matrices of each group.

In Chapter 4 we consider the case where a data set consists of two classes (populations) and propose two alternative classifiers that are often superior to the DLDA classifier and its competing variants because our proposed classifiers preserve off-diagonal classificatory information by simultaneously diagonalizing the sample covariance matrix of each class. Using four well-known microarray data sets, we demonstrate that our proposed classifiers can yield superior classification performance compared to the *DLDA* classifier, the shrinkage-based *DLDA* (*SDLDA*) classifier from Pang, Tong, and Zhao [2009], and the shrinkage-mean-based *DLDA* (*SmDLDA*) classifier from Tong, Chen, and Zhao [2012]. Furthermore, in the derivation of our proposed *SimDiag* classifier, we provide a direct generalization of a result from Fukunaga [1990] to simultaneously diagonalize two positive-semidefinite real symmetric matrices in a feature subspace.

In Chapter 5 we consider the case where a data set is generated from three or more classes. We propose a classifier that utilizes a whitening transform [Duda, Hart,

and Stork, 2001] to diagonalize a pooled sample covariance matrix estimator to improve the diagonal covariance matrix assumption. We also employ two simultaneous diagonalization algorithms from Asfari [2006] and Souloumiac [2009] to nearly simultaneously diagonalize the sample covariance matrices from each class to improve the naive diagonal assumption prior to applying the *DLDA* classifier. In short, we first reduce the naiveté of the diagonal covariance matrix assumption before employing the *DLDA* classifier to improve its classification performance. Using two well-known microarray data sets, we demonstrate that our proposed classifiers yield improved classification performance compared to the DLDA classifier, the *SDLDA* classifier, and the *SmDLDA* classifier. We have developed the `R` package `diagdiscrim`, which contains an implementation of the classifiers that we have proposed in Chapters 4 and 5 as well as the competing diagonal classifiers.

We remark that the evaluation of statistical and machine learning methods is of particular interest to researchers at the Pacific Northwest National Laboratory (PNNL) in Richland, Washington. Researchers at PNNL are especially interested in the evaluation of semi-supervised and unsupervised learning methods, including clustering algorithms, for high-dimensional data sets, where no gold standard or classification labels are present. We have worked jointly with PNNL researchers to develop our clustering stability evaluation statistic proposed in Chapter 2 that can be deployed using the open-source statistical computing environment `R` and the powerful PNNL Institutional Computing (PIC) platform. Furthermore, we conducted all simulations reported in this dissertation using the PIC platform.

CHAPTER TWO

Cluster Stability Evaluation of Gene Expression Data via Cluster Omission

2.1 Introduction

The advent of genomics technology, in particular microarray experiments, has yielded extremely large and complex data sets. As an initial exploratory step of microarray data, unsupervised learning techniques, such as clustering algorithms, are frequently used to identify apparent groups of functional gene relationships. Additionally, the application of clustering algorithms facilitates insightful discovery, such as uncovering prognostic subclasses or tumor subtypes of cancer [McLachlan, Do, and Ambroise, 2004]. Consequently, the identification of useful and accurate group structures is essential. Numerous clustering algorithms have been proposed in the literature of many disciplines, including machine learning, knowledge discovery, image processing, and bioinformatics. Thus, a researcher has a vast and perhaps overwhelming collection of methods from which to choose when exploring a microarray data set. Clustering algorithms may potentially find nonexistent relationships in microarray data, thereby impeding research progress. Furthermore, two distinct clustering algorithms might suggest opposing structures in the data. Thus, the assessment of discovered clusters is imperative and is as important as the discovered clusters themselves [Yeung et al., 2001]. Typical approaches to assessing cluster quality include the use of prior biological knowledge, visual inspection, and biological experimentation, but early in a research investigation these assessments may be limited or ambiguous.

Over the last decade, a number of clustering assessment methods have been proposed in the bioinformatics literature. These methods are classified into one of three categories: internal validity, external validity, and stability [Handl, Knowles,

and Kell, 2005]. Internal validity measures evaluate intrinsic information in the data, such as compactness, connectedness, and separation. External validity measures utilize a gold standard data set that is compared to the determined clusters. Finally, stability validation methods measure the consistency and variability of a clustering algorithm when applied to a given microarray data set.

Stability validation methods effectively consider the original clustering on the unperturbed data as the baseline ground truth to which similarity comparisons are made with clusterings on resampled data, providing a reasonable assessment of the original clustering as ground truth. Hence, we can often view stability validation methods as approximate external validation methods. If the similarity scores on proposed clusters are far from optimal, then we can usually infer that the proposed clustering is poor. As discussed by Hennig [2007] and Handl et al. [2005], cluster stability alone does not indicate good clustering, but when paired with other clustering assessment methods, stability validation methods can provide useful insight about the proposed clusters.

The present clustering literature lacks a solid theoretical foundation for many of the proposed clustering evaluation criteria. Many of these methods, whether internal validity, external validity, or stability, are a function of the data, the determined clusters, and, if applicable, an external validation set [Handl et al., 2005]. However, the sampling distribution or even an estimated standard error is seldom provided for these statistics. Consider, for example, the *FOM* statistic from Yeung et al. [2001], who have used an approach that resembles the jackknife method [Efron and Tibshirani, 1994] to assess the predictive power of a clustering algorithm. Although Yeung et al. [2001] have provided the expectation of the *FOM* statistic under mild assumptions, they have not provided the sampling distribution for the *FOM* statistic or even an approximate sampling distribution. Additionally, the *FOM* statistic lacks interpretability and results in a relative score, so that the observed *FOM* statistics

for two distinct clustering algorithms can be compared only for a given data set to determine a relative best clustering; researchers have no method to determine if the minimum observed score actually indicates a trustworthy clustering. Also, without a sampling distribution, researchers have no scale by which they can determine if two observed *FOM* scores are statistically or practically different.

Despite the lack of statistical foundation that we often see in the clustering validation literature, some researchers have attempted to formulate a statistical distribution for some clustering evaluation statistics [Dudoit and Fridlyand, 2003, Datta and Datta, 2006, Hennig, 2007]. To our knowledge, the majority of assessment methods that provide a sampling distribution do so via a resampling approach. Bootstrapping, which is often used when a statistic's sampling distribution is complex or unknown, is a useful tool to approximate the sampling distribution of a statistic with only the original data in hand [Efron, 1979].

We propose a clustering stability method based on a subset of the decision-theoretic admissibility conditions proposed by Fisher and Van Ness [1971], who have provided guidelines for a reasonable clustering algorithm. Their guidelines have established a systematic foundation that is often lacking in the evaluation of clustering algorithms. Specifically, based on the cluster omission admissibility condition from Fisher and Van Ness [1971], we propose the *ClustOmit* cluster stability statistic and approximate its sampling distribution statistic using a stratified, nonparametric bootstrapping method. Furthermore, we use the apparent variability in the sampling distribution as a diagnostic tool for further evaluation of the proposed clusters.

We compare our proposed *ClustOmit* statistic with the *FOM* statistic from Yeung et al. [2001] and a clustering stability method from Hennig [2007]. With the *ClustOmit* statistic, we utilize the Jaccard similarity coefficient [Jaccard, 1912] to provide a clear interpretation of the cluster assessment that is lacking from the *FOM* statistic. Furthermore, we demonstrate that our proposed clustering stability

method is better suited to comparing clustering algorithms than the *FOM* statistic because the Jaccard similarity coefficient provides an absolute scale by which we can indicate if a proposed clustering is reasonable as well as compare two distinct clustering algorithms.

The cluster stability method from Hennig [2007] also utilizes the Jaccard similarity coefficient with a bootstrapping scheme comparable to the *ClustOmit* method. After obtaining a clustering of the original data, Hennig's method computes the maximum of the Jaccard similarity coefficients between a given cluster from the original clustering and the clusters obtained from a bootstrapped data set. For several bootstrapped data sets, the mean maximum Jaccard coefficient is computed for each of the original clusters. In this manner, Hennig has attempted to overcome the well-known label-switching problem [Yao, 2012, Jakobsson and Rosenberg, 2007, Stephens, 2000, Richardson and Green, 1997], where clusters are arbitrarily labeled across bootstrap replicates, to determine the average stability of individual clusters. We argue that this *ad hoc* cluster matching adds additional variability to the bootstrapped Jaccard coefficient. We show that our proposed *ClustOmit* statistic yields similar results to Hennig's method but with reduced variability because the *ClustOmit* statistic does not employ cluster matching.

We have organized the remainder of our paper as follows. In Section 2 we present necessary notation and preliminaries to facilitate our proposed method. In Section 3 we discuss the admissibility conditions from Fisher and Van Ness [1971], and in Section 4 we present our proposed cluster stability method. We compare our proposed method with the competing validation methods from Yeung et al. [2001] and Hennig [2007] on three simulated data models and a microarray data set from Khan et al. [2001] in Section 5. Finally, we conclude with a brief discussion in Section 6.

2.2 Preliminaries

Suppose we have n observations, and let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})' \in \mathbb{R}_{p\times 1}$ be the ith observation with p gene expression levels for $i = 1, \ldots, n$. Let $\mathcal{L} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ denote the learning (training) data set. We assume the pair $(\boldsymbol{x}_i, y_i)$ is a realization from a mixture distribution of M populations (classes), where $y_i \in \mathcal{M} = \{1, 2, \ldots, M\}$ denotes the true, unique population membership of observation $\boldsymbol{x}_i$. In practice, we do not know the true population membership of each observation in $\mathcal{L}$; additionally, the number of populations M is typically unknown. Hence, we must not only determine the membership of each $\boldsymbol{x}_i \in \mathcal{L}$, but we must also estimate M. See Jain [2010] for more information concerning the estimation of M.

In this work, we consider clustering algorithms that first require the specification of a fixed number of clusters K and then determine the cluster membership of each $\boldsymbol{x}_i \in \mathcal{L}$. Formally, we define a clustering procedure $\mathcal{P} : \mathcal{L} \to \mathcal{K}$, where $\mathcal{K} = \{1, \ldots, K\}$ denotes the candidate clustering labels. Let $C_k = \{\boldsymbol{x}_i \in \mathcal{L} | \mathcal{P}(\boldsymbol{x}_i) = k, k \in \mathcal{K}\}$ denote the kth cluster such that $\cup_{k=1}^{K} C_k = \mathcal{L}$ and $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$, i.e., the clusters comprise mutually exclusive and exhaustive sets in $\mathcal{L}$. Let $\mathcal{L}_{\setminus k} = \{\boldsymbol{x}_i \in \mathcal{L} | \boldsymbol{x}_i \notin C_k\}$ be the set of observations from $\mathcal{L}$ that remain after the kth cluster is omitted. Also, let $\mathcal{C} = \{\mathcal{P}(\boldsymbol{x}_1), \mathcal{P}(\boldsymbol{x}_2), \ldots, \mathcal{P}(\boldsymbol{x}_n)\}$ denote the sequence of clustering labels assigned by the clustering algorithm $\mathcal{P}$, and let $\mathcal{C}_{\setminus k}$ be the sequence of clustering labels with the kth cluster omitted.

Following Tibshirani and Walther [2005], we say that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are *comembers* of the cluster C_k if $\boldsymbol{x}_i \in C_k$ and $\boldsymbol{x}_j \in C_k$. Suppose $\mathcal{U} = \{\mathcal{P}_U(\boldsymbol{x}_1), \ldots, \mathcal{P}_U(\boldsymbol{x}_n)\}$ and $\mathcal{V} = \{\mathcal{P}_V(\boldsymbol{x}_1), \ldots, \mathcal{P}_V(\boldsymbol{x}_n)\}$ are two partitions of $\mathcal{L}$ such that $\mathcal{P}_U : \mathcal{L} \to \mathcal{K}_\mathcal{U}$, $\mathcal{P}_V : \mathcal{L} \to \mathcal{K}_\mathcal{V}$ with $\mathcal{K}_\mathcal{U} = \{1, \ldots, K_U\}$, $\mathcal{K}_\mathcal{V} = \{1, \ldots, K_V\}$, and K_U and K_V are not necessarily equal. For $1 \leq i, j \leq n$, we write the comemberships of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as $U_{ij} = I[P_U(\boldsymbol{x}_i) = P_U(\boldsymbol{x}_j)]$ and $V_{ij} = I[P_V(\boldsymbol{x}_i) = P_V(\boldsymbol{x}_j)]$ to indicate if the observations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are clustered together in $\mathcal{U}$ and $\mathcal{V}$, respectively, where

$I[z] = 1$ if the Boolean statement z is true, and 0 otherwise. We remark that other authors, such as Monti, Tamayo, Mesirov, and Golub [2003], refer to comembership as *connectivity*, while Jackson, Somers, and Harvey [1989] use the term *co-occurrence*.

One cannot always determine a best clustering algorithm when multiple clustering algorithms have been applied to the same data set. Instead, one can employ a similarity measure that yields an agreement level between the resulting clusters from two clusterings of $\mathcal{L}$. We use the Jaccard similarity coefficient, $\mathcal{J} \in [0, 1]$, to measure the similarity between the two clusterings $\mathcal{U}$ and $\mathcal{V}$. We write $\mathcal{J}$ as

$$\mathcal{J} = \frac{\sum_{i<j} U_{ij} V_{ij}}{\sum_{i<j} (U_{ij} + V_{ij} - U_{ij} V_{ij})}, \tag{2.1}$$

which is the proportion of observation pairs in $\mathcal{L}$ that have been clustered together in both $\mathcal{U}$ and $\mathcal{V}$ such that each pair has been clustered together in at least one of the two partitions. Often (2.1) is written as $\mathcal{J} = \frac{|\mathcal{U} \cap \mathcal{V}|}{|\mathcal{U} \cup \mathcal{V}|}$ to emphasize the overlap of the partitions $\mathcal{U}$ and $\mathcal{V}$. We define $\mathcal{J} = 0$ for $|\mathcal{U} \cup \mathcal{V}| = 0$ (i.e., each of $\mathcal{U}$ and $\mathcal{V}$ consists of singleton clusters).

Values of $\mathcal{J}$ near zero suggest that little agreement exists between the two partitions, whereas values near one suggest strong agreement. Furthermore, as discussed by Fligner, Verducci, and Blower [2002], the Jaccard coefficient can be viewed as a proportion from a binomial model, thereby providing a natural probabilistic interpretation of the similarity between clusterings. We utilize the Jaccard similarity coefficient because it is widely known and is easy to interpret, but other similarity measures, such as the (adjusted) Rand index or the Minkowski score, can be used in our proposed cluster evaluation method. We refer the reader to Handl et al. [2005] for a concise discussion of other similarity measures and their applications to clustering evaluation. Also, see Hennig [2007] for more details regarding the Jaccard similarity coefficient.

To assess clustering stability, we are specifically interested in the case where $\mathcal{U}$ is a clustering of $\mathcal{L}$ and $\mathcal{V}$ is a clustering of a resampled data set from $\mathcal{L}$. Following Hennig [2008] and Hennig [2010], we consider a stable clustering to have a Jaccard value of at least 0.75, with values above 0.85 suggesting a highly stable clustering. Furthermore, we consider values between 0.6 and 0.75 to suggest structure within the data but with inexact and unreliable cluster membership. Jaccard similarity values less than 0.6 suggest that the original clustering is untrustworthy.

2.2.1 The Competing Clustering-Evaluation Methods.

We compare our proposed *ClustOmit* statistic that is discussed in Section 2.4 with two widely-known clustering-evaluation methods: the *FOM* method from Yeung et al. [2001] and the cluster stability method from Hennig [2007]. The *FOM* method resembles the jackknife and leave-one-out (LOO) cross-validation methods. Rather than omitting each observation in sequence as with *LOO* cross-validation, the *FOM* method aggregates the average distance of each observation to its cluster centroid (typically, the cluster mean) after removing a single gene expression level across all samples. The reported *FOM* score is the average of the aggregated distances corresponding to each gene expression level being removed. Yeung et al. [2001] have discussed that the *FOM* statistic can be used only for relative comparisons of clustering algorithms on the same data set for a specified value of K. The lack of interpretation and relative scale of the *FOM* statistic is a drawback in practice because one can determine only a relative best clustering on a given data set for considered clustering algorithms. Moreover, no scale has been provided by which one can determine if two *FOM* scores are statistically or practically different, nor can one determine if an observed *FOM* estimate alone suggests that the clustering is adequate.

Hennig [2007] has proposed a bootstrapping method to evaluate the performance of clustering algorithms with the assumption that a small change in the data should yield a similar clustering. For each bootstrap replicate, the specified similarity measure between a specified cluster from the original clustering and each of the bootstrapped clusters is computed, and the maximum similarity value is recorded. Although the *ad hoc* recording of the maximum similarity measure is intuitive and can be effective, this approach adds an additional source of variability to the Jaccard similarity statistic. We refer to the method from Hennig [2007] as the Hennig method. Our proposed method, discussed in Section 2.4, avoids the *ad hoc* selection of similarity scores and, therefore, exhibits lower variability, thereby providing a more accurate and reliable assessment of clustering stability.

2.3 Clustering Admissibility Conditions

Fisher and Van Ness [1971] have provided nine admissibility conditions based on decision-theoretic principles that restrict algorithms to admissible decision rules for clustering algorithms. These criteria yield a set of properties that any reasonable clustering algorithm should satisfy and, more importantly, provide guidelines with which one can determine if a clustering method gives unreasonable results. Fisher and Van Ness' work has established a necessary foundation for the assessment of clustering methods applied to microarray data so that investigators do not base future research on unreasonable clusters that may arise from a particular clustering algorithm. However, Fisher and Van Ness [1971] did not provide statistical methods to assess each of their proposed admissibility conditions but instead provided an initial set of principles for assessing clustering algorithms.

In Section 2.4 we discuss our proposed statistical method that extends a portion of Fisher and Van Ness' work to a practical setting. To our knowledge, we are the first to do so. We focus specifically on the cluster omission admissibility condi-

tion to evaluate the stability of cluster boundaries and to provide reasonable doubt towards any set of anomalous clusters. We define the cluster omission admissibility condition as follows: when a clustering procedure is applied to $\mathcal{L}$ resulting in clusters $C_1, \ldots, C_K$, then, for each k, if the same procedure is applied to $\mathcal{L}_{\setminus k}$, the same $K-1$ clusters $C_1, \ldots, C_{k-1}, C_{k+1}, \ldots, C_K$ should result.

The assessment of cluster omission admissibility and other admissibility criteria from Fisher and Van Ness [1971] are clearly suited to a bootstrapping approach because several of these criteria suggest that clusterings should be robust to the addition or omission of clusters or observations.

2.4 Cluster Omission Stability Method

We present our clustering-evaluation method based on the cluster omission admissibility condition provided by Fisher and Van Ness [1971] to assess the K clusters determined by a specified clustering algorithm, $\mathcal{P}$, using B bootstrap replications. We refer to our proposed method as the cluster omission (*ClustOmit*) stability method and provide our method here:

(1) Fix the number of clusters, K.

(2) Apply the clustering method $\mathcal{P}$ to $\mathcal{L}$ to obtain the clusters, $C_1, \ldots, C_K$, and the corresponding set of cluster labels $\mathcal{C}$.

(3) For $k = 1, \ldots, K$:

  (a) Construct the set, $\mathcal{L}_{\setminus k} = \mathcal{L} \setminus C_k$, and the corresponding set of cluster labels $\mathcal{C}_{\setminus k}$.

(4) For $b = 1, \ldots, B$:

  (a) For $k = 1, \ldots, K$:

    (i) Draw a bootstrap sample $\mathcal{L}_{\setminus k}^{(b)}$ from the original data set $\mathcal{L}_{\setminus k}$.

(ii) Apply the clustering method $\mathcal{P}$ to the bootstrap sample $\mathcal{L}_{\setminus k}^{(b)}$ to obtain $K-1$ clusters and denote the corresponding cluster labels as $\mathcal{C}_{\setminus k}^{(b)}$.

(iii) Compute the Jaccard similarity, $\mathcal{J}_{\setminus k}^{(b)} = \mathcal{J}(\mathcal{C}_{\setminus k}, \mathcal{C}_{\setminus k}^{(b)})$.

(b) Compute the average Jaccard similarity, $\overline{\mathcal{J}}_b = \sum_{k=1}^{K} w_k \mathcal{J}_{\setminus k}^{(b)}$,

where $w_k = |\mathcal{C}_k|/n$ denotes the proportion of observations assigned to cluster $\mathcal{C}_k$, and $|\mathcal{A}|$ denotes the cardinality of the set, $\mathcal{A}$. The above algorithm generates a sequence $\overline{\mathcal{J}}_b$ $(b = 1, \ldots, B)$ that is an approximate sampling distribution for the average Jaccard similarity between the original clustering $\mathcal{C}$ and the clustering $\mathcal{C}_{\setminus k}$ after each of the k $(k = 1, \ldots, K)$ clusters are sequentially removed.

To draw the bth $(b = 1, \ldots, B)$ bootstrap sample in step 4.a.i of our algorithm, we draw a sample with replacement from $\mathcal{L}_{\setminus k}$ to obtain the set of observations $\{\boldsymbol{x}_1^{(b)}, \boldsymbol{x}_2^{(b)}, \ldots, \boldsymbol{x}_s^{(b)}\}$, where s is the number of observations in $\mathcal{L}_{\setminus k}$. We employ a stratified sampling scheme to ensure that $\mathcal{C}_{k'}$ maintains the same number of observations for each bootstrap replication $(k' = 1, \ldots, k-1, k+1, \ldots, K)$. We remark that the application of clustering to repeated observations is related to the point proportion admissible condition from Fisher and Van Ness [1971] and, therefore, deserves further attention in a future study.

A key strength of the *ClustOmit* method is that we can examine the stability of individual clusters along with the overall clustering. In particular, for each $k = 1, \ldots, K$, in step 4.a.iii of the *ClustOmit* method, we can retain the Jaccard score for the kth cluster. As we will see in Section 2.5, we can utilize the bootstrapped Jaccard scores for each cluster to identify individual unstable clusters.

We restrict our method to $K \geq 3$ because the omission of $K = 1$ cluster omits each $\boldsymbol{x}_i \in \mathcal{L}$. Also, for $K = 2$, the Jaccard coefficient yields one for each bootstrap replicate because only one cluster remains after the omission of the second cluster.

2.5 Data Sets and Results

We compared the performance of the *ClustOmit*, the *FOM*, and the Hennig methods on three simulated data models and the well-known microarray data set from Khan et al. [2001]. In our study we used the implementation of the *FOM* statistic in the `clValid` R package [Brock, Datta, Datta, and Pihur, 2008] and the implementation of the Hennig method with the `clusterboot` function from the `fpc` package. We used version 2.15.0 of the open source statistical software `R` for the simulations presented in this section and used the `R` package `ggplot2` [Wickham, 2009] to create our summary plots. For the K-means algorithm, we used the `kmeans` function in `R` with the method proposed by Hartigan and Wong [1979] and 20 random starts. Also, we applied the `mclust` function and its default arguments from the `Mclust` package [Fraley and Raftery, 2006] for the *MBC* algorithm [Fraley and Raftery, 2002], and we utilized the `diana` function available in the `cluster` package with a Euclidean dissimilarity matrix. Each of the `R` packages that we used is available on CRAN.

2.5.1 Simulated Data Sets

For our simulation study, we generated n_m $(m = 1, \ldots, M)$ observations from population Π_m so that the Euclidean distance between each of the population centroids and the origin was equal and was scaled by $\Delta \geq 0$. We constructed the M populations such that for $\Delta = 0$, we had the configuration $\Pi_1 = \Pi_2 = \ldots = \Pi_M$. That is, for $\Delta = 0$ we had M identical populations, and the obtained clusters from each clustering algorithm for $K > 1$ were in error. Additionally, for small values of Δ, the separation among the populations was negligible. The consideration of small values of Δ is imperative to reflect that real data often lack separation. However, for $K = M$ and for sufficiently large Δ, we expected each clustering algorithm to correctly cluster the generated observations.

For our simulated data sets, we assessed the efficacy of each clustering-evaluation method with three different data-generation models, where each model consisted of $M = 5$ populations drawn from a family of probability distributions with $\Delta = 0.0, 0.5, \ldots, 2.5, 3.0$. For a specified value of Δ, we generated 1000 data sets and applied the three clustering algorithms to each data set with $K = 4$, 5, and 6 to examine the correct and incorrect specification of the number of clusters. Then, for each clustering algorithm, we computed the three clustering evaluation scores for each value of K. Specifically, for each clustering algorithm, we constructed an approximate sampling distribution for our *ClustOmit* statistic with $B = 100$ and stored the average Jaccard similarity value as discussed in Section 2.4. Similarly, for comparative purposes, we computed the Hennig scores for each cluster with $B = 100$ bootstrap replications and stored a weighted average of the stability scores with weights equaling the proportion of observations in each cluster. Then, we averaged the 1000 computed values for each clustering-evaluation method. We remark that the aggregation of the observed values of the Hennig method across each cluster was not Hennig's original intent.

2.5.1.1 *Model I – Multivariate Uniform.* Let $\boldsymbol{x} = (X_1, \ldots, X_p)'$ be a multivariate uniformly distributed random vector such that $X_j \sim U(a_j, b_j)$ is an independently distributed uniform random variable with $a_j < b_j$ for $j = 1, \ldots, p$. We generated $n_m = 25$ observations from each population, where

$$\Pi_1 = U(-1/2, 1/2) \times U(\Delta - 1/2, \Delta + 1/2) \times U(-1/2, 1/2) \times U(-1/2, 1/2),$$

$$\Pi_2 = U(\Delta - 1/2, \Delta + 1/2) \times U(-1/2, 1/2) \times U(-1/2, 1/2) \times U(-1/2, 1/2),$$

$$\Pi_3 = U(-1/2, 1/2) \times U(-\Delta - 1/2, -\Delta + 1/2) \times U(-1/2, 1/2) \times U(-1/2, 1/2),$$

$$\Pi_4 = U(-1/2, 1/2) \times U(-1/2, 1/2) \times U(-\Delta - 1/2, -\Delta + 1/2) \times U(-1/2, 1/2),$$

$$\Pi_5 = U(-1/2, 1/2) \times U(-1/2, 1/2) \times U(-1/2, 1/2) \times U(\Delta - 1/2, \Delta + 1/2).$$

Here, the support of each population Π_m $(m = 1, \ldots, M)$ is a unit hypercube, and for $\Delta \geq 1$, the populations are mutually exclusive.

2.5.1.2 *Model II – Multivariate Normal.* We generated $n_m = 25$ observations from the p-dimensional multivariate normal distribution $N_p(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, where $\boldsymbol{\mu}_m \in \mathbb{R}_{p\times 1}$ is the population mean vector and $\boldsymbol{\Sigma}_m \in \mathbb{R}_{p\times p}$ is the positive-definite covariance matrix for population $m = 1, \ldots, M$. Here, we considered $M = 5$ populations with $p = 50$ features, where the sample sizes for each population were half the number of features.

To define the mth population mean vector, let $\boldsymbol{e}_m \in \mathbb{R}_{p\times 1}$ be the mth standard basis vector such that the mth element of $\boldsymbol{e}_m$ is one and the remaining elements are zero $(m = 1, \ldots, M)$. Then, we defined the mth population mean vector as

$$\boldsymbol{\mu}_m = \Delta \sum_{j=1}^{p/M} \boldsymbol{e}_{(p/M)(m-1)+j}.$$

Effectively, we translated the first mean in the first 10 dimensions, the second mean in the second 10 dimensions, and so on.

Also, we considered equal intraclass covariance (correlation) matrices such that $\boldsymbol{\Sigma}_m = \sigma^2(1 - \rho_m)\boldsymbol{J}_p + \rho_m \boldsymbol{I}_p$, where $-(p-1)^{-1} < \rho_m < 1$, $\boldsymbol{I}_p$ is the $p \times p$ identity matrix, and $\boldsymbol{J}_p$ denotes the $p \times p$ matrix of ones. We chose $\rho_m = 0.9$ for $m = 1, \ldots, M$ to examine the effect of highly correlated, ellipsoidal data because Handl et al. [2005] have demonstrated the difficulty that some clustering algorithms, including the K-means algorithm, have with this type of data. For simplicity, we let $\sigma^2 = 1$.

2.5.1.3 *Model III – Multivariate Student's t.* We utilized $M = 5$ multivariate Student's t populations to explore the impact of heavy tails on clustering algorithms, which tend to result in singleton or small anomalous clusters – a common issue in clustering microarray data. Let $\boldsymbol{x} \sim T_p(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, c_m)$ denote the multivariate

Student's t distribution, where $\boldsymbol{\mu}_m \in \mathbb{R}_{p\times 1}$ is the mth population location vector, $\boldsymbol{\Sigma}_m \in \mathbb{R}_{p\times p}$ is the positive-definite covariance matrix, and c_m is the degrees of freedom for the mth population ($m = 1, \ldots, M$). For small values of c_m, the tails are heavier, and, therefore, the average number of outlying observations is increased.

For $m = 1, \ldots, M$, we generated observations with $\boldsymbol{\mu}_m = \Delta(\boldsymbol{e}_m + \boldsymbol{e}_{2m})$ to translate the location vector in two dimensions, and we used a common covariance matrix $\boldsymbol{\Sigma}_m \equiv \boldsymbol{I}_p$ for all populations. Also, we generated more observations from the populations that have heavier tails to increase the expected number of outlying observations. For $m = 1, \ldots, 3$, we set $n_m = 25$ and $c_m = 10$, and for $m = 4, 5$, we chose $n_m = 50$ and $c_m = 3$.

2.5.1.4 *Simulation Results.* In Figures 2.1, 2.2, and 2.3, we display the average *FOM*, Hennig, and *ClustOmit* statistics, respectively, as a function of Δ for each data-generating model for $K = 4$, 5, and 6. For the Uniform data model, the average *FOM* statistics appeared to increase as Δ increased and were nearly the same for each clustering algorithm. For $K = 5$ and $1.0 \leq \Delta \leq 3.0$, the Diana algorithm exhibited the least average predictive power, but we could conclude only that the Diana algorithm was outperformed by the K-means and *MBC* algorithms. Applying the Hennig and *ClustOmit* methods for the true value of $K = 5$, we can conclude that the K-means and *MBC* algorithms became highly stable for $1.0 \leq \Delta \leq 3.0$, while the Diana algorithm did not become highly stable until $1.5 \leq \Delta \leq 3.0$.

The average *FOM* statistics were nearly the same for all considered values of K and Δ for the Normal model. Applying the Hennig and *ClustOmit* clustering-evaluation methods, we determined that the K-means algorithm was the most stable for the true number of clusters $K = 5$. As with the Uniform data model, for $1 \leq \Delta \leq 2$, we can claim only that the *MBC* algorithm was outperformed by the other two clustering algorithms with respect to the *FOM* statistic, whereas utilizing

the Hennig and the *ClustOmit* methods, we concluded that the *MBC* algorithm yielded unstable clusterings.

For the Student's t data, we were specifically interested in the impact of outlying observations on the three clustering algorithms considered here. Applying the Hennig and *ClustOmit* clustering-evaluation methods, we determined that the Diana algorithm incorrectly yet consistently identified the outliers as clusters, but, as expected, the K-means and *MBC* increased in stability as Δ increased. Hence, using the Hennig and *ClustOmit* clustering-evaluation methods, we concluded that the Diana algorithm should be avoided for data sets where outliers are suspected.

We were unable to gain much insight into the behavior of the *FOM* statistic in part because Yeung et al. [2001] have suggested that the *FOM* statistic can be considered only for a fixed K. We add that the predictive power of the *FOM* statistic cannot be examined as a function of population separation because the *FOM* score apparently increases as the population separation increases. Finally, for the three data-generating models, we conclude that the Hennig and *ClustOmit* methods yield similar results with better interpretability than the *FOM* statistic provides.

2.5.2 A Microarray Data Set – Khan et al. [2001]

We consider the small, round blue cell tumor (SRBCT) gene expression data set provided by Khan et al. [2001]. The SRBCT cancers include four distinct diagnostic categories: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), and the Ewing family of tumors (EWS). The data set consists of 63 observations that were obtained from cDNA microarrays containing 6567 genes. Furthermore, the 63 observations include 12 NB observations, 20 RMS observations, 8 NHL observations, and 23 EWS observations. From the `R` package `pamr`, we obtained a filtered SRBCT data set that has been reduced to 2308 genes. We further filtered the data set to 250 genes using the variable selection method from Dudoit,

Fridlyand, and Speed [2002] to reduce the amount of computation involved in our bootstrap simulations.

Because the *FOM* statistic results in a single value without reference to a sampling distribution, we approximated the sampling distribution of the *FOM* statistic via nonparametric bootstrapping [Efron and Tibshirani, 1994] by randomly sampling with replacement from the data set to obtain 500 bootstrapped data sets. We then applied the *FOM* method to each of the data sets for the considered values of K. Additionally, we applied the Hennig and *ClustOmit* methods to the filtered data set with $B = 500$. We assumed the true number of populations to be $M = 4$ and compared the efficacy of the *FOM*, Hennig, and *ClustOmit* clustering-evaluation methods for $K = 3, \ldots, 6$.

2.5.2.1 *Microarray Data Results.* In Figure 2.4 we display boxplots of the bootstrapped *FOM* scores for each value of K considered here. The differences between the *FOM* statistics of the clustering algorithms for each value of K appeared statistically insignificant because the variability of the *FOM* statistics applied to each clustering algorithm was too large to indicate a best clustering algorithm. Without any interpretability of the *FOM* statistic and with the large observed variability, we were unable to obtain any conclusive results from the *FOM* statistic.

In Figures 2.5 and 2.6, we present boxplots of the $B = 500$ values obtained with the Hennig and *ClustOmit* statistics, respectively, for the Khan data set. With respect to these two clustering-evaluation methods for the correct number of clusters $K = 4$, we concluded that the K-means and *MBC* algorithms consistently were highly stable, while the Diana algorithm was more variable and less stable. Although the average values of the Hennig and *ClustOmit* methods were approximately equal for $K = 4$, 5, and 6, the *ClustOmit* statistic exhibited less variability than the Hennig statistic.

Examining the average Hennig scores for the $K = 4$ individual clusters proposed by the Diana algorithm in Figure 2.7, we noticed that Clusters 1 and 2 were unstable. Furthermore, Clusters 3 and 4 appeared stable for the majority of the $B = 500$ bootstrap replications, but both clusters exhibited instances of instability when the data were resampled. In Figure 2.8 we observed that the *ClustOmit* method yielded similar results. With respect to the *ClustOmit* method, Clusters 1 and 2 often appeared stable, while Clusters 3 and 4 were frequently unstable. We emphasize that the two stable clusters were unstable at times with respect to both the Hennig and *ClustOmit* clustering-evaluation methods and remark that these two clusters should be examined more closely. Additionally, for the individual cluster stabilities, notice that the Hennig scores had larger variability than the *ClustOmit* scores.

2.6 Discussion

Clustering is an important part of class discovery and the initial exploratory data analysis of microarray data, and numerous clustering algorithms are available to aid the researcher in finding groups for future investigation. However, the application of different clustering algorithms to a data set can yield different and contradictory results that can hinder research progress. Thus, we must assess the proposed clusterings of the candidate clustering algorithms to promote the determination of valid and reasonable clusterings. Several clustering-evaluation methods have recently been proposed in the literature to assess clustering algorithms for a given data set, but many of these methods result in an obscure efficacy score with no reference to a sampling distribution or to a magnitude scale. Thus, one often has difficulty in differentiating among the scores of several clustering algorithms and cluster sizes under consideration.

In this paper, we have formulated a clustering stability assessment statistic that is similar to the method of Hennig [2007] and is based on the decision theoretic admissibility criteria from Fisher and Van Ness [1971]. Specifically, we have developed a method to evaluate Fisher and Van Ness' cluster omission admissibility criterion for a clustering of gene expression data. We have utilized a stratified, nonparametric bootstrapping approach to approximate the sampling distribution of our proposed *ClustOmit* statistic. Our proposed method provides a helpful visual aid for clustering evaluation and is useful for determining the relative stability of a clustering algorithm applied to microarray data.

We have compared our proposed *ClustOmit* statistic with the *FOM* statistic from Yeung et al. [2001] and the cluster stability statistic from Hennig [2007] using three simulated data models and a microarray data set from Khan et al. [2001]. We have found that the *FOM* statistic is difficult to use because it yields values that lack interpretability. Furthermore, as noted by Yeung et al. [2001], we are unable to compare the *FOM* scores for different values of K and, hence, cannot utilize the *FOM* statistic as indirect or direct evidence that a given value of K is reasonable. Furthermore, the *FOM* statistic tends to increase as the separation among population increases, so that the *FOM* statistic cannot directly evaluate clustering algorithms as a function of population separation.

We have found that the method from Hennig [2007] often works well and agrees with ground truth, provided the ground truth is known. We have also determined that the *ClustOmit* statistic often yields similar results to those of the Hennig method. However, the *ClustOmit* method can avoid the *ad hoc* selection of the maximum Jaccard similarity coefficient that the Hennig method employs. Furthermore, with the microarray data set from Khan et al. [2001], our proposed method exhibited less variability than the aggregated Hennig method. Although we do not advise the direct estimation of K with either the Hennig or *ClustOmit* methods, we have

seen that both methods can corroborate the true value of K and, therefore, provide evidence for the estimation of K. Hence, our proposed *ClustOmit* statistic is useful in practice when paired with methods such as the Gap statistic from Tibshirani, Walther, and Hastie [2001], to estimate K.

As with most clustering-evaluation methods, our *ClustOmit* statistic is not intended to declare a best clustering algorithm on any specific data set, as is typically performed in a supervised learning study, but rather to present reasonable doubt towards misleading clustering results. The *ClustOmit* method allows researchers to reduce the set of candidate clusters from various clustering algorithms on a given data set to a manageable set for additional research. Furthermore, we have demonstrated that our proposed *ClustOmit* statistic can be effectively used to identify anomalous clusters.

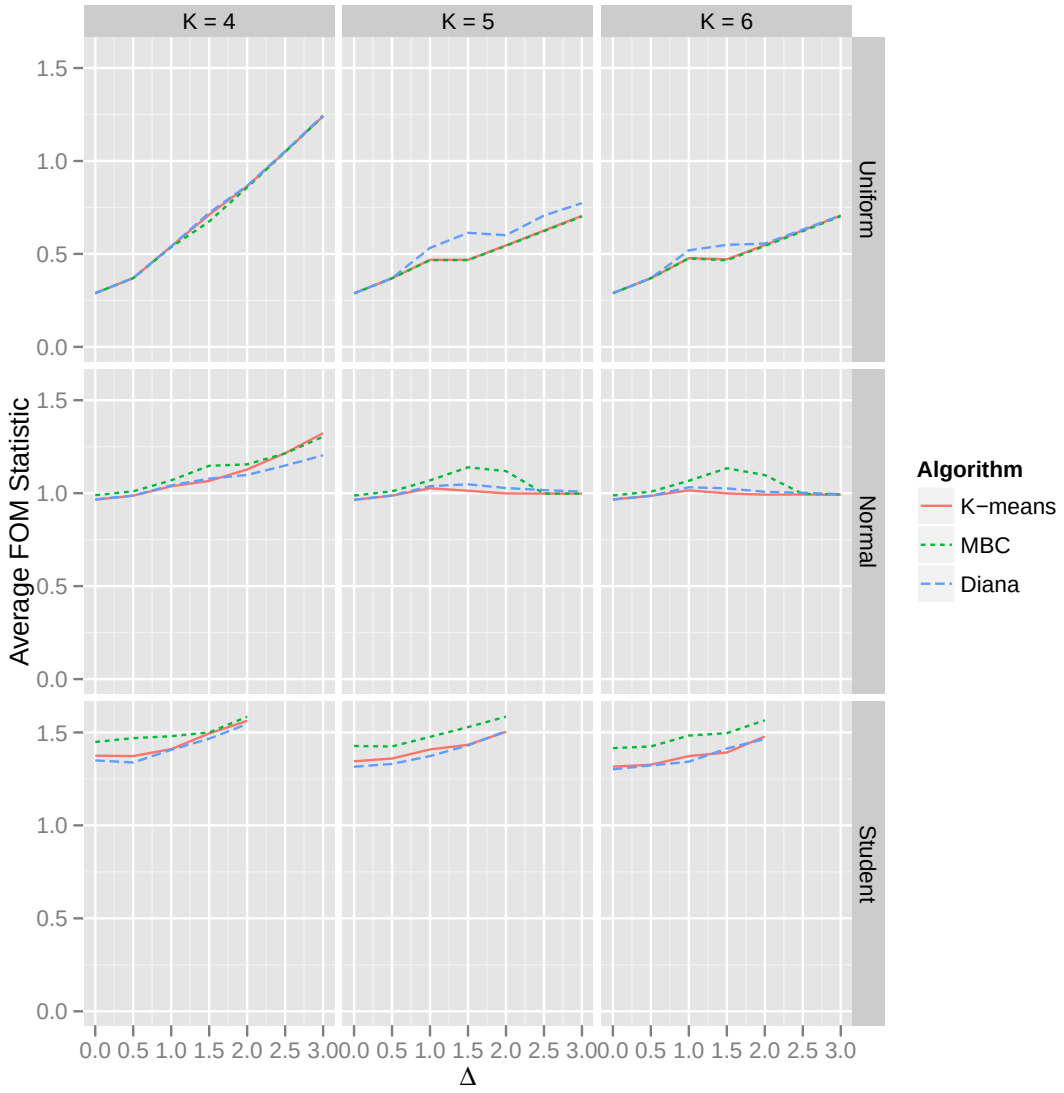

Figure 2.1: The average *FOM* statistic for the three competing clustering algorithms applied to the three data models as a function of the population separation Δ.

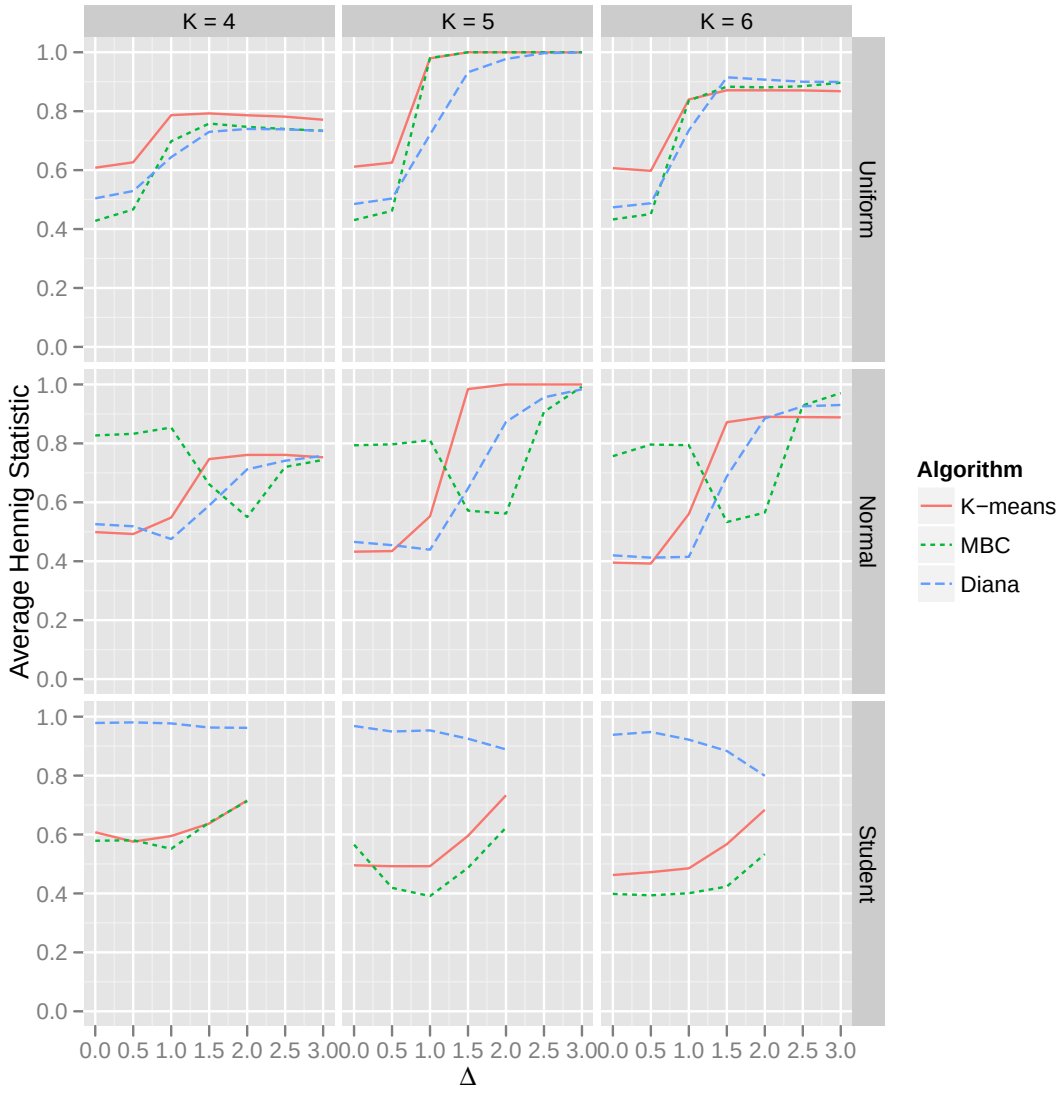

Figure 2.2: The average Hennig statistic for the three competing clustering algorithms applied to the three data models as a function of the population separation Δ.

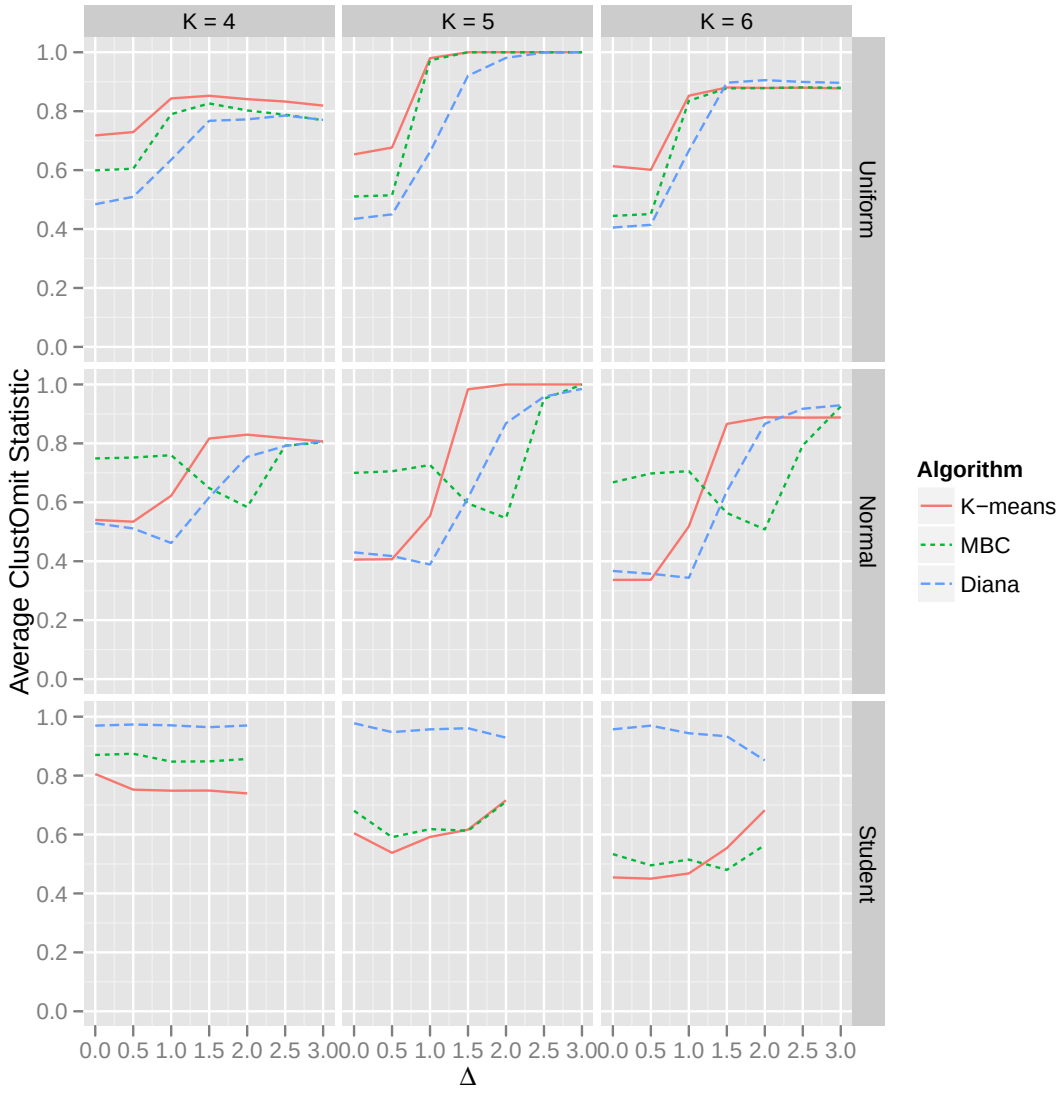

Figure 2.3: The average *ClustOmit* statistic for the three competing clustering algorithms applied to the three data models as a function of the population separation Δ.

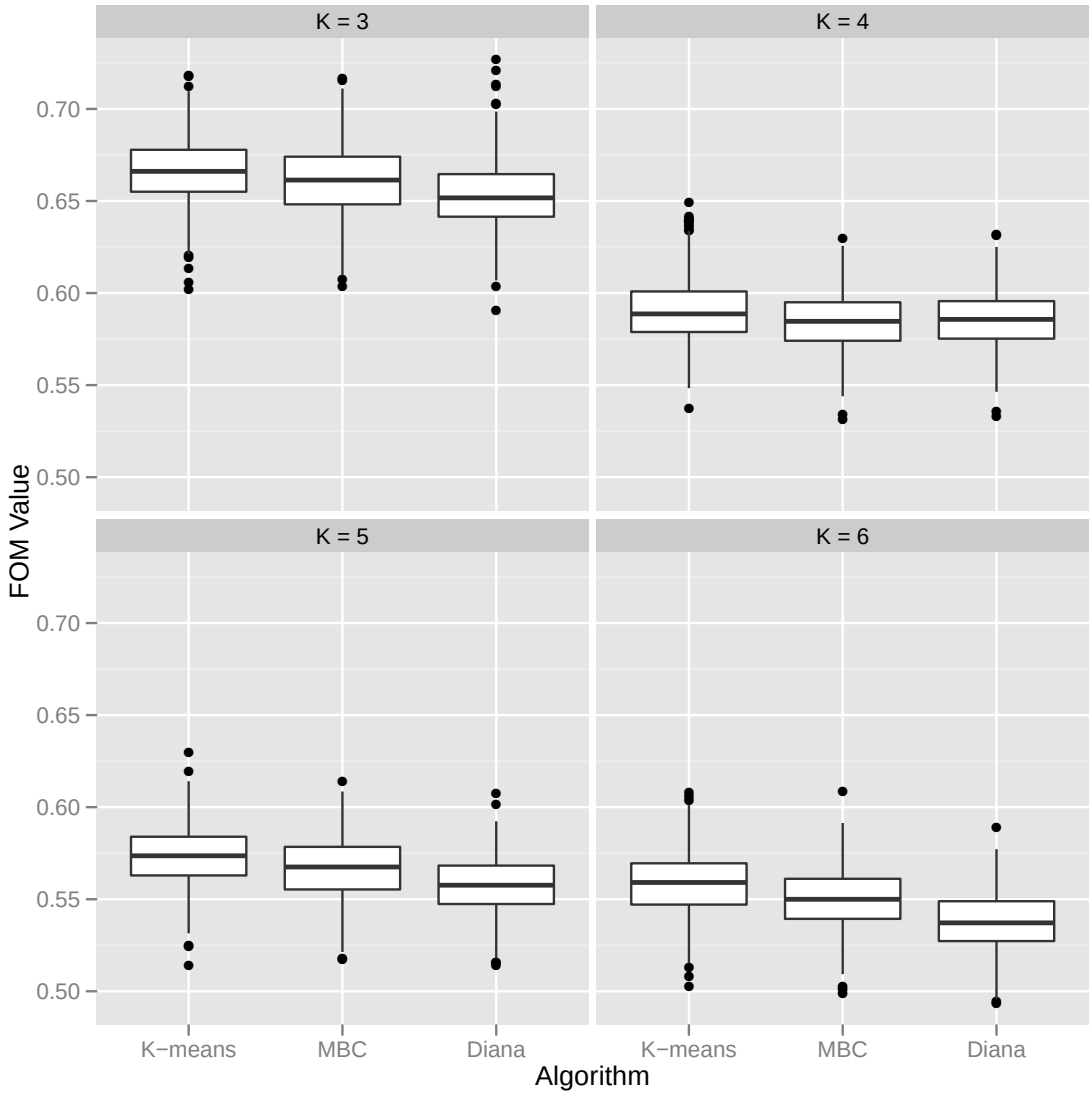

Figure 2.4: Box plots of the *FOM* statistic for 500 bootstrap replications of the Khan data set for $K = 3, \ldots, 6$.

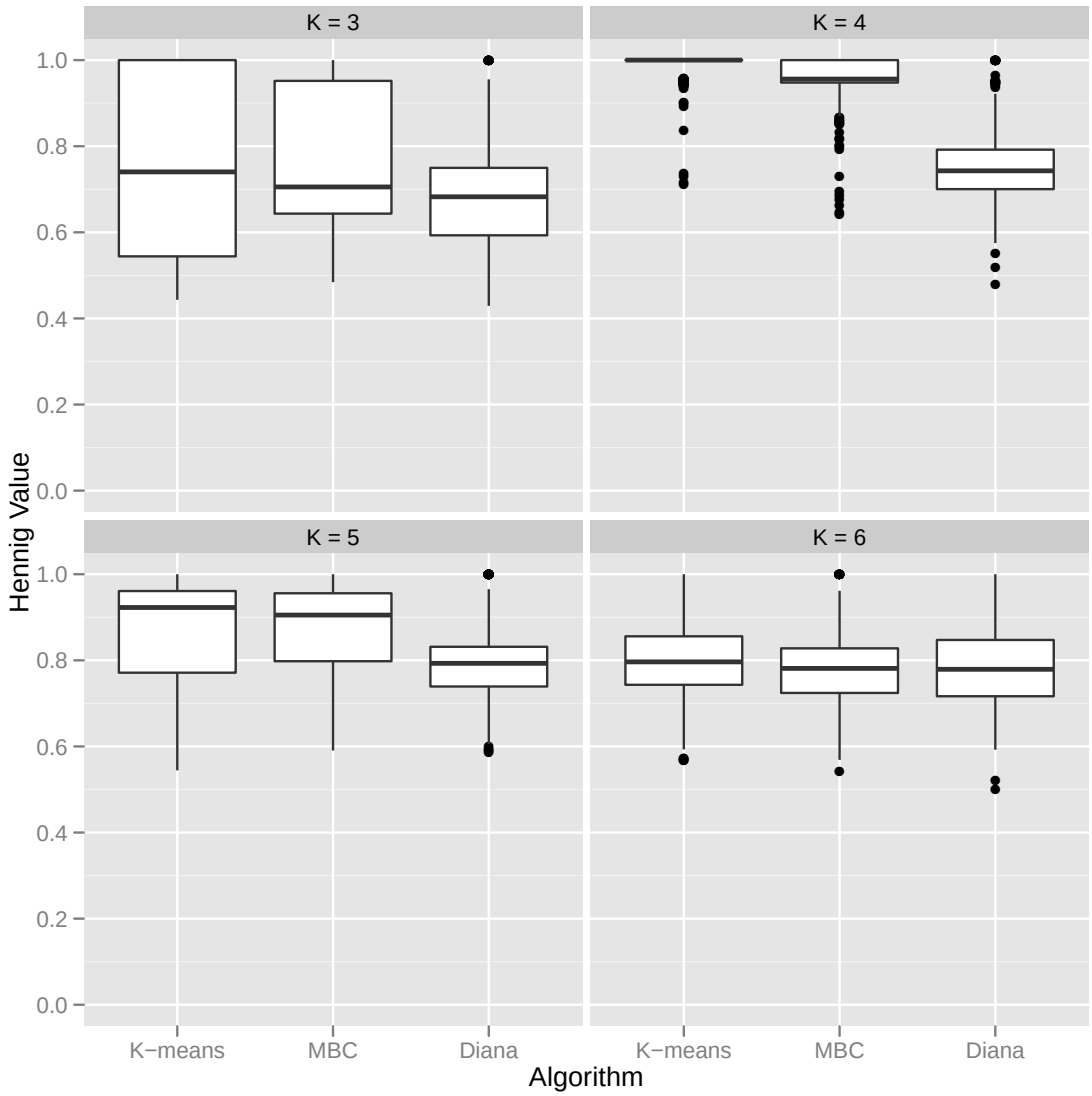

Figure 2.5: Box plots of the Hennig statistic obtained from the Khan data set for $B = 500$.

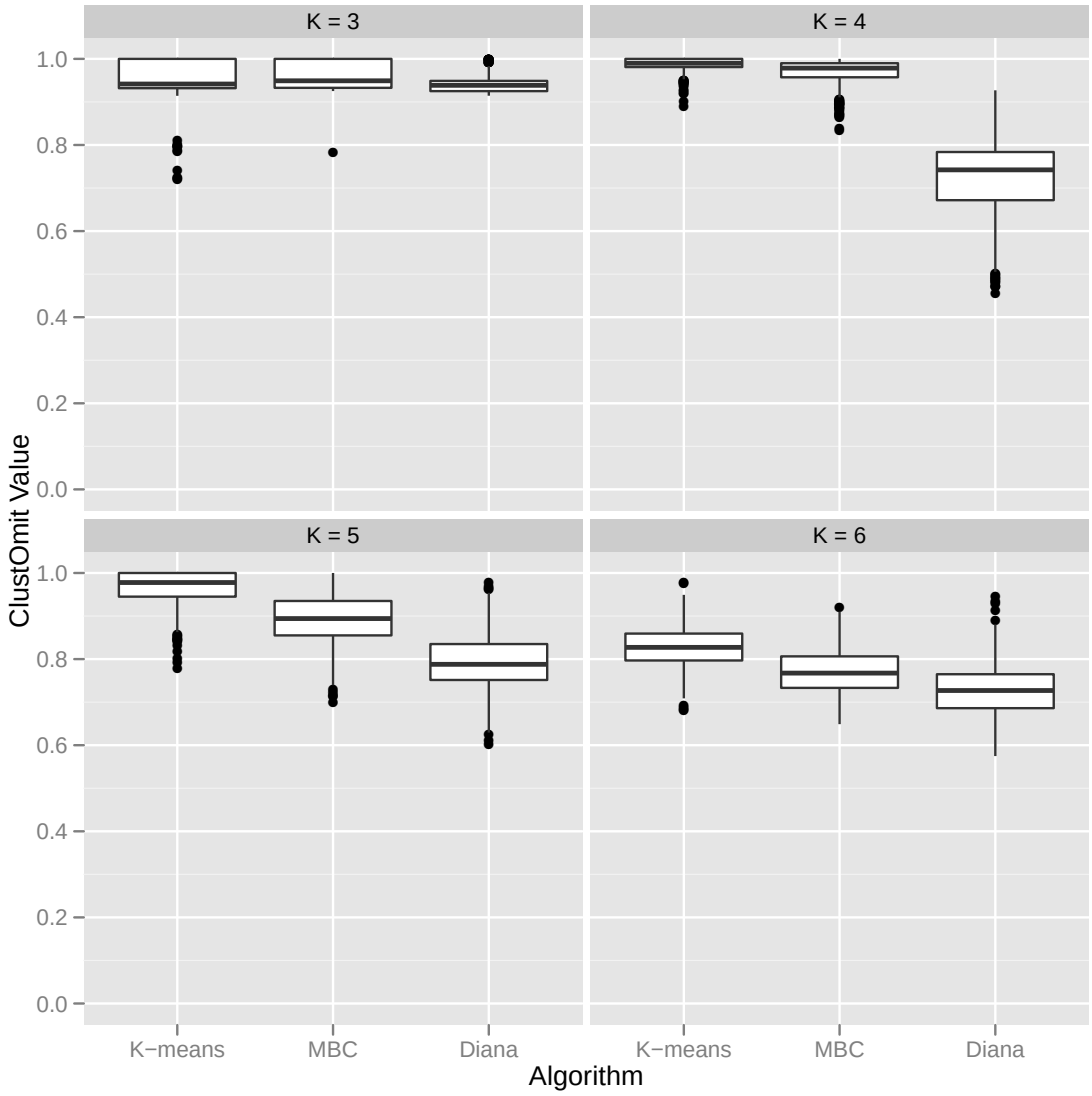

Figure 2.6: Box plots of the *ClustOmit* statistic obtained from the Khan data set for $B = 500$.

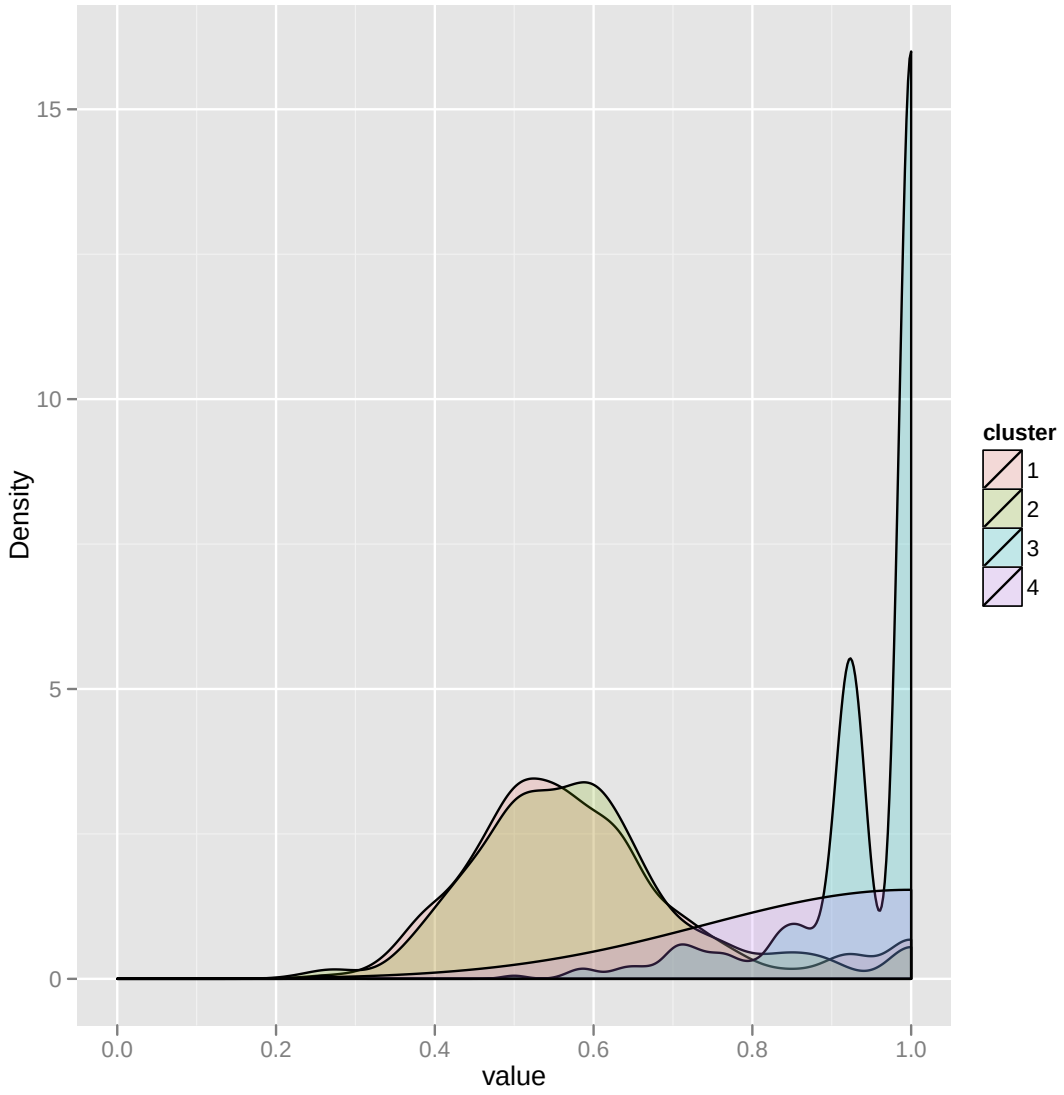

Figure 2.7: Density plots of the Hennig statistic applied to the Khan data set for the Diana algorithm for $K = 4$ and $B = 500$.

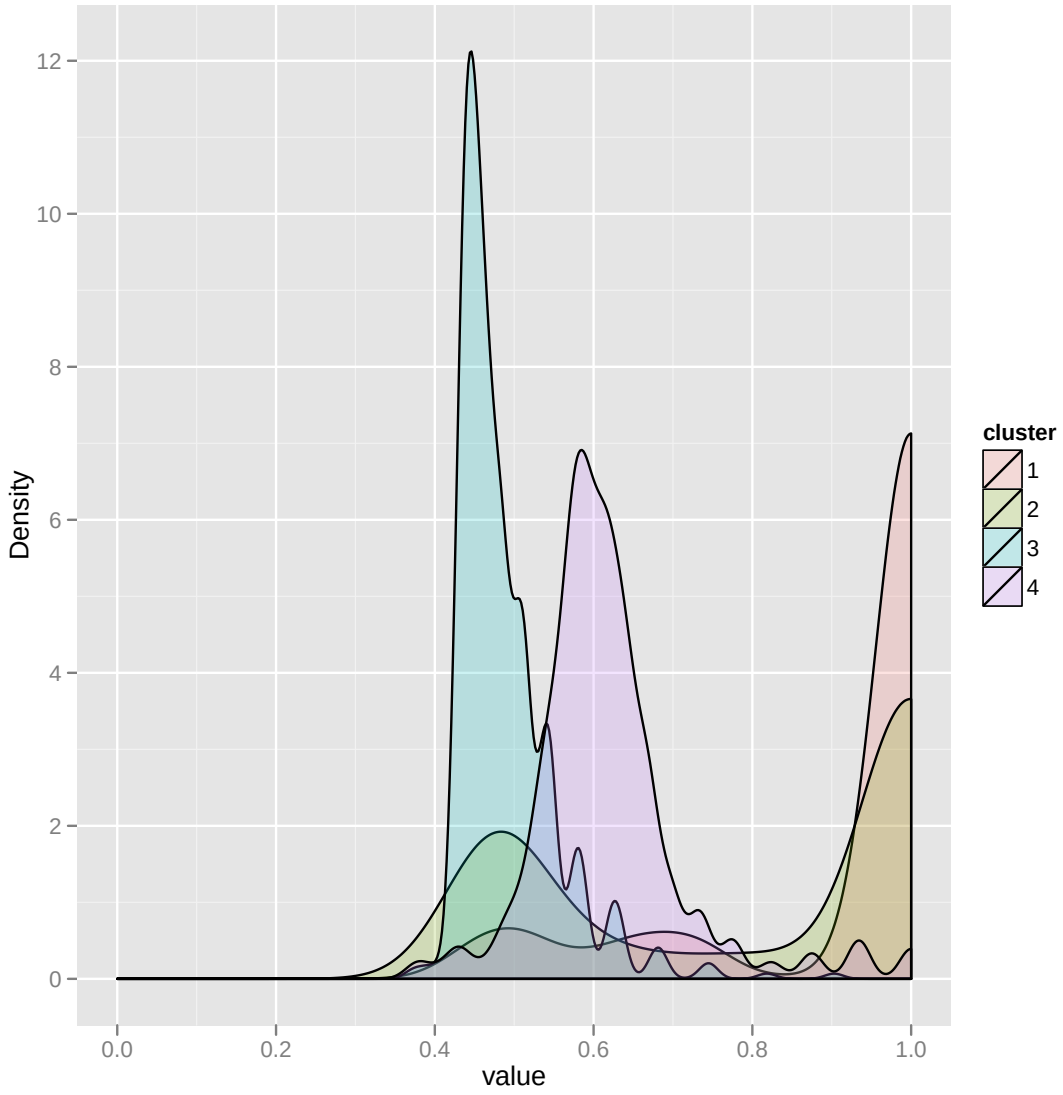

Figure 2.8: Density plots of the *ClustOmit* statistic applied to the Khan data set for the Diana algorithm for $K = 4$ and $B = 500$.

CHAPTER THREE

On Model Selection with Regularized Discriminant Analysis

3.1 Introduction

Friedman [1989] has proposed the *RDA* classifier as an alternative to the well-known linear discriminant analysis (*LDA*) and quadratic discriminant analysis (*QDA*) classifiers by incorporating a biased covariance matrix estimator that has been shown to improve classification performance. The *RDA* classifier incorporates two tuning parameters that are typically estimated when we minimize an empirical loss function, such as a conditional error-rate estimator. Cross-validation is typically employed by default to estimate the training error rate in the selection of tuning parameters with supervised classification models (hereafter, *classifiers*). Friedman has suggested that the *LOO* error rate estimator be used for model selection with the *RDA* classifier. However, the *LOO* estimator is well-known to exhibit large variance [Izenman, 2008] and can yield ill-advised model selections because multiple tuning parameter pairs can yield the minimum *LOO* error-rate estimate [Aeberhard, Coomans, and Vel, 1993].

Currently, our choice of conditional error-rate estimator for model selection with the *RDA* classifier is unclear. We consider five alternative conditional error-rate estimators to the *LOO* estimator to improve the *RDA* model selection. For a thorough listing of additional proposed parametric and nonparametric error-rate estimators, we recommend Schiavo and Hand [2000], Hand [1997], Molinaro, Simon, and Pfeiffer [2005], Toussaint [1974], and Wehberg and Schumacher [2004].

Additionally, other empirical loss functions have been considered for model selection for the *RDA* classifier. For instance, Aeberhard et al. [1993] have claimed that their appreciation function yields improved model selection in terms of classifi-

cation performance for the *RDA* classifier because their proposed function does not rely on error counting. Instead, Aeberhard et al. utilize estimators of the *a posteriori* probabilities of class membership to provide a *smooth* empirical loss function. However, we do not consider such functions in this paper. Instead, we consider only conditional error-rate estimators for *RDA* model selection.

Using four small-sample, high-dimensional microarray data sets, we compare the *RDA* model selection with the *LOO* estimator, the 10-fold *CV* estimator, the 632 and 632+ estimators, the bootstrap estimator considered by Jain, Dubes, and Chen [1987] and Efron and Tibshirani [1994], and the *BCV* estimator from Fu et al. [2005]. Based on our Monte Carlo simulations, we prefer the 10-fold *CV* estimator for *RDA* model selection. Although its classification performance is comparable to the model selection with the *LOO*, bootstrap, and *BCV* estimators, we remark that the 10-fold *CV* estimator is less computationally demanding. Furthermore, we comment that the bootstrap and *BCV* estimators yield the smallest model selection variability among the considered error-rate estimators, but these two estimators require substantially more computation than the competing error-rate estimators. Thus, if classification performance is our main goal, we recommend the 10-fold *CV* estimator. However, if we also desire small variability in model selection and if the additional computation is warranted, we recommend the *BCV* estimator for *RDA* model selection in the small-sample, high-dimensional setting.

We have organized the remainder of the paper as follows. In Section 2 we review the *RDA* classifier from Friedman [1989]. In Section 3 we describe the model selection techniques for the *RDA* classifier and then discuss error-rate estimators that we examine in Section 4. In Section 5 we briefly describe four small-sample, high-dimensional microarray data sets, our Monte Carlo simulation design, and the simulation results. We then briefly discuss the results in Section 6.

3.2 Regularized Discriminant Analysis

In discriminant analysis, also known as supervised learning, we wish to correctly assign an unlabeled p-dimensional observation vector $\boldsymbol{x}$ to one of K unique, known classes (or populations) by constructing a classifier from n training observations that can accurately predict the class membership of $\boldsymbol{x}$. Let $\mathcal{T} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ denote a training data set, and let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}_{p\times 1}$ be the ith observation ($i = 1, \ldots, n$), where $\mathbb{R}_{m\times n}$ denotes the matrix space of all $m \times n$ matrices over the real field $\mathbb{R}$. We assume that $(\boldsymbol{x}_i, y_i)$ is a realization from a mixture distribution $p(\boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x}|\omega_k)p(\omega_k)$, where $p(\boldsymbol{x}|\omega_k)$ is the probability density function (PDF) of the kth class, $p(\omega_k)$ is prior probability of class membership of the kth class, and $y_i \in \mathcal{K} = \{\omega_1, \ldots, \omega_K\}$ denotes the true, unique membership of sample $\boldsymbol{x}_i$.

If we assume that the K distributions are multivariate normal with known parameters, the optimal Bayesian classifier with respect to a $0-1$ loss function is the well-known *QDA* classifier. We say that the kth class consists of p-dimensional multivariate normal vectors with the PDF

$$p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right\}, \tag{3.1}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}_{p\times 1}$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{>}_{p\times p}$ are the known mean vector and covariance matrix, respectively, for class ω_k, where $\mathbb{R}^{>}_{m\times m}$ denotes the cone of real $n \times n$ positive definite matrices.

In practice, we estimate the unknown parameters with their maximum likelihood estimators (MLEs). With the MLEs for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, we assign an unlabeled observation $\boldsymbol{x}$ to class ω_k using the following decision rule:

$$D(\boldsymbol{x}) = \arg\min_k (\boldsymbol{x} - \bar{\boldsymbol{x}}_k)'\widehat{\boldsymbol{\Sigma}}_k^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_k)' + \ln|\widehat{\boldsymbol{\Sigma}}_k| - 2\ln P(\omega_k), \tag{3.2}$$

where $\bar{\boldsymbol{x}}_k$ and $\widehat{\boldsymbol{\Sigma}}_k$ denote the MLEs for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, respectively.

If we assume that the covariance matrix parameters are equal for each class (i.e., $\boldsymbol{\Sigma}_k \equiv \boldsymbol{\Sigma}$) in (3.1), then (3.2) simplifies to the *LDA* classifier,

$$D(\boldsymbol{x}) = \arg\min_k (\boldsymbol{x} - \bar{\boldsymbol{x}}_k)' \widehat{\boldsymbol{\Sigma}}_{pool}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_k)' - 2 \ln P(\omega_k), \tag{3.3}$$

where $\widehat{\boldsymbol{\Sigma}}_{pool}$ is the pooled sample covariance matrix MLE with

$$\widehat{\boldsymbol{\Sigma}}_{pool} = \frac{1}{n} \sum_{k=1}^{K} n_k \widehat{\boldsymbol{\Sigma}}_k. \tag{3.4}$$

To estimate the covariance matrices well, we require a large number of observations relative to the dimension p. Recall that the spectral decomposition of $\widehat{\boldsymbol{\Sigma}}_k^{-1}$ is given by

$$\widehat{\boldsymbol{\Sigma}}_k^{-1} = \sum_{j=1}^{p} \boldsymbol{v}_j \boldsymbol{v}_j' / e_j,$$

where e_j is the jth largest eigenvalue of $\widehat{\boldsymbol{\Sigma}}_k$ and $\boldsymbol{v}_j$ is the associated eigenvector [Harville, 2008]. We remark that the smallest eigenvalues and the directions associated with their eigenvectors highly influence the estimator of $\boldsymbol{\Sigma}_k^{-1}$. The eigenvalues of $\widehat{\boldsymbol{\Sigma}}_k^{-1}$ are well known to be biased such that the smallest eigenvalues are underestimated [Seber, 2004] and that this bias increases as the training-sample size n decreases relative to the feature dimensionality p. Consequently, $\widehat{\boldsymbol{\Sigma}}_k^{-1}$ used in (3.2) is highly variable for small values of n/p. Moreover, for $p > n$, (3.2) is incalculable because $\widehat{\boldsymbol{\Sigma}}_k^{-1}$ does not exist. Thus, although more feature information is available to discriminate among the K classes as p increases, classification accuracy decreases unless one obtains enough training-sample observations to reliably estimate the increased number of parameters.

Several regularization methods, such as Guo, Hastie, and Tibshirani [2007], Mkhadri [1995], and Xu, Brock, and Parrish [2009], have been proposed in the literature to stabilize the eigenvalues of the sample covariance matrices used in (3.2). A typical regularized sample covariance matrix resembles the ridge estimator

$$\widehat{\boldsymbol{\Sigma}}_k(\gamma) = \widehat{\boldsymbol{\Sigma}}_k + \gamma \boldsymbol{I}_p, \tag{3.5}$$

where $\boldsymbol{I}_p \in \mathbb{R}_{p \times p}$ is the identity matrix and γ is a positive scalar that must be estimated with training data. This technique effectively *shrinks* the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_k$ towards $\boldsymbol{I}_p$ and increases the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_k$ away from zero.

Friedman [1989] has extended the shrinkage technique by first computing a weighted average of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_k$ for class ω_k and the pooled sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{pool}$ to estimate the covariance matrix for class ω_k with $\widehat{\boldsymbol{\Sigma}}_k(\lambda)$, where

$$\begin{aligned} n_k(\lambda) &= (1-\lambda)n_k + \lambda n, \\ \boldsymbol{S}_k &= n_k \widehat{\boldsymbol{\Sigma}}_k, \\ \boldsymbol{S} &= \sum_{k=1}^{K} \boldsymbol{S}_k, \\ \boldsymbol{S}_k(\lambda) &= (1-\lambda)\boldsymbol{S}_k + \lambda \boldsymbol{S}, \\ \widehat{\boldsymbol{\Sigma}}_k(\lambda) &= \boldsymbol{S}_k(\lambda)/n_k(\lambda) \end{aligned} \tag{3.6}$$

with $\lambda \in [0,1]$. We can interpret (3.6) as a covariance matrix estimator for class ω_k that borrows from (5.3) to better estimate $\boldsymbol{\Sigma}_k$. Notice that for $\lambda = 0$, (3.6) corresponds to the covariance matrix estimator used in (3.2). Also, notice that for $\lambda = 1$, (3.6) corresponds to the covariance matrix estimator used in (5.4), in which we implicitly assume that $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_K$.

Friedman [1989] has also used the regularization parameter γ to shrink the eigenvalues of (3.6) towards the p-dimensional identity matrix in order to stabilize the inverse of (3.6), resulting in the biased covariance matrix estimator

$$\widehat{\boldsymbol{\Sigma}}_k(\lambda,\gamma) = (1-\gamma)\widehat{\boldsymbol{\Sigma}}_k(\lambda) + \gamma \frac{\mathrm{tr}\{\widehat{\boldsymbol{\Sigma}}_k(\lambda)\}}{p}\boldsymbol{I}_p, \tag{3.7}$$

for class ω_k where $\mathrm{tr}\{\cdot\}$ is the trace operation and $\gamma \in [0,1]$. Notice that for $\gamma = 0$ we have a scalar times the identity matrix $\boldsymbol{I}_p$, which corresponds to a modified Euclidean classifier [Marco, Young, and Turner, 1987].

Finally, if we substitute (3.7) in place of $\widehat{\boldsymbol{\Sigma}}_k$ in (3.2), we obtain the *RDA* classifier

$$D_k(\boldsymbol{x}) = \arg\min_k (\boldsymbol{x} - \bar{\boldsymbol{x}}_k)' \widehat{\boldsymbol{\Sigma}}_k(\lambda, \gamma)^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_k)' + \ln |\widehat{\boldsymbol{\Sigma}}_k(\lambda, \gamma)| - 2 \ln P(\omega_k), \quad (3.8)$$

In summary, the *pooling* parameter λ controls the amount that we borrow from $\widehat{\boldsymbol{\Sigma}}_{pool}$, and the *shrinkage* parameter γ determines the amount of applied shrinkage.

3.3 Model Selection for the RDA Classifier

A main disadvantage of the *RDA* classifier discussed in the literature is model selection because no closed-form estimators for λ and γ are available. Friedman [1989] has suggested that a training error-rate estimator be computed for each parameter pair in a grid of candidate values of (λ, γ). We construct the model selection grid as the Cartesian product of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_G)'$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_G)'$, where γ_i, $\lambda_i \in [0, 1]$ for $i = 1, 2, \ldots, G$. We select the pair $(\widehat{\lambda}, \widehat{\gamma})$ that minimizes the considered training error-rate estimator.

The grid model selection method is computationally intensive because we must estimate the training error rate for each candidate pair (λ, γ). However, this procedure is an *embarrassingly parallel* computational situation for which modern parallel programming techniques are readily available to reduce the length of the computational runtime [Foster, 1995]. Rather than employing parallel computation, Friedman [1989] has derived a "down-dating" method that alleviates much of the computational burden when one is computing the *LOO* error-rate estimator for each parameter pair in the *RDA* model grid. This characteristic suggests traversing through different values of γ for a fixed value of λ because the down-dated formula is independent of γ.

Because we must calculate the training error rate for each grid pair, Ye and Wang [2006] view this process as prohibitive for a large number of candidate parameter values. Hence, we must consider the precision of the grid to find an estimator

that results in small error rate. Arguably we should use a large grid size and consider a large number of values for (λ, γ) for model selection. However, by increasing the grid precision, we will considerably increase the amount of computation.

3.4 Conditional Error-Rate Estimators

In this section, we discuss various conditional error rate estimators that we will compare for model selection with the *RDA* classifier. First, we provide some necessary notation to discuss the conditional error-rate estimators in detail. We borrow extensively from the notation of Hastie, Tibshirani, and Friedman [2008].

Let f be a decision function that maps an unlabeled observation $\boldsymbol{x} \in \mathbb{R}_{p\times 1}$ to its predicted class label, $y \in \mathcal{K}$. We say that the pair $(\boldsymbol{x}, y)$ is sampled from $\mathcal{F}$, the joint distribution of the data. From a training data set, $\mathcal{T}$, we train an estimator $\widehat{f}$ of the decision function f to obtain a sample-based classifier. The *RDA* classifier in (3.8) is an estimator for the *QDA* decision rule in (3.2).

We require a loss function $L\{y, f(\boldsymbol{x})\}$ for penalizing classification errors. Following the typical approach in supervised classification studies, we consider the $0-1$ loss function $L\{y, f(\boldsymbol{x})\} = I[f(\boldsymbol{x}) = y]$, where $I[z]$ is 1 if the Boolean statement z is true and 0, otherwise. Next, we define the conditional (test) error rate (CER) of the classifier, f, as

$$CER = E_{\boldsymbol{x}^0, y^0}[L\{y^0, f(\boldsymbol{x}^0)\}|\mathcal{T}], \tag{3.9}$$

where the pair $(\boldsymbol{x}^0, y^0)$ is a test observation sampled from $\mathcal{F}$, and $E_{\boldsymbol{z}}$ denotes the expectation with respect to the distribution of the random vector $\boldsymbol{z} \in \mathbb{R}_{p\times 1}$. We remark that for (3.9), the training data set $\mathcal{T}$ is fixed. By averaging over all training data sets, we obtain the expected error rate (EER) as

$$\begin{aligned} EER &= E_{\mathcal{T}} E_{\boldsymbol{x}^0, y^0}[L\{y^0, f(\boldsymbol{x}^0)\}|\mathcal{T}] \\ &= E_{\mathcal{T}}[CER]. \end{aligned} \tag{3.10}$$

The majority of error-rate estimators estimate (3.10) by partitioning the data into training and test data sets. However, these estimators are often used to estimate (3.9).

Some classifiers, such as the *RDA* classifier, are indexed by the tuning parameters $\boldsymbol{\theta}$. Thus, we write the trained classifier as $\widehat{f}(\boldsymbol{x}, \boldsymbol{\theta})$. We then estimate $\boldsymbol{\theta}$ by selecting $\widehat{\boldsymbol{\theta}}$, which yields the minimum estimate of (3.9). For brevity, we write $\widehat{f}(\boldsymbol{x}) = \widehat{f}(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})$ in our discussion of error rate estimators below.

3.4.1 The Apparent Error Rate (AER) Estimator

The AER estimator is perhaps one of the simplest conditional error-rate estimators because it is the proportion of misclassified observations with the usage of the original data set as both the training and test data set. We write the AER estimator as

$$\overline{\text{err}} = \sum_{i=1}^{n} L(y_i, \widehat{f}(\boldsymbol{x}_i)). \tag{3.11}$$

The AER estimator is overly optimistic because the classifier is first adapted to the training data and then predictions are performed with the same data set. Hence, (3.11) generally estimates (3.9) with a downward bias. Although we do not consider (3.11) in our simulation studies, we emphasize it here because a subset of the CER estimators we consider here functionally depend on it.

3.4.2 The M-fold Cross-validation (MCV) Estimator

A reasonable attempt to overcome the downward bias of (3.11) is to partition the original training data into M mutually exclusive and exhaustive sets, also known as folds, that have approximately the same number of observations. Then, for $m = 1, \ldots, M$, we classify the observations in the mth fold by training a classifier on the remaining $M-1$ folds. We then calculate the proportion of misclassified observations across the M folds to obtain the MCV estimator of (3.9).

Next, we define the MCV estimator more precisely. Similar to Hastie et al. [2008], we define the indexing function $\kappa : \{1, \ldots, n\} \to \{\omega_1, \ldots, \omega_K\}$ to randomly assign the ith observation to the mth fold. Let $\widehat{f}^{-m}(\boldsymbol{x})$ be the trained classifier on the observations the mth fold omitted from the data. Then, we write the MCV estimator as

$$\widehat{\mathrm{Err}}^{(CV)} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \widehat{f}^{-\kappa(i)}(\boldsymbol{x}_i)). \tag{3.12}$$

In the literature, we have seen that typical choices of M are 5, 10, and n. We note that the case $M = n$ corresponds to the LOO estimator. In this case, we have $\kappa(i) = i$, so that the classifier is trained on every observation except $\boldsymbol{x}_i$. The LOO estimator is well known to be an approximately unbiased estimator for the EER but can exhibit a large variance because of the similarity of the $M = n$ training sets [Hastie et al., 2008]. For other choices of M, we can significantly reduce the variance of the MCV estimator, although we have a trade-off of adding bias to the estimator. Despite the common choices for M that we have seen in the literature, we argue that a *best* value for M is difficult to identify because the choice depends on the classifier, the training sample size, and the distributions from which each population is realized. We consider $M = 10$ in our simulation studies because it is commonly used in the literature and is relatively quick to compute.

3.4.3 The Bootstrap Estimator

Rather than partitioning the data set as with the MCV estimator, many estimators utilize the bootstrapping paradigm first to sample with replacement from the data set and then to classify the unsampled training observations. Here, we consider one of the simplest bootstrapping error-rate estimators, which has been considered by Jain et al. [1987] and Efron and Tibshirani [1994]. Let $\widehat{f}^{*b}(\boldsymbol{x}_i)$ be our

classification of $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ for the bth bootstrap sample with $b = 1, \ldots, B$. Then, we define

$$\widehat{\text{Err}}_{boot} = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n} L(y_i, \widehat{f}^{*b}(\boldsymbol{x}_i)). \tag{3.13}$$

to be the bootstrap error rate estimator. Notice that the training and test data sets overlap. Hence, (3.13) is a downward-biased estimator of the EER that can be used to estimate the CER. To overcome the downward bias, Efron and Tibshirani [1997] have proposed the leave-one-out bootstrap ($LOO\text{-}Boot$) error-rate estimator. Contrary to (3.13), the $LOO\text{-}Boot$ error rate estimator uses only the observations that are not sampled with replacement in a bootstrap sample as the test observations. Let

$$C^{-i} \equiv \{b \in \{1, \ldots, B\} \mid \boldsymbol{x}_i \notin \boldsymbol{Z}^{*b}\}.$$

We write the $LOO\text{-}Boot$ estimator as

$$\widehat{\text{Err}}^{(LOO-Boot)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \widehat{f}^{*b}(\boldsymbol{x}_i)). \tag{3.14}$$

For small values of B, $|C^{-i}| = 0$ may hold for at least one $i = 1, \ldots, n$. For example, consider the case $B = 1$. Then, approximately 36.8% of the observations are not sampled. Hence, $|C^{-i}| = 0$ for each i such that $\boldsymbol{x}_i \notin \boldsymbol{Z}^{*1}$. Therefore, we must choose a sufficiently large B.

3.4.4 The .632 Estimator

Although (3.14) corrects for the over-optimism of (3.13), it is susceptible to a biased estimation of EER similar to the MCV estimator because a subset of the training data set might not be fully representative of the data-generating populations. However, as the sample size of the data increases, the training data can be considered more representative of the underlying populations, and therefore, the bias of (3.13) is typically smaller.

For each bootstrap sample, we have that the average number of distinct observations is approximately $0.632n$ because

$$Pr(\boldsymbol{x}_i \in \boldsymbol{Z}^{*b}) = 1 - \left(1 - \frac{1}{n}\right)^n$$
$$\approx 0.632. \tag{3.15}$$

For a fixed value of $b = 1, \ldots, B$, we note that the $Pr(\boldsymbol{x}_i \in \boldsymbol{Z}^{*b})$ rapidly approaches 0.632 as n increases. For instance, if we consider the values of $n = 10$, 20, and 30, then we have that the probabilities that an observation is any one bootstrap sample are approximately 0.651, 0.642, and 0.638, respectively. From (3.15), we can conclude that the estimation bias of (3.14) is similar to two-fold cross-validation because we can expect approximately half of the original training observations to be sampled in each bootstrap sample. Hence, (3.14) is likely to yield upward bias. To correct for the bias, Efron and Tibshirani [1994] have proposed the .632 estimator

$$\widehat{\text{Err}}^{(.632)} = 0.368 \cdot \overline{\text{err}} + 0.632 \cdot \widehat{\text{Err}}^{(LOO-Boot)}, \tag{3.16}$$

which is a weighted average of (3.11) and (3.14). Intuitively, we see that (3.16) corrects for the upward bias of (3.14) by including the downward bias of (3.11). We note that the coefficients in (3.16) sum to unity and have been chosen corresponding to (3.15).

3.4.5 The .632+ Estimator

While (3.16) has been shown to estimate the EER well when the specified classifier does not overfit the training data, it is often biased when the trained classifier overfits the training data. Efron and Tibshirani [1997] have proposed the .632+ estimator, which is intended to correct the bias of (3.16) in cases of overfitting. Efron and Tibshirani have defined the *no-information error rate*, η, as the error rate

if the feature vectors and the class labels are independent. Efron and Tibshirani have estimated η by

$$\widehat{\eta} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} L(y_i, \widehat{f}(\boldsymbol{x}_{i'})), \tag{3.17}$$

which evaluates the classifier at all possible combinations of the class labels y_i and the feature vectors $\boldsymbol{x}_{i'}$ $(1 \leq i, i' \leq n)$. Alternatively, we have the equivalent and more intuitive expression

$$\widehat{\eta} = \sum_{k=1}^{K} \widehat{p}_k (1 - \widehat{q}_k),$$

where $\widehat{p}_k$ is the observed proportion of observations that are members of class ω_k and $\widehat{q}_k$ is the proportion of observations that are classified into class ω_k $(k = 1, \ldots, K)$. The computed value for $\widehat{\eta}$ is invariant to the ordering of the classes.

Next, Efron and Tibshirani have defined the *relative overfitting rate* as

$$\widehat{R} = \frac{\widehat{\text{Err}}^{(LOO-Boot)} - \overline{\text{err}}}{\widehat{\eta} - \overline{\text{err}}}, \tag{3.18}$$

which is defined on $[0, 1]$. If $\widehat{R} = 0$, which implies that $\widehat{\text{Err}}^{(LOO-Boot)} = \overline{\text{err}}$, then we say that overfitting has occurred. Similarly, if $\widehat{R} = 1$, then we have that the $\widehat{\text{Err}}^{(LOO-Boot)} = \overline{\text{err}}$, which again suggests that the classifier has overfit the data.

Efron and Tibshirani [1997] have defined the .632+ estimator as

$$\widehat{\text{Err}}^{(.632+)} = (1 - \widehat{w}) \cdot \overline{\text{err}} + \widehat{w} \cdot \widehat{\text{Err}}^{(LOO-Boot)}, \tag{3.19}$$

where the weight $\widehat{w} = \frac{.632}{1 - .368\widehat{R}} \in [0.632, 1]$. Hence, the values of (3.19) range from $\widehat{\text{Err}}^{(.632)}$ when $\widehat{w} = .632$ to $\widehat{\text{Err}}^{(LOO-Boot)}$ when $\widehat{w} = 1$. Similar to (3.16), (3.19) yields a compromise between (3.14) and (3.11), but (3.19) is intended also to account for possible overfitting.

3.4.6 The Bootstrap Cross-Validation (BCV) Estimator

Fu et al. [2005] have discussed that (3.12) can exhibit large variability and poor estimation of the EER for small sample sizes. Fu et al. have reasoned that, due to

the small sample size, the data are not adequately representative of the underlying distribution, $\mathcal{F}$, so that the distance between the partitioned training and test data sets is often large. Fu et al. have proposed the BCV error-rate estimator to improve (3.12) by combining (3.12) and (3.13) under the bagging framework [Breiman, 1996]. Formally, we write the BCV estimator as

$$\widehat{\mathrm{Err}}^{(BCV)} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{n}\sum_{i=1}^{n} L(y_i^*, \widehat{f}^{-\kappa_b(i)}(\boldsymbol{x}_i^*)). \tag{3.20}$$

To compute (3.20), one randomly draws B bootstrap samples from $\mathcal{D}$ and then computes (3.12) on the bth sample for $b = 1, \ldots, B$. Fu et al. [2005] have argued that (3.20) estimates EER well because the held-out fold of the bth bootstrap sample is reasonably close to the $M-1$ remaining folds upon which the classifier is trained. They have also argued that (3.20) is superior to (3.14) because the latter estimator counts only errors from test data sets that have no overlap with training data sets. Moreover, the authors have argued that (3.20) estimates the EER better than (3.16) and (3.19) because the latter two estimators heavily weight (3.14).

Fu et al. have recommended that B be between 50 and 200 and have required that at least three distinct observations be present in each class of the bootstrapped held-out folds. Because (3.20) is computationally burdensome if any of n, M, or B is large, we choose $M = 10$ and $B = 100$ in our simulation study below.

3.5 Monte Carlo Simulation Design and Results

Our goal in this paper is to compare the model selection of the *RDA* classifier using the conditional error-rate estimators described in Section 4 with small-sample, high-dimensional data sets. We compare the model selection methods in two ways. First, we wish to identify the error-rate estimators that yield the smallest variability in terms of the selected models for data generated from the same joint distribution $\mathcal{F}$. Ideally, an error-rate estimator will consistently result in the same model. Second, we compare the classification performance of the models selected using each error

rate estimator. We desire to know if one of the error-rate estimators described in Section 4 yields the best model in terms of the average classification performance.

To study our two stated goals, we consider four well-known, small-sample, high-dimensional microarray data sets. We repeat the following procedure 250 times. First, we randomly partition the given data set $\mathcal{D}$ into a training data set, $\mathcal{T}$, and a validation (test) data set, $\mathcal{V}$, such that $\mathcal{D} = \mathcal{T} \cup \mathcal{V}$ and $\mathcal{T} \cap \mathcal{V} = \emptyset$. We partition $\mathcal{D}$ so that $\mathcal{T}$ contains four-fifths of the observations, while $\mathcal{V}$ contains the remaining one-fifth of the observations. We preserve the proportions of each class as observed in $\mathcal{D}$. We also reduce the dimension of the observations before applying the *RDA* classifier because the log-determinant in (3.8) is numerically unstable and is often incalculable when shrinkage is applied. We apply the variable selection method from Dudoit, Fridlyand, and Speed [2002] to $\mathcal{D}$ before randomly partitioning $\mathcal{D}$ to reduce the data from p features to $q = 50$ features. We are aware of the classification error rate bias that can be induced [Ambroise and McLachlan, 2002], but following Fu et al. [2005], we do not consider the bias here because our focus is not on gene selection.

Next, we train the *RDA* classifier, $\widehat{f}$, using the observations in $\mathcal{T}$. For each candidate model, using a parameter grid with $G = 11$, we compute each of the considered error-rate estimators and select the model that yields the minimum CER. As discussed by Zhou and Mao [2006], an error rate estimator can yield multiple models with the minimum CER (i.e., ties are present), in which case we randomly choose the model from multipe best models. We then calculate CER as the proportion of observations in $\mathcal{V}$ that we misclassify with the trained classifier $\widehat{f}(\boldsymbol{x})$. For each error-rate estimator, we report the selected model $\widehat{\boldsymbol{\theta}}$ along with EER.

Here, we briefly describe four microarray data sets analyzed in our Monte Carlo simulations.

3.5.1 Alon Data Set

Alon et al. [1999] have examined the gene expression profiles measured with an Affymetrix oligonucleotide array for $n_1 = 40$ tumor and $n_2 = 22$ normal colon tissues for 6,500 human genes. Following Alon et al. [1999], we restrict the data set to the 2,000 genes with the highest minimal intensity across the samples.

3.5.2 Chiaretti Data Set

Chiaretti et al. [2004] have presented a data set that contains the gene expression levels for 128 individuals with acute lymphoblastic leukemia (ALL). As in Xu et al. [2009], we consider the $n_1 = 42$ observations labeled *NEG* and the $n_2 = 37$ observations labeled *BCR/ABL*. For each of the $n = 79$ observations that have been obtained from Affymetrix human 95Av2 arrays, the robust multichip average (*RMA*) normalization method has been applied to all 12,625 gene expression levels.

3.5.3 Golub Data Set

Golub et al. [1999] have examined the gene expression levels for $n_1 = 47$ patients with acute lymphoblastic leukemia (ALL) and $n_2 = 25$ patients with acute myeloid leukemia (AML). Bone marrow samples from each patient were assayed using Affymetrix Hgu6800 chips. We use the merged version of the data set that is available in the `golubEsets` package on Bioconductor.

3.5.4 Singh Data Set

Singh et al. [2002] have presented a data set that consists of 235 radical prostatectomy specimens from surgery patients between 1995 and 1997, and oligonucleotide microarrays containing probes for approximately 12,600 genes and expressed sequence tags were used. We consider 102 of the radical prostatectomy specimens that the authors reported as high-quality to obtain a data set consisting of $n_1 = 52$ prostate tumor samples and $n_2 = 50$ non-tumor prostate samples.

3.5.5 Simulation Results

For the microarray data sets, we computed EERs for the selected *RDA* models with respect to each of the six error-rate estimators described in Section 4. In Table 3.1 we summarized the EER estimates (i.e., the average of the CER estimates) along with approximate standard errors for each model selection method. In Figures 1-4, we plot heatmaps of the 250 *RDA* models that were selected with respect to each of the considered error-rate estimators for the four microarray data sets. Furthermore, in Figures 5-8, we plot average training error rates for each of the six error-rate estimators for each *RDA* models considered.

In Table 3.1 we see that the models selected with the *LOO*, 10-fold *CV*, bootstrap, and *BCV* estimators yielded similar EER estimates and estimated standard errors. The models selected with the .632 and .632+ estimators yielded much larger EER estimates with increased estimated standard errors. For instance, with the Alon data set, the difference in classification performance of the *RDA* classifier between the *RDA* model selection with the .632 estimator and the 10-fold *CV* estimator was 0.235; notice the larger estimated standard error for the .632 estimator compared to the 10-fold *CV* estimator.

With the Golub data set, the *RDA* models selected with the .632 and .632+ estimators were similar. These models yielded EER estimates that were 0.148 larger than the models selected with the *LOO* and 10-fold *CV* estimators. Once again, the estimated standard errors of the EER estimates were larger for the .632 and .632+ error rate estimators than with the *LOO* and 10-fold *CV* estimators.

The Chiaretti data set was the only case where each of the model selection methods yielded similar EER estimates. However, the estimated standard errors of the EER estimates obtained for the models selected using the .632 and .632+ estimators are larger here than the estimated standard errors of the competing error-rate estimators.

Recall that our goal was to identify the error-rate estimators that yielded small variability in the selected *RDA* models. Consider the selected models for the Chiaretti data set in Figure 3.2. We can see that the models selected using the bootstrap and *BCV* error-rate estimators were concentrated about the parameters $\widehat{\lambda} = 0.2$ and $\widehat{\gamma} = 0.4$. Hence, although the training data sets are different for each of the 250 Monte Carlo iterations, we can be confident in our selected *RDA* model when employing it with the Chiaretti data set. We contrast this instance with the models selected with the .632+ estimator. Although we can see some concentration about the *RDA* model corresponding to $\widehat{\lambda} = 0.2$ and $\widehat{\gamma} = 0.4$, the .632+ estimator exhibited large variability in the selected *RDA* models. Moreover, using the .632+ estimator, we often selected $\widehat{\lambda} = 0$ and $\widehat{\gamma} = 0$, which correspond to the *QDA* classifier, but in other instances we selected $\widehat{\lambda} = 1$ and $\widehat{\gamma} = 0$, which correspond to the *LDA* classifier.

We argue that this inconsistency in classifier model selection using the .632 and .632+ estimators is problematic in practice. For the above example, we ask, "Should we employ the *LDA* classifier, the *QDA* classifier, or the *RDA* model corresponding to $\lambda = 0.2$ and $\gamma = 0.4$?" If we were to employ the .632+ estimator for model selection with the *RDA* classifier in practice, our selection would be unclear, due to the large variability. Contrarily, if we were to use the bootstrap or *BCV* estimators instead, we could be much more confident due to the smaller variability in Figure 3.2. Hence, for the Chiaretti data set, we prefer the bootstrap and *BCV* estimators, especially because they yielded EER estimates that were comparable to the *LOO* and 10-fold *CV* estimates.

Here, we examine the poor classification performance of the models selected with the .632 and .632+ estimators for the Golub data set. First, in Figure 3.3 we notice that the *LOO*, 10-fold *CV*, bootstrap, and *BCV* estimators resulted primarily in models with $\widehat{\gamma} = 0.1$, which yielded excellent classification performance. Contrarily, we see that the .632 and .632+ estimators resulted in chosen models with either

$\widehat{\gamma} = 0.0$ or $\widehat{\gamma} = 0.1$ with the latter being selected more frequently, yielding larger EER estimates and standard errors. In Figure 3.7, we gain additional insight about the computed .632 and .632+ training error-rate estimates for the Golub data set. Notice that the models for $\widehat{\gamma} \geq 0.2$ yielded relatively large training error estimates for each error-rate estimator and were not selected. Also, notice that the subset of models that have the smallest training error-rate estimates using the *LOO*, 10-fold *CV*, bootstrap, and *BCV* estimators yielded regions that were distinctly different from those of other possible models.

Next, we consider the average training error grids for the Alon data set in Figure 3.5. We can see that the average of the training error rates corresponding to the .632 and .632+ estimators were similar for the candidate models. However, notice that the *LOO* and 10-fold *CV* estimators yielded smaller regions with lower training error rates. Hence, the *LOO* and 10-fold *CV* estimators resulted in models from these smaller regions. The average training error rates with respect to the .632 and .632+ estimators were similar for all considered models. That is, the difference in the training error rates among the possible models with respect to the .632 and .632+ estimators was small. Thus, the chosen models were not immediately obvious, which, in the case of the Alon data set, yielded larger CER estimates with greater variability. In other words, we have evidence that the small difference in the .632 estimates for these *RDA* models yielded larger EER estimates and estimated standard errors.

However, the impact on classification performance was not as pronounced for the Chiaretti data set. In Figure 3.6 we see that the difference in training error-rate estimates among the possible models was small for the .632 and .632+ estimators, and the resulting EER estimates differed only slightly from the EER estimates obtained from the models with competing error-rate estimators.

3.6 Discussion

We have considered the model selection of the *RDA* classifier from Friedman [1989] using six well-known error-rate estimators with four small-sample, high-dimensional microarray data sets. Friedman has proposed the estimation of the *RDA* classifier's two tuning parameters by minimizing the *LOO* error-rate estimator. With its large variability, the *LOO* estimator frequently yields multiple models that attain the observed minimum *LOO* estimate. For this reason, we investigated other choices of error-rate estimators that are known to have reduced variability to identify a better model selection method for the *RDA* classifier that overcomes the issue of ties. We expected the models selected with the .632 and .632+ estimators to have relatively good classification performance because several empirical studies, such as the ones from Glele Kakaï and Palm [2009], Wehberg and Schumacher [2004], and Fitzmaurice, Krzanowski, and Hand [1991], have demonstrated the superior estimation of these estimators. However, we determined that these two estimators yielded large model selection variability because the differences in error-rate estimates were small for large subsets of these models. Moreover, the models selected with the .632 and .632+ estimators resulted in degraded classification performance relative to the other error-rate estimators.

As discussed by Fu et al. [2005], the *BCV* error-rate estimator yields smaller variability than estimators based on the $LOO - Boot$ estimator, including the .632 and .632+ estimators, for small sample sizes. In our study, we also conclude that the BCV estimator outperforms the .632 and .632+ estimator for model selection with the *RDA* classifier. However, we did not anticipate that the *LOO*, 10-fold *CV*, and bootstrap error-rate estimators would yield relatively good classification performance. Of these latter four estimators, we prefer the 10-fold *CV* estimator for *RDA* model selection because it is the simplest and fastest to compute; furthermore, we have found that the 10-fold *CV* estimator is easier to explain to collaborators and

clients, in practice, than the resampling-based estimators. On the other hand, the 10-fold *CV* estimator resulted in larger variability in model selection than the *BCV* and bootstrap error-rate estimators. However, if we also desire small variability in model selection and if the additional computation is warranted, we recommend the *BCV* estimator for *RDA* model selection in the small-sample, high-dimensional setting.

Although our empirical comparison of error-rate estimators has been restricted to the model selection of the *RDA* classifier, we welcome similar studies for other classifiers, such as the support vector machine classifiers with both polynomial and radial basis functions.

Table 3.1: The estimated $\widehat{\text{EER}}$ for the selected *RDA* models obtained from each error-rate estimator with approximate standard errors in parentheses.

	Estimators	Alon	Chiaretti	Golub	Singh
1	LOO	0.142 (0.022)	0.055 (0.014)	0.027 (0.010)	0.075 (0.017)
2	CV10	0.146 (0.022)	0.054 (0.014)	0.027 (0.010)	0.069 (0.016)
3	BOOT	0.160 (0.023)	0.048 (0.013)	0.053 (0.014)	0.066 (0.016)
4	632	0.381 (0.031)	0.072 (0.016)	0.175 (0.024)	0.104 (0.019)
5	632+	0.207 (0.026)	0.069 (0.016)	0.175 (0.024)	0.091 (0.018)
6	BCV	0.157 (0.023)	0.047 (0.013)	0.044 (0.013)	0.066 (0.016)

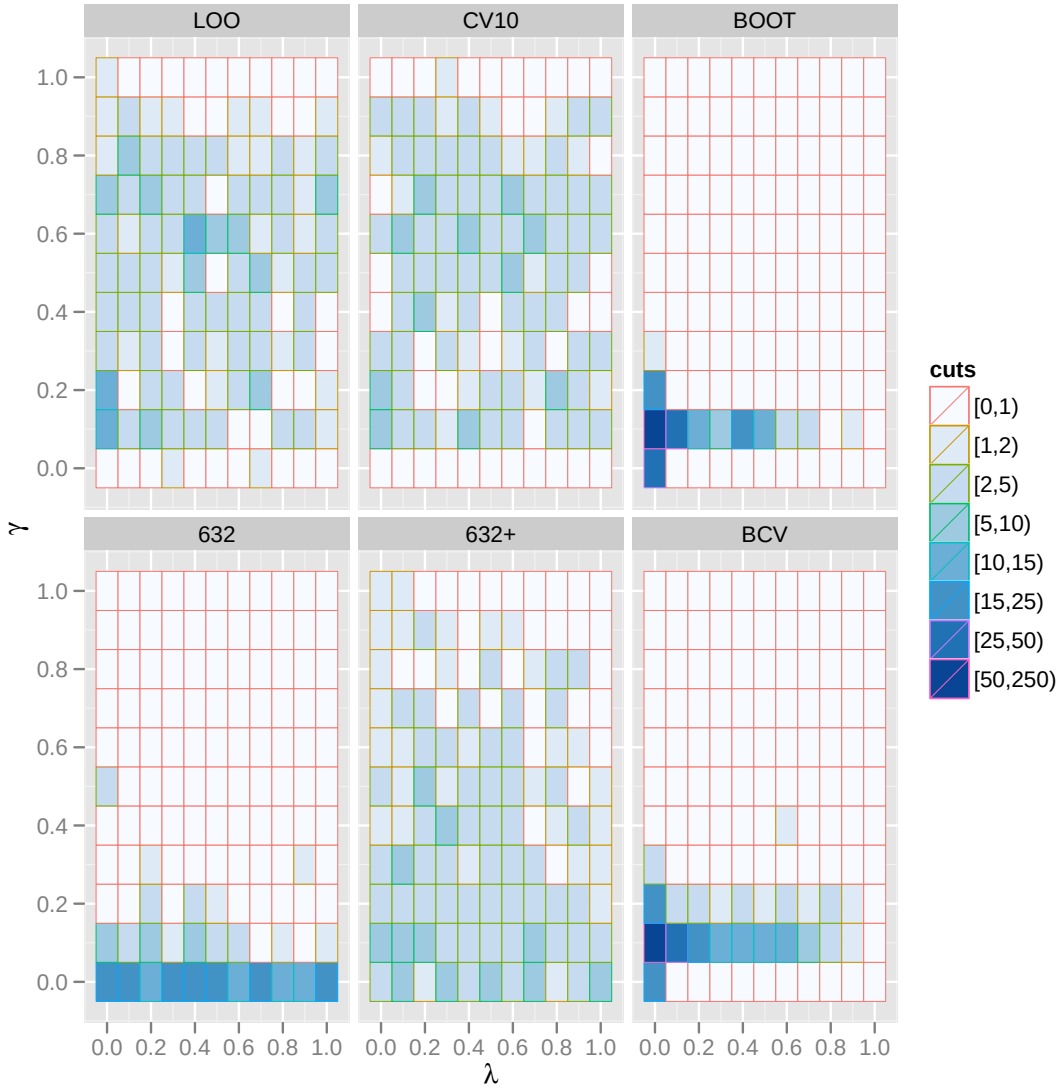

Figure 3.1: Heat maps of the number of times an *RDA* model is selected with respect to the competing error-rate estimators for the Alon data set.

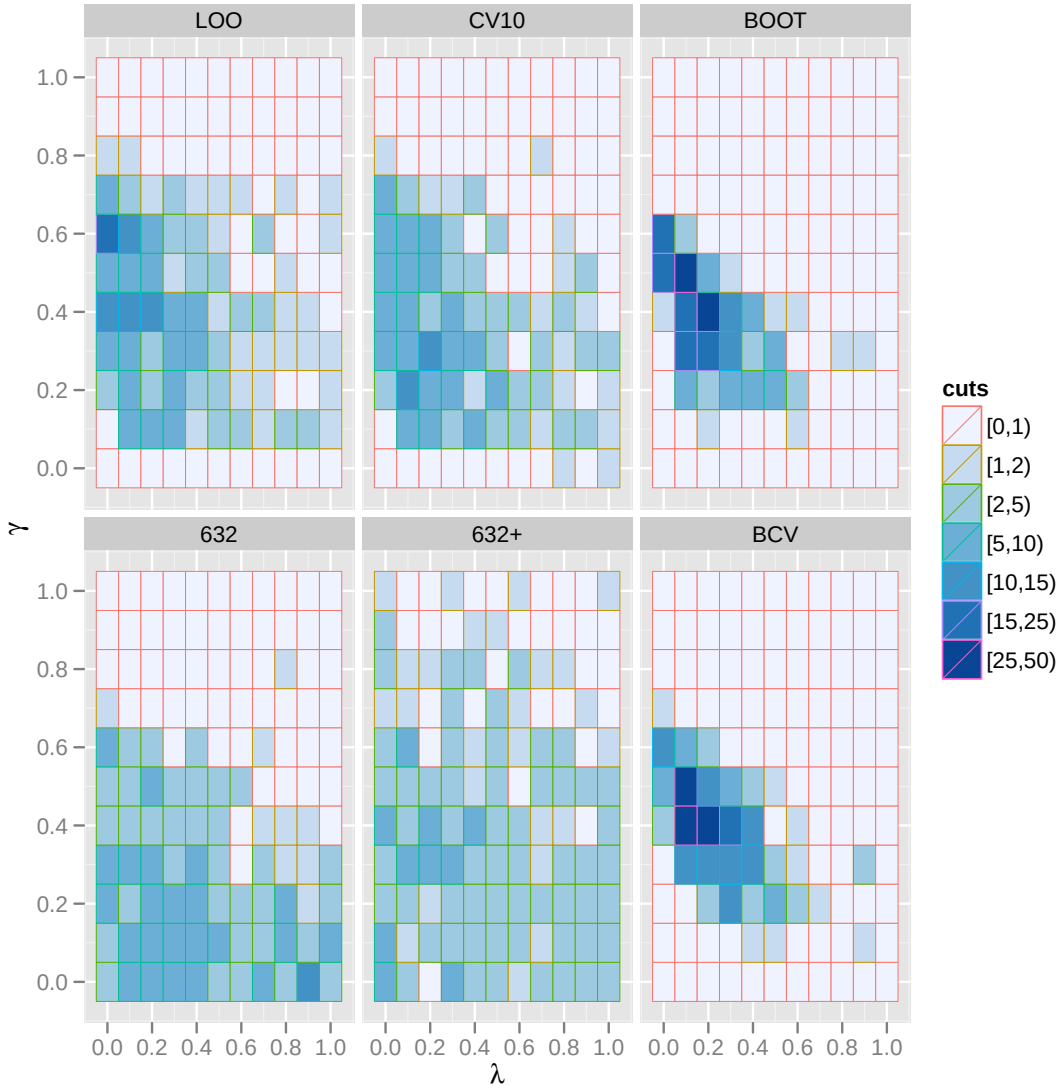

Figure 3.2: Heat maps of the number of times an *RDA* model is selected with respect to the competing error-rate estimators for the Chiaretti data set.

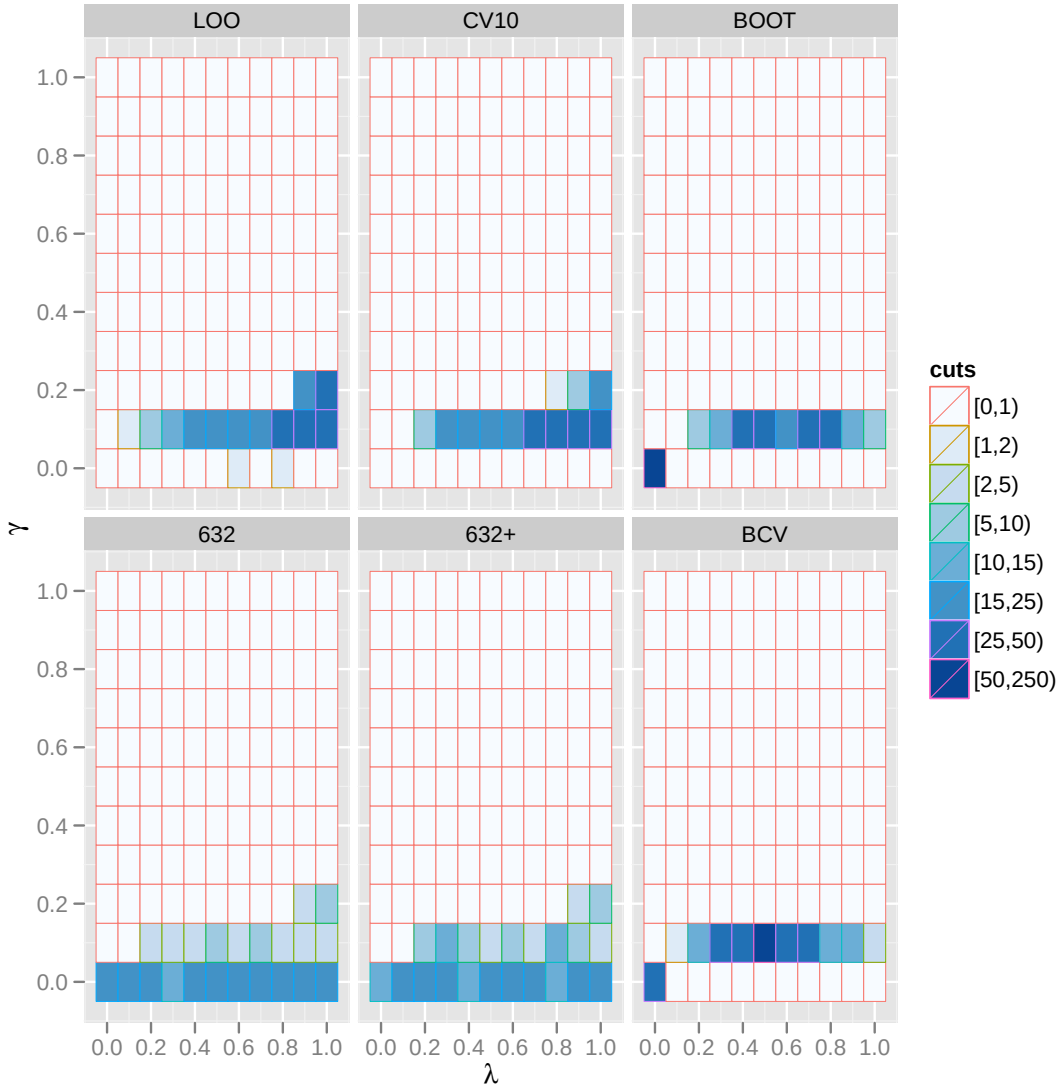

Figure 3.3: Heat maps of the number of times an *RDA* model is selected with respect to the competing error-rate estimators for the Golub data set.

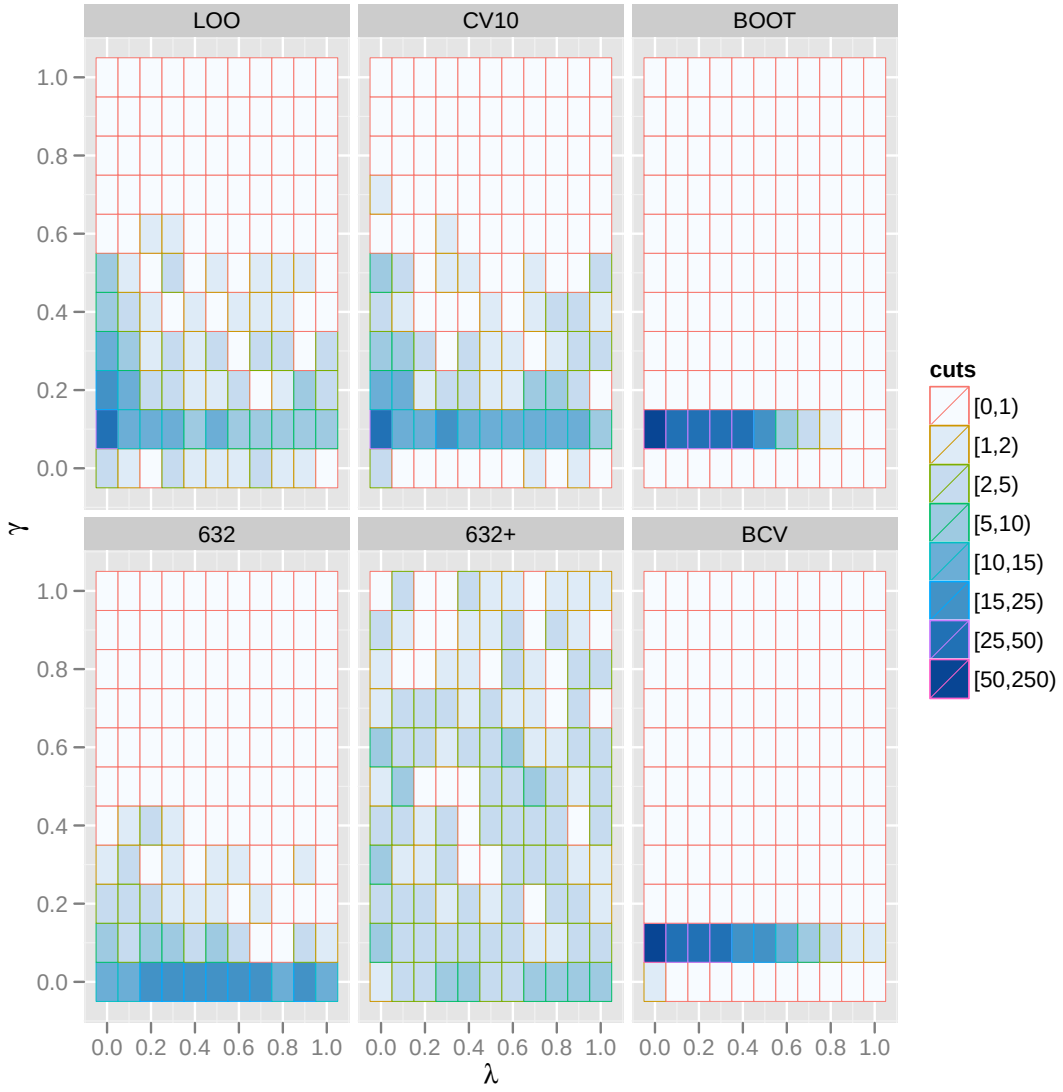

Figure 3.4: Heat maps of the number of times an *RDA* model is selected with respect to the competing error-rate estimators for the Singh data set.

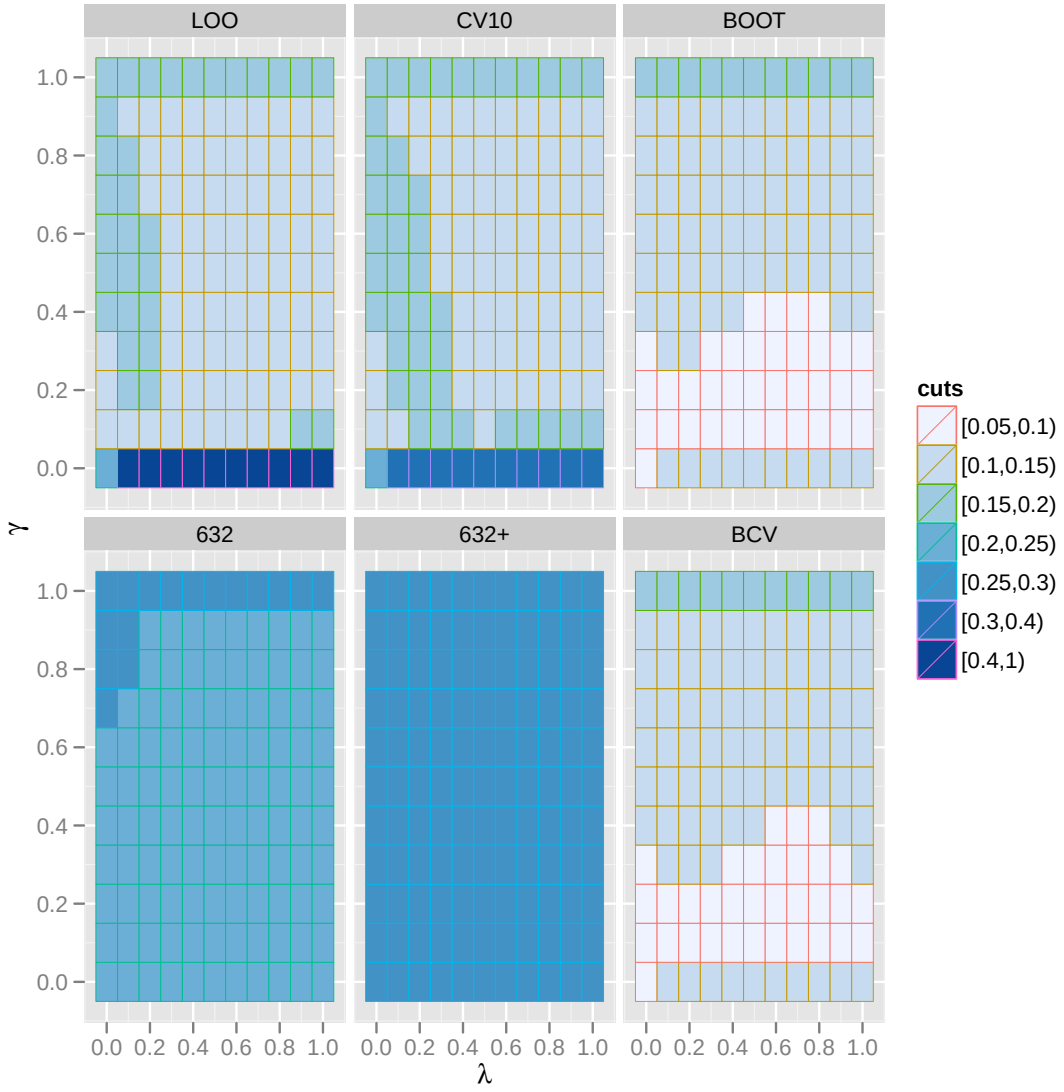

Figure 3.5: Heat maps of the average training error rate of the *RDA* models with respect to the competing error-rate estimators for the Alon data set.

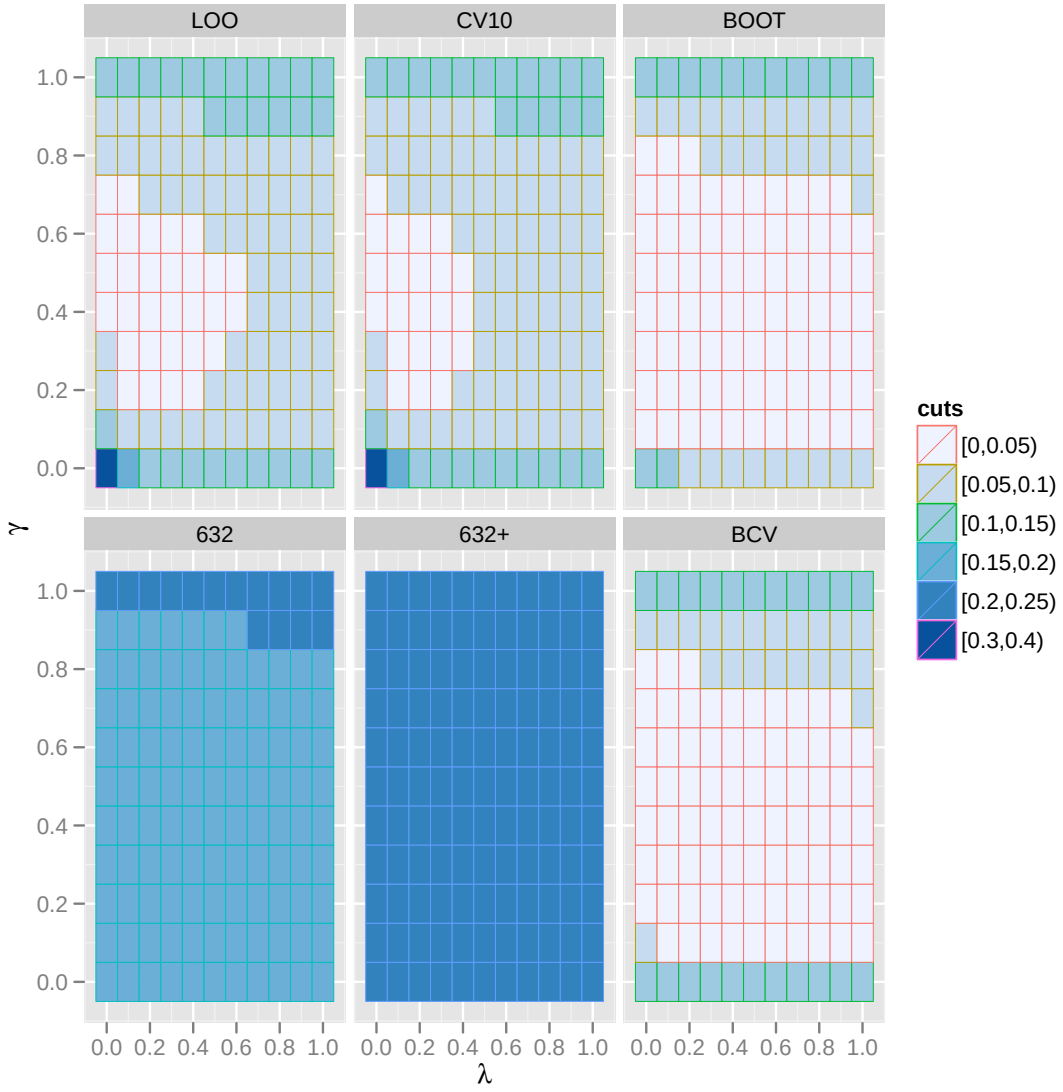

Figure 3.6: Heat maps of the average training error rate of the *RDA* models with respect to the competing error-rate estimators for the Chiaretti data set.

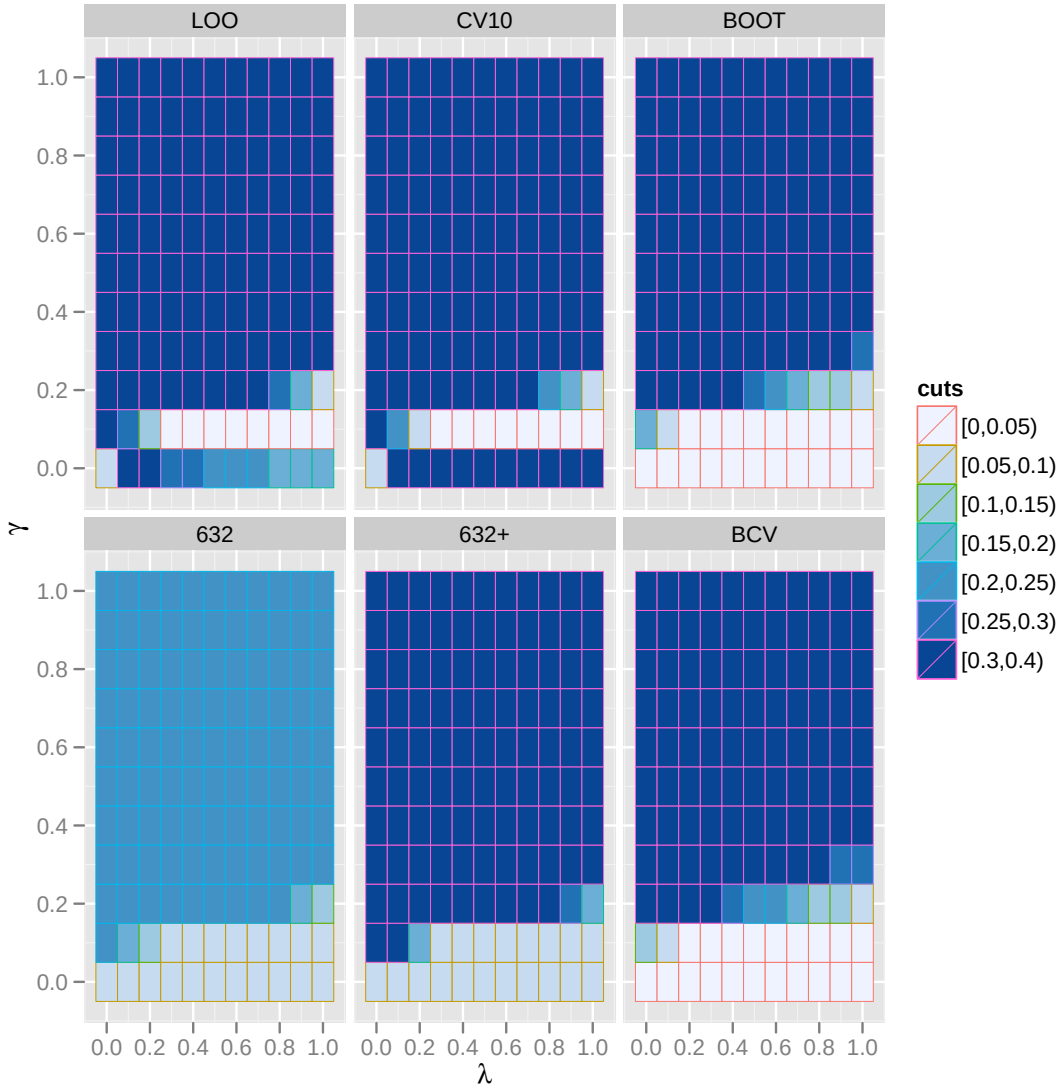

Figure 3.7: Heat maps of the average training error rate of the *RDA* models with respect to the competing error-rate estimators for the Golub data set.

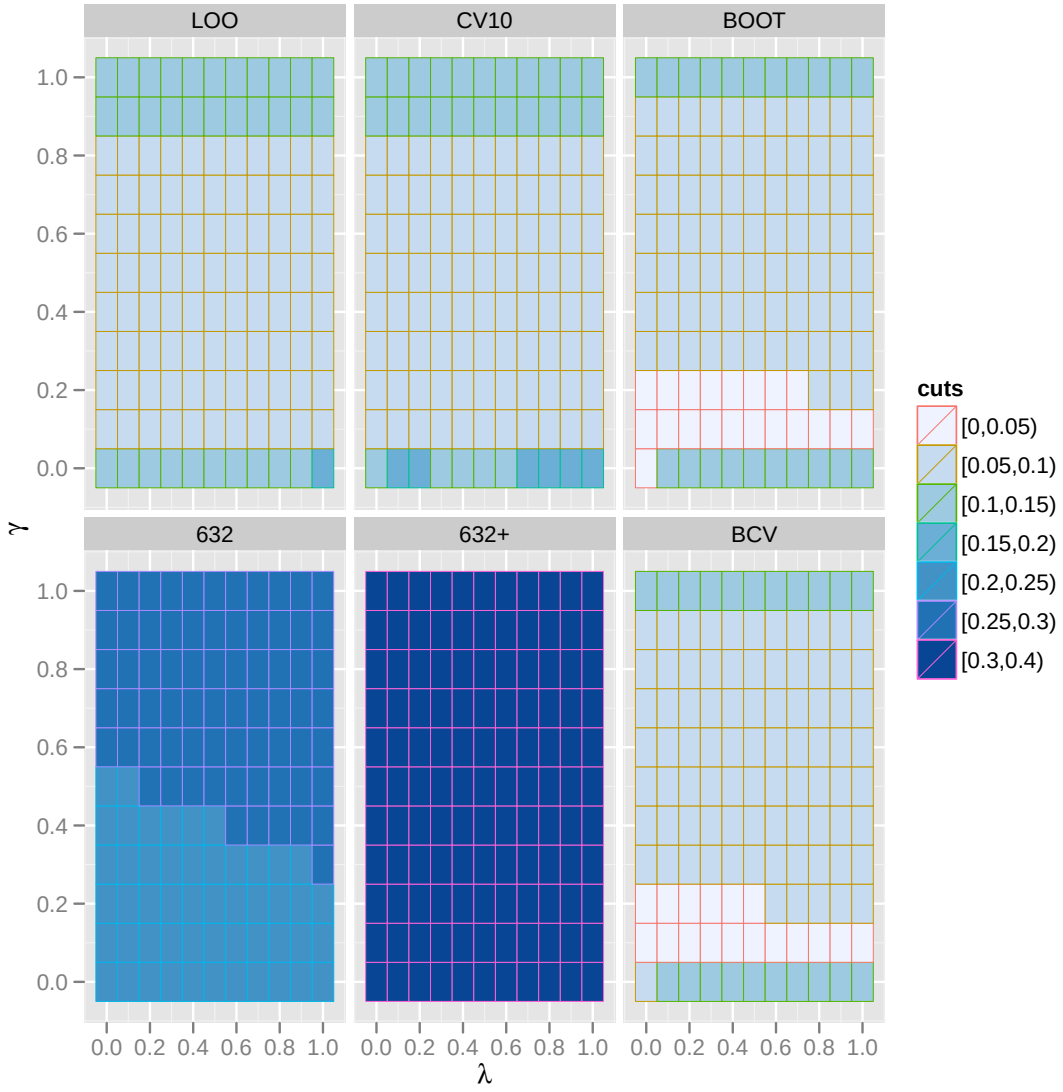

Figure 3.8: Heat maps of the average training error rate of the *RDA* models with respect to the competing error-rate estimators for the Singh data set.

CHAPTER FOUR

SimDiag: An Alternative to Diagonal Discriminant Analysis

4.1 Introduction

Pang et al. [2009] have claimed that classification of patients into known classes is one of the most important statistical problems in cancer genomics. The task is important yet arduous because high-throughput microarrays yield gene expression data sets with p gene expression levels and N observations, where typically $p \gg N$. Modern gene expression data sets often have between 10,000 and 50,000 probes or probe sets obtained from a relatively small number of patients (e.g., 25-50 patients per group). Furthermore, the classification performance of standard supervised learning methods, such as the LDA classifier, degrades due to the *curse of dimensionality* [Bellman, 1961]. As a result, many researchers emphasize simple, parsimonious models to avoid the estimation of a large number of parameters relative to the training sample size N. In fact, seemingly naive models can often perform well due to their simplicity. For example, with a naive assumption that features are uncorrelated within each class, Dudoit et al. [2002] have used a modification of the LDA classifier where the off-diagonal elements of the class covariance matrices are assumed to be zero. We refer to this supervised learning method as the *DLDA* classifier.

Dudoit et al. [2002], Pang et al. [2009], and Tong et al. [2012] have shown that the *DLDA* classifier performs well with high-dimensional gene expression data. Furthermore, Bickel and Levina [2004] have shown that the *DLDA* classifier is asymptotically superior to the LDA classifier for two multivariate normal populations with equal covariance matrices. Also, Pang et al. [2009] have proposed the *SDLDA* classifier and have shown that it can outperform the *DLDA*, support vector machines, and k-nearest neighbors classifiers in many small-sample, high-dimensional situa-

tions. Moreover, Tong et al. [2012] have asserted that their proposed *SmDLDA* classifier is superior to the *DLDA* classifier in terms of classification performance because of an improved mean estimator for high-dimensional data.

Although the *DLDA* classifier and its variants can perform well when $p \gg N$, we contend that classificatory information is present in the off-diagonal elements of the covariance matrices, which, if preserved, can yield gains in classification performance with the *DLDA* classifier. In this paper, we aim to preserve much of the classificatory information present in the covariances between gene expressions while taking advantage of the reduction in the number of estimated model parameters and the computational efficiency of the *DLDA* classifier. We propose two alternative classifiers that can yield superior classification performance to the *DLDA* classifier and its variants from Pang et al. [2009] and Tong et al. [2012]. Specifically, prior to applying the *DLDA* classifier, we simultaneously diagonalize the sample covariance matrices from two classes by constructing a linear projection matrix that transforms the feature space so that gene expressions are uncorrelated within each class. Thus, we propose the *SimDiag* and *Pool-Diag* classifiers, which first reduce the naiveté of the diagonal covariance matrix assumption prior to employing the *DLDA* classifier and, hence, can improve the classification performance of the *DLDA* classifier. Our new classifiers can substantially reduce the number of parameters to incorporate within the constructed classifiers, thereby greatly reducing the variance of the estimated classifiers. Moreover, we demonstrate that the improved assumption of diagonal covariance matrices often yields gains in classification performance compared to the improved variance estimators employed with the *SDLDA* classifier and the improved mean estimators applied with the *SmDLDA* classifier.

We have organized the remainder of the paper as follows. In Section 2, we present and discuss the supervised classification problem and the DLDA, SDLDA, and SmDLDA classifiers. In Section 3, we discuss simultaneous diagonalization of co-

variance matrices and our proposed *SimDiag* and *Pool-Diag* classifiers. We compare the classification performance of our proposed classifiers with the DLDA, SDLDA, and SmDLDA classifiers with four microarray data sets in Section 4. Finally, we conclude with a brief discussion in Section 5.

4.2 Discriminant Analysis

In discriminant analysis, also known as supervised learning, we attempt to classify an unlabeled p-dimensional observation $\boldsymbol{x} = (x_1, \ldots, x_p)'$ into one of K known groups or classes, where we assume that $\boldsymbol{x}$ belongs to class k with *a priori* probability π_k for $k = 1, 2, \ldots, K$ with $\sum_{k=1}^{K} \pi_k = 1$. We assume that the *a priori* probabilities π_k are equal for $k = 1, \ldots, K$. Let $\mathbb{R}_{m\times n}$, $\mathbb{R}^{>}_{m\times m}$, and $\mathbb{R}^{\geq}_{m\times m}$ denote the matrix space of all $m \times n$ matrices over the real field $\mathbb{R}$, the cone of $m \times m$ positive definite real matrices, and the cone of $m \times m$ positive semidefinite real matrices, respectively. Also, let $\boldsymbol{A}'$, $\boldsymbol{A}^{-}$, and $\boldsymbol{A}^{+}$ denote the transpose, generalized inverse, and Moore-Penrose pseudoinverse of the matrix $\boldsymbol{A} \in \mathbb{R}_{m\times n}$, respectively. Let $\boldsymbol{A}^{+/2} = (\boldsymbol{A}^{+})^{1/2}$, and let $\mathcal{C}(\boldsymbol{A})$ denote the column space of $\boldsymbol{A}$. We say that $\boldsymbol{A}$ is idempotent if and only if $\boldsymbol{A}^2 = \boldsymbol{A}$. We denote the squared-Frobenius norm of $\boldsymbol{A}$ by $||\boldsymbol{A}||_F^2 = \text{tr}\{\boldsymbol{A}'\boldsymbol{A}\}$, where $\text{tr}\{\cdot\}$ is the matrix trace operation. Let $\boldsymbol{B} \oplus \boldsymbol{D}$ and $\boldsymbol{B} \circ \boldsymbol{D}$ denote the direct sum [Lütkepohl, 1996, Chapter 1] and the element-wise Hadamard product of $\boldsymbol{B}, \boldsymbol{D} \in \mathbb{R}_{r\times r}$, respectively [Harville, 2008]. We denote by $\boldsymbol{I}_p$ the $p \times p$ identity matrix. Additionally, we assume that we have drawn n_k independently and identically distributed (IID) random vectors from the kth class

$$\boldsymbol{x}_{k,1} \ldots, \boldsymbol{x}_{k,n_k} \overset{IID}{\sim} N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\boldsymbol{x}_{k,i} \in \mathbb{R}_{p\times 1}$ is the ith sample of gene expressions from class k and $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}_k \in \mathbb{R}_{p\times 1}$ and covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{>}_{p\times p}$. We construct a supervised classifier from the $N = n_1 + n_2 + \ldots + n_K$ training observations to predict the class membership

of $\boldsymbol{x}$. Typically, with gene expression data, the groups are the diseases or disease subtypes under consideration. In particular, we focus on the $K = 2$ case, where typically one class is a diseased group while the other class is a control group.

We estimate the unknown parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ with their maximum likelihood estimators (MLEs), $\widehat{\boldsymbol{\mu}}_k = (\widehat{\mu}_{k1}, \ldots, \widehat{\mu}_{kp})'$ and $\widehat{\boldsymbol{\Sigma}}_k$, respectively, where

$$\widehat{\boldsymbol{\mu}}_k = n_k^{-1} \sum_{i=1}^{n_k} \boldsymbol{x}_{k,i} \tag{4.1}$$

and

$$\widehat{\boldsymbol{\Sigma}}_k = n_k^{-1} \sum_{i=1}^{n_k} (\boldsymbol{x}_{k,i} - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_{k,i} - \widehat{\boldsymbol{\mu}}_k)'. \tag{4.2}$$

We further assume that $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} \in \mathbb{R}^{>}_{p \times p}$, $k = 1, 2$. The pooled sample covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{k=1}^{K} n_k \widehat{\boldsymbol{\Sigma}}_k \tag{4.3}$$

is the MLE of $\boldsymbol{\Sigma}$. The sample *LDA* classifier is defined as follows: assign an unlabeled observation $\boldsymbol{x}$ to class k using

$$\widehat{D}(\boldsymbol{x}) = \underset{k \in \{1,2\}}{\arg\min} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k)' \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k). \tag{4.4}$$

4.2.1 Diagonal Linear Discriminant Analysis

For $p > n$, the pooled sample covariance matrix estimator in (5.3) is singular, and, hence, (5.4) is ill-posed and incalculable. Often, one employs some combination of variable selection, dimension reduction, and covariance matrix regularization to ensure that (5.3) is positive definite [Ramey and Young, 2011]. By assuming that $\boldsymbol{\Sigma}$ is a diagonal matrix, we have another effective approach for stabilizing (5.4). In particular, we assume that the gene expressions within class k are uncorrelated and, therefore, that $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k \circ \boldsymbol{I}_p = \text{ diag}(\sigma_{k1}^2, \sigma_{k2}^2, \ldots, \sigma_{kp}^2)$, $k = 1, 2$, where σ_{kj}^2 is the true marginal variance of the jth gene expression of the kth class for $j = 1, \ldots, p$. Thus,

we estimate $\boldsymbol{\Sigma}_k$ with $\widehat{\boldsymbol{\Sigma}}_k \circ \boldsymbol{I}_p = \text{ diag}(\widehat{\sigma}^2_{k1}, \widehat{\sigma}^2_{k2}, \ldots, \widehat{\sigma}^2_{kp})$, where the MLE of σ^2_{kj} is $\widehat{\sigma}^2_{kj} = n_k^{-1} \sum_{i=1}^{n_k} (x_{ij} - \overline{x}_{kj})^2$. If we further assume that $\boldsymbol{\Sigma}_k \equiv \boldsymbol{\Sigma}$, then we estimate $\boldsymbol{\Sigma}$ with the diagonal pooled sample covariance matrix $\widehat{\boldsymbol{\Sigma}} \circ \boldsymbol{I}_p = \text{ diag}(\widehat{\sigma}^2_1, \widehat{\sigma}^2_2, \ldots, \widehat{\sigma}^2_p)$, where $\widehat{\sigma}^2_j = N^{-1} \sum_{k=1}^{K} n_k \widehat{\sigma}^2_{kj}$. Hence, with the diagonal covariance matrix assumption, (5.4) reduces to the *DLDA* classifier

$$\widehat{D}^{DLDA}(\boldsymbol{x}) = \underset{k \in \{1,2\}}{\arg\min} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k)' \left(\widehat{\boldsymbol{\Sigma}} \circ \boldsymbol{I}_p \right)^{-1} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k). \tag{4.5}$$

Our assumption of equal, diagonal covariance matrices reduces the number of covariance parameters to estimate from $Kp(p+1)/2$ to p. Furthermore, the inverse of $\widehat{\boldsymbol{\Sigma}}$ exists and (4.5) is easily and quickly calculated as a dot product.

4.2.2 Shrinkage-based Diagonal Linear Discriminant Analysis

To improve the estimation of the diagonal covariance matrices for $p >> N$, Pang et al. [2009] have proposed the *SDLDA* classifier, based on a family of shrinkage-based estimators from Tong and Wang [2007]. Following Pang et al. [2009], we write the shrinkage-based estimator for σ_j^{2t} as

$$\tilde{\sigma}_j^{2t}(\alpha) = \left\{ h_{\nu,p}(t) \widehat{\sigma}^{2t}_{pool} \right\}^{\alpha} \left\{ h_{\nu,1}(t) \widehat{\sigma}^{2t}_j \right\}^{1-\alpha}, \tag{4.6}$$

where $t \in \mathbb{R}$, $\alpha \in [0,1]$, $\nu = N - K$, $\widehat{\sigma}^{2t}_{pool} = \prod_{j=1}^{p} \left(\widehat{\sigma}^2_j \right)^{t/p}$ is a pooled variance estimator,

$$h_{\nu,p}(t) = (\nu/2)^t \left(\frac{\Gamma(\nu/2)}{\Gamma(\nu/2 + t/2)} \right)^p, \tag{4.7}$$

and $\Gamma(\cdot)$ is the gamma function.

The term (5.7) is a bias-correction term such that $h_{\nu,1}(t) \widehat{\sigma}^{2t}_j$ is an unbiased estimator for σ_j^{2t}, and $h_{\nu,p}(t) \widehat{\sigma}^{2t}_{pool}$ is an unbiased estimator for σ^{2t} when $\sigma^2_j \equiv \sigma^2$. The shrinkage parameter α controls the amount of shrinkage from the individual variance estimator toward the bias-corrected pooled estimator. Tong and Wang [2007] have shown that for fixed p, ν, and $t > -\nu/2$, there exists a unique optimal

α with respect to the risk function $R_{Stein}(\boldsymbol{\sigma}^{2t}, \tilde{\boldsymbol{\sigma}}^{2t})$ corresponding to the Stein loss function $L_{Stein}(\sigma^2, \tilde{\sigma}^2) = \tilde{\sigma}^2/\sigma^2 - \ln(\tilde{\sigma}^2/\sigma^2) - 1$, where $\boldsymbol{\sigma}^{2t} = (\sigma_1^{2t}, \ldots, \sigma_p^{2t})'$ and $\tilde{\boldsymbol{\sigma}}^{2t} = (\tilde{\sigma}_1^{2t}, \ldots, \tilde{\sigma}_p^{2t})'$. For brevity we do not include the Stein risk function but note that it is implicitly a function of α.

To estimate α, Pang et al. [2009] have proposed that α^* be chosen from a grid of candidate parameters such that

$$\alpha^* = \underset{\alpha \in [0,1]}{\arg\min}\, R_{Stein}(\boldsymbol{\sigma}^{2t}, \tilde{\boldsymbol{\sigma}}^{2t}). \tag{4.8}$$

Even for large p, the empirical minimization of the Stein risk function requires little computation time. Using (5.8) and $t = -1$, Pang et al. [2009] have estimated $(\boldsymbol{\Sigma} \circ \boldsymbol{I}_p)^{-1}$ with

$$\widehat{\boldsymbol{\Sigma}}^{-1}(\alpha^*) = \text{diag}(\tilde{\sigma}_1^{-2}(\alpha^*), \ldots, \tilde{\sigma}_p^{-2}(\alpha^*)). \tag{4.9}$$

Pang et al. [2009] have substituted (4.9) for $(\boldsymbol{\Sigma} \circ \boldsymbol{I}_p)^{-1}$ into (4.5) to obtain the *SDLDA* classifier

$$\widehat{D}^{SDLDA}(\boldsymbol{x}) = \underset{k \in \{1,2\}}{\arg\min} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k)' \widehat{\boldsymbol{\Sigma}}^{-1}(\alpha^*)(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k).$$

In our Monte Carlo simulations in Section 4.4, we estimate α using a grid of 21 equidistant candidate values between 0 and 1, inclusively.

4.2.3 The DLDA Classifier with Improved Mean Estimation

While Pang et al. [2009] have sought to improve the *DLDA* classifier with improved estimators of the marginal variances, Tong et al. [2012] have argued that the MLE for $\boldsymbol{\mu}_k$, $k = 1, 2$, is unreliable within the *DLDA* classifier for $p \gg N$. Instead, Tong et al. [2012] have considered improved mean estimators with optimal shrinkage parameters under quadratic loss as $p \to \infty$. Tong et al. have discussed that James-Stein mean estimators of the form

$$\widehat{\boldsymbol{\mu}}_k^{(JS)} = \left(1 - \frac{(p-2)/n_k}{||\widehat{\boldsymbol{\mu}}_k||^2_{\widehat{\boldsymbol{\Sigma}}}}\right) \widehat{\boldsymbol{\mu}}_k,$$

where $||\boldsymbol{z}||^2_{\boldsymbol{A}} = \boldsymbol{z}'\boldsymbol{A}^{-1}\boldsymbol{z}$ for $\boldsymbol{z} \in \mathbb{R}_{p\times 1}$ and $\boldsymbol{A} \in \mathbb{R}^{>}_{p\times p}$, cannot be used when $p > N$ because $\widehat{\boldsymbol{\Sigma}}$ is singular. Hence, Tong et al. [2012] have considered a similar family of estimators where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma} \circ \boldsymbol{I}_p$. Under the equal diagonal covariance matrix assumption, Tong et al. have proposed the mean estimator

$$\widehat{\boldsymbol{\mu}}_k(\widehat{r}_k) = \left(1 - \frac{\widehat{r}_k}{||\widehat{\boldsymbol{\mu}}_k||^2_{\widehat{\boldsymbol{\Sigma}}\circ\boldsymbol{I}_p}}\right)\widehat{\boldsymbol{\mu}}_k, \tag{4.10}$$

where $\widehat{r}_k = (n_k - 1)(p-2)/\{n_k(n_k-3)\}$ is a shrinkage estimator optimized under quadratic loss with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma} \circ \boldsymbol{I}_p$ as $p \to \infty$. Substituting (5.9) for $\widehat{\boldsymbol{\mu}}_k$ in (4.5), we have the *SmDLDA* classifier function

$$\widehat{D}^{SmDLDA}(\boldsymbol{x}) = \underset{k\in\{1,2\}}{\arg\min}\{\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k(\widehat{r}_k)\}'\left(\widehat{\boldsymbol{\Sigma}} \circ \boldsymbol{I}_p\right)^{-1}\{\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k(\widehat{r}_k)\}. \tag{4.11}$$

4.3 Simultaneous Diagonalization of Two Covariance Matrices

As we have discussed, the *DLDA*, *SDLDA*, and *SmDLDA* classifiers have been developed for small-sample, high-dimensional microarray data sets with the naive assumption of diagonal covariance matrices. However, we contend that if we can preserve classificatory information present in the off-diagonal elements of the covariance matrices prior to applying the *DLDA* classifier, then we can improve its classification performance. In particular, we desire to determine a linear transformation $\boldsymbol{Q}\boldsymbol{x} \sim N_q(\boldsymbol{Q}\boldsymbol{\mu}_k, \boldsymbol{Q}\boldsymbol{\Sigma}_k\boldsymbol{Q}' \circ \boldsymbol{I}_q)$ for $k = 1, 2$, where $\boldsymbol{Q} \in \mathbb{R}_{q\times p}$ $(q \leq p)$. Rather than restricting the covariance matrices of the original populations to be diagonal, we desire that the covariance matrices of the transformed populations are diagonal. Hence, we wish to determine a $\boldsymbol{Q} \in \mathbb{R}_{q\times p}$ that improves our goal of diagonal covariance matrices. We say that the set of matrices $\{\boldsymbol{\Sigma}_k | k = 1, 2\}$ is simultaneously diagonalizable if $\boldsymbol{Q}\boldsymbol{\Sigma}_k\boldsymbol{Q}' = \boldsymbol{Q}\boldsymbol{\Sigma}_k\boldsymbol{Q}' \circ \boldsymbol{I}_q$, $k = 1, 2$, and that $\boldsymbol{Q}$ is a *simultaneous diagonalizer*. From Anderson [2003] and Fukunaga [1990], we have the following well-known result:

Result 1. *Let* $\boldsymbol{A}_1 \in \mathbb{R}^{>}_{p\times p}$ *and* $\boldsymbol{A}_2 \in \mathbb{R}^{\geq}_{p\times p}$ *be symmetric matrices. Then, there exists* $\boldsymbol{Q} \in \mathbb{R}_{p\times p}$ *such that* $\boldsymbol{Q}\boldsymbol{A}_1\boldsymbol{Q}' = \boldsymbol{\Lambda}$ *and* $\boldsymbol{Q}\boldsymbol{A}_2\boldsymbol{Q}' = \boldsymbol{I}_p$*, where* $\boldsymbol{\Lambda} \in \mathbb{R}_{p\times p}$ *is a diagonal matrix.*

In particular, Anderson [2003] has provided a solution for $\boldsymbol{Q}$ using the Cholesky decomposition. Fukunaga [1990] has shown that $\boldsymbol{Q} = \boldsymbol{\Phi}$ simultaneously diagonalizes $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, where $\boldsymbol{\Lambda}$ is the matrix of the eigenvalues of $\boldsymbol{A}_1^{-1}\boldsymbol{A}_2$ and the columns of $\boldsymbol{\Phi}$ are the corresponding eigenvectors of $\boldsymbol{A}_1^{-1}\boldsymbol{A}_2$. Also, one can obtain a solution for $\boldsymbol{Q}$ from the generalized eigenvalue problem, $\boldsymbol{A}_1\boldsymbol{x} = \lambda\boldsymbol{A}_2\boldsymbol{x}$ [Golub and van Loan, 1996].

With high-dimensional microarray data we do not satisfy the necessary conditions in the above result because we typically have $p > N$, which implies that the covariance matrix estimators for each class are singular. That is, with microarray data, we have that $\text{rank}(\widehat{\boldsymbol{\Sigma}}_k) < p$, $k = 1, 2$, whereas the above result requires that at least one of the two sample covariance matrices has full rank. Hence, we can apply neither the result from Anderson [2003] nor the result from Fukunaga [1990]. Instead, we require a generalization of the above result to simultaneously diagonalize two positive semidefinite matrices. From Harville [2008] we require three lemmas to generalize the above result and give them here without proof.

Lemma 1 (Harville, 2008, Chapter 5). *Let* $\boldsymbol{A} \in \mathbb{R}_{r\times m}$*. Then,* $||\boldsymbol{A}||_F^2 = 0$ *if and only if* $\boldsymbol{A} = \boldsymbol{0}$*, a matrix of zeroes.*

Lemma 2 (Harville, 2008, Chapter 10). *Let* $\boldsymbol{A} \in \mathbb{R}_{r\times m}$ *with rank*$(\boldsymbol{A}) = m$ *and* $\boldsymbol{B} \in \mathbb{R}_{m\times s}$*. Then, we have that* $\boldsymbol{A}^{-}\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^{-}$ *is idempotent.*

Lemma 3 (Harville, 2008, Chapter 10). *Let* $\boldsymbol{A} \in \mathbb{R}_{m\times m}$ *be idempotent with rank*$(\boldsymbol{A}) = m$*. Then,* $\boldsymbol{A} = \boldsymbol{I}_m$*.*

Here, we provide an additional lemma that we will use in the proof of our main result.

Lemma 4. *Let* $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{R}_{p\times q}$ *with rank*$(\boldsymbol{U}_k) = q$ *and* $\boldsymbol{U}_k'\boldsymbol{U}_k = \boldsymbol{I}_q$ *for* $k = 1, 2$. *Then,* $\boldsymbol{U}_1\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2' = \boldsymbol{I}_q \oplus \boldsymbol{0}_{p-q}$ *and* $\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{U}_1 = \boldsymbol{I}_q$.

Proof. First, notice that $\text{rank}(\boldsymbol{U}_1\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2') = q$. Hence, applying Lemma 2 with $\boldsymbol{A} = \boldsymbol{U}_1'$ and $\boldsymbol{B} = \boldsymbol{U}_2$, we see that $\boldsymbol{U}_1\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2'$ is idempotent. Therefore, by Lemma 3, $\boldsymbol{U}_1\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2' = \boldsymbol{I}_q \oplus \boldsymbol{0}_{p-q}$. Next, we have

$$\begin{aligned}
||\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{U}_1 - \boldsymbol{I}_q||_F^2 &= \text{tr}\{\boldsymbol{U}_1\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{U}_1\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2' - 2\boldsymbol{U}_1\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2'\} + \text{tr}\{\boldsymbol{I}_q\} \\
&= q - \text{tr}\{\boldsymbol{I}_q \oplus \boldsymbol{0}_{p-q}\} \\
&= 0.
\end{aligned}$$

Hence, by Lemma 1 we have that $\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{U}_1 - \boldsymbol{I}_q = \boldsymbol{0}$, which implies that $\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{U}_1 = \boldsymbol{I}_q$. □

Now, we prove our main result by utilizing Lemmas 1–4.

Theorem 1. *Let* $\boldsymbol{A}_k \in \mathbb{R}^{\geq}_{p\times p}$ *be symmetric with rank*$(\boldsymbol{A}_k) = q_k$ *for* $k = 1, 2$. *Without loss of generality, we assume* $1 \leq q_2 \leq q_1 \leq p$. *Then, there exists* $\boldsymbol{Q}^{(q)} \in \mathbb{R}_{q\times p}$ *such that*

$$\begin{aligned}
\boldsymbol{Q}^{(q)}\boldsymbol{A}_1\boldsymbol{Q}^{(q)'} &= \boldsymbol{\Lambda}^{(q)} \\
\boldsymbol{Q}^{(q)}\boldsymbol{A}_2\boldsymbol{Q}^{(q)'} &= \boldsymbol{I}_{q^*} \oplus \boldsymbol{0}_{q-q^*},
\end{aligned}$$

where $q^* = \min\{q, q_2\}$.

Proof. First, we consider the spectral decomposition of $\boldsymbol{A}_k = \boldsymbol{U}_k\boldsymbol{\Lambda}_k\boldsymbol{U}_k'$ for $k = 1, 2$, where $\boldsymbol{\Lambda}_k = \text{diag}(\lambda_{k1}, \ldots, \lambda_{kp})$ is the diagonal matrix of eigenvalues of $\boldsymbol{A}_k$ with corresponding eigenvectors as the columns of $\boldsymbol{U}_k$ such that $\lambda_{k1} \geq \ldots \geq \lambda_{kq_k} > 0$ and $\lambda_{kq_k+1} = \ldots = \lambda_{kp} = 0$. For $1 \leq q \leq p$, let $\boldsymbol{\Lambda}_2^{(q)} \in \mathbb{R}_{q\times q}$ be the diagonal matrix of the q largest eigenvalues of $\boldsymbol{A}_2$ such that the corresponding eigenvectors are the columns of $\boldsymbol{U}_2^{(q)} \in \mathbb{R}_{p\times q}$, and notice that $\boldsymbol{\Lambda}_2 = \boldsymbol{\Lambda}_2^{(q)} \oplus \boldsymbol{I}_{p-q}$. Then, let $\boldsymbol{Q}_2^{(q)} = \left\{\boldsymbol{\Lambda}_2^{(q)}\right\}^{+/2}\boldsymbol{U}_2^{(q)'}$. Because $\boldsymbol{Q}_2^{(q)}\boldsymbol{A}_2\boldsymbol{Q}_2^{(q)'} = \boldsymbol{I}_{q^*} \oplus \boldsymbol{0}_{q-q^*}$, $\boldsymbol{Q}_2^{(q)}$ resembles a whitening transform of $\boldsymbol{A}_2$

[Duda, Hart, and Stork, 2001]. Now, let $\boldsymbol{Q}_1^{(q)} = \boldsymbol{U}_1^{(q)'}\boldsymbol{U}_2^{(q)}$, which implies that $\boldsymbol{Q}_1^{(q)}\boldsymbol{Q}_1^{(q)'} = \boldsymbol{I}_q$ by Lemma 4. Choosing $\boldsymbol{Q}^{(q)} = \boldsymbol{Q}_1^{(q)}\boldsymbol{Q}_2^{(q)} = \left\{\boldsymbol{\Lambda}_2^{(q)}\right\}^{+/2}\boldsymbol{U}_1^{(q)'}\boldsymbol{U}_2^{(q)}\boldsymbol{U}_2^{(q)'}$, we have that

$$\begin{aligned}
\boldsymbol{Q}^{(q)}\boldsymbol{A}_1\boldsymbol{Q}^{(q)'} &= \left\{\boldsymbol{\Lambda}_2^{(q)}\right\}^{+/2}\boldsymbol{U}_1^{(q)'}\boldsymbol{U}_2^{(q)}\boldsymbol{U}_2^{(q)'}\boldsymbol{U}_1^{(q)}\boldsymbol{\Lambda}_1\boldsymbol{U}_1^{(q)'}\boldsymbol{U}_2^{(q)}\boldsymbol{U}_2^{(q)'}\boldsymbol{U}_1^{(q)}\left\{\boldsymbol{\Lambda}_2^{(q)}\right\}^{+/2} \\
&= \left\{\boldsymbol{\Lambda}_2^{(q)}\right\}^{+/2}\boldsymbol{\Lambda}_1^{(q)}\left\{\boldsymbol{\Lambda}_2^{(q)}\right\}^{+/2} \\
&= \left\{\boldsymbol{\Lambda}_2^{(q)}\right\}^{+}\boldsymbol{\Lambda}_1^{(q)}
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{Q}^{(q)}\boldsymbol{A}_2\boldsymbol{Q}^{(q)'} &= \left\{\boldsymbol{\Lambda}_2^{(q)}\right\}^{+/2}\boldsymbol{U}_1^{(q)'}\boldsymbol{U}_2^{(q)}\boldsymbol{U}_2^{(q)'}\boldsymbol{U}_2^{(q)}\boldsymbol{\Lambda}_2\boldsymbol{U}_2^{(q)'}\boldsymbol{U}_2^{(q)}\boldsymbol{U}_2^{(q)'}\boldsymbol{U}_1^{(q)}\left\{\boldsymbol{\Lambda}_2^{(q)}\right\}^{+/2} \\
&= \boldsymbol{I}_{q^*} \oplus \boldsymbol{0}_{q-q^*}.
\end{aligned}$$

Therefore, $\boldsymbol{Q}^{(q)}$ simultaneously diagonalizes $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ in a q-dimensional subspace.

□

From Theorem 1, we can transform two sample covariance matrices so that the classificatory information contained in the off-diagonal elements is preserved in the first q^* transformed features of each class. Specifically, the variances of the first q^* features contain the classificatory information, while the remaining $p - q^*$ transformed features have zero variance. Additionally, in the following corollary, we show that for $q = q_2$, our main result generalizes directly the result from Fukunaga [1990] for the case of $K = 2$ positive-semidefinite sample covariance matrices.

Corollary 1. *Let $\boldsymbol{A}_k \in \mathbb{R}^{\geq}_{p\times p}$ be symmetric with rank$(\boldsymbol{A}_k) = q_k$ for $k = 1, 2$. Without loss of generality, we assume $1 \leq q_2 \leq q_1 \leq p$. Then, there exists $\boldsymbol{Q} \in \mathbb{R}_{q_2\times p}$ such that*

$$\begin{aligned}
\boldsymbol{Q}\boldsymbol{A}_1\boldsymbol{Q}' &= \boldsymbol{\Lambda} \\
\boldsymbol{Q}\boldsymbol{A}_2\boldsymbol{Q}' &= \boldsymbol{I}_{q_2} \oplus \boldsymbol{0}_{p-q_2},
\end{aligned}$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_2^{+}\boldsymbol{\Lambda}_1$ is the diagonal matrix of eigenvalues of $\boldsymbol{A}_2^{+}\boldsymbol{A}_1$.

Proof. In the above theorem, set $q = q_2$ and $\boldsymbol{Q} = \boldsymbol{Q}^{(q)}$. Hence, we have that $\boldsymbol{Q}\boldsymbol{A}_1\boldsymbol{Q}' = \boldsymbol{\Lambda}$ and $\boldsymbol{Q}\boldsymbol{A}_2\boldsymbol{Q}' = \boldsymbol{I}_{q_2} \oplus \boldsymbol{0}_{p-q_2}$. Next, we show that the diagonal matrix $\boldsymbol{\Lambda}_2^+\boldsymbol{\Lambda}_1$ consists of the eigenvalues of $\boldsymbol{A}_2^+\boldsymbol{A}_1$. Writing $\boldsymbol{A}_2^+\boldsymbol{A}_1 = \boldsymbol{U}_2\boldsymbol{\Lambda}_2^+\boldsymbol{U}_2'\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1'$, we write the eigenvalue formulation of $\boldsymbol{A}_2^+\boldsymbol{A}_1$ as $\boldsymbol{U}_2\boldsymbol{\Lambda}_2^+\boldsymbol{U}_2'\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1'\boldsymbol{x} = \lambda\boldsymbol{x}$, where λ is an eigenvalue of $\boldsymbol{A}_2^+\boldsymbol{A}_1$ and $\boldsymbol{x}$ is the orthonormal eigenvector corresponding to λ. Hence, we have that $\boldsymbol{U}_2'\boldsymbol{U}_1\boldsymbol{\Lambda}_2^+\boldsymbol{\Lambda}_1\boldsymbol{U}_1'\boldsymbol{x} = \lambda\boldsymbol{U}_2'\boldsymbol{x}$. Now, because $\boldsymbol{x} \in \mathcal{C}(\boldsymbol{A}_2^+\boldsymbol{A}_1)$, we have $\boldsymbol{x} \in \mathcal{C}(\boldsymbol{A}_2^+)$, which implies that $\boldsymbol{x} \in \mathcal{C}(\boldsymbol{U}_2)$. Thus, there exists $\boldsymbol{z} \in \mathbb{R}_{p\times 1}$ such that $\boldsymbol{x} = \boldsymbol{U}_2\boldsymbol{z}$, which implies that $\boldsymbol{U}_2'\boldsymbol{U}_1\boldsymbol{\Lambda}_2^+\boldsymbol{\Lambda}_1\boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{z} = \lambda\boldsymbol{\Lambda}_2^+\boldsymbol{\Lambda}_1\boldsymbol{z} = \lambda\boldsymbol{z}$ by Lemma 4. Therefore, λ is an eigenvalue of $\boldsymbol{\Lambda}_2^+\boldsymbol{\Lambda}_1$, which implies that the diagonal entries of $\boldsymbol{\Lambda}_2^+\boldsymbol{\Lambda}_1$ consist of the eigenvalues of $\boldsymbol{A}_2^+\boldsymbol{A}_1$. □

4.3.1 The SimDiag Classifier

Here, we describe our proposed *SimDiag* classifier. By generalizing the results from Anderson [2003] and Fukunaga [1990], we can simultaneously diagonalize two singular sample covariance matrices with $\boldsymbol{Q} = \boldsymbol{U}_1'\boldsymbol{U}_2\boldsymbol{U}_2'\boldsymbol{\Lambda}_2^{+/2}$. Thus, by choosing $\boldsymbol{A}_k = \widehat{\boldsymbol{\Sigma}}_k$ for $k = 1, 2$ in Theorem 1, we have that the assumption of diagonal covariance matrices is more reasonable and no longer as naive. Therefore, we can discard the off-diagonal elements of the transformed covariance matrices. We select the value of q_k as the number of nonzero eigenvalues of $\widehat{\boldsymbol{\Sigma}}_k$ for $k = 1, 2$; if the features in the kth class are linearly independent, we have that $q_k = n_k - 1$. After employing the simultaneous diagonalizer $\boldsymbol{Q}$, we transform (5.1) to obtain $\widehat{\boldsymbol{\mu}}_K^{SD} = \boldsymbol{Q}\widehat{\boldsymbol{\mu}}_k$ for $k = 1, 2$. Therefore, we write our proposed *SimDiag* classifier as

$$\widehat{D}_k^{SimDiag}(\boldsymbol{x}) = \underset{k\in\{1,2\}}{\arg\min}(\boldsymbol{Q}\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k^{SD})' \left(\boldsymbol{Q}\widehat{\boldsymbol{\Sigma}}_k\boldsymbol{Q}'\right)^{-1} (\boldsymbol{Q}\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k^{SD}).$$

4.3.2 The Pool-Diag Classifier

For $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, $k = 1, 2$, consider the spectral decomposition of $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p) \in \mathbb{R}_{p\times p}$ is the matrix of eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{U} \in \mathbb{R}_{p\times p}$ is the matrix of the corresponding eigenvectors of $\boldsymbol{\Sigma}$, where $\lambda_1 \geq \ldots \geq \lambda_p > 0$ and

$\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_p$. Thus, the whitening transform $\boldsymbol{B} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}' = \boldsymbol{U}'\boldsymbol{\Lambda}^{-1/2}$ diagonalizes $\boldsymbol{\Sigma}$ such that $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}' = \boldsymbol{I}_p = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}' \circ \boldsymbol{I}_p$ [Duda, Hart, and Stork, 2001].

As discussed in Section 5.2, $\widehat{\boldsymbol{\Sigma}}^{-1}$ does not exist for $p > N$ because $\text{rank}(\widehat{\boldsymbol{\Sigma}}) = q < p$, where $\widehat{\boldsymbol{\Sigma}}$ is given in (5.2). Let $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}'$ be the spectral decomposition of $\widehat{\boldsymbol{\Sigma}}$, where $\widehat{\boldsymbol{\Lambda}} = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p) \in \mathbb{R}_{p\times p}$ is the matrix of eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{U}} \in \mathbb{R}_{p\times p}$ is the matrix of the corresponding eigenvectors of $\widehat{\boldsymbol{\Sigma}}$, where $\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}} = \boldsymbol{I}_p$. Recall that $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_q > 0$ and $\widehat{\lambda}_{q+1} = \ldots = \widehat{\lambda}_p = 0$. We partition $\widehat{\boldsymbol{U}} = [\widehat{\boldsymbol{U}}_q, \widehat{\boldsymbol{U}}_{p-q}]$, where the columns of $\widehat{\boldsymbol{U}}_q \in \mathbb{R}_{p\times q}$ correspond to the q nonzero eigenvalues of $\widehat{\boldsymbol{\Sigma}}$, and the columns of $\widehat{\boldsymbol{U}}_{p-q} \in \mathbb{R}_{p\times(p-q)}$ correspond to the latter $p-q$ eigenvalues. Furthermore, recall that $\widehat{\boldsymbol{\Sigma}}^{+} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}^{+}\widehat{\boldsymbol{U}}'$, where $\widehat{\boldsymbol{\Lambda}}^{+} = \text{diag}(\widehat{\lambda}^1, \ldots, \widehat{\lambda}^p)$ and

$$\widehat{\lambda}^j = \begin{cases} 1/\widehat{\lambda}_j, & \widehat{\lambda}_j > 0 \\ 0, & \widehat{\lambda}_j = 0 \end{cases}$$

for $j = 1, \ldots, p$. Now, writing $\boldsymbol{Q}_{PD} = \widehat{\boldsymbol{U}}_q'\widehat{\boldsymbol{\Lambda}}^{+/2}$, we have that $\boldsymbol{Q}_{PD}$ is a simultaneous diagonalizer in a q-dimensional feature subspace so that $\boldsymbol{Q}_{PD}\widehat{\boldsymbol{\Sigma}}\boldsymbol{Q}_{PD}' = \boldsymbol{I}_q$. Thus, by transforming the feature vectors from the $K = 2$ classes with $\boldsymbol{Q}_{PD}$, we improve the equal diagonal covariance matrix assumption. We denote the linearly transformed MLE of $\boldsymbol{\mu}_k$, $k = 1, 2$, by $\widehat{\boldsymbol{\mu}}_k^{PD} = \boldsymbol{Q}_{PD}\widehat{\boldsymbol{\mu}}_K \in \mathbb{R}_{q\times 1}$. Our proposed *Pool-Diag* classifier is

$$\widehat{D}^{PD}(\boldsymbol{x}) = \underset{k\in\{1,2\}}{\arg\min}(\boldsymbol{Q}_{PD}\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k^{PD})'(\boldsymbol{Q}_{PD}\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k^{PD}).$$

4.4 Monte Carlo Simulations

We contrasted the classification performance of our proposed *SimDiag* and *Pool-Diag* classifiers with the *DLDA*, *SDLDA*, and *SmDLDA* classifiers using four well-known microarray data sets, each consisting of $K = 2$ classes. To compute an error rate for classifier comparison, we followed a similar approach to that of Pang et al. [2009], Ramey and Young [2011], and Dudoit et al. [2002]. We randomly

partitioned a microarray data set into a training data set, where the kth class contained n_k observations, $k = 1, 2$, and a test data set comprised of the remaining observations. Then, we applied the variable selection approach from Dudoit et al. [2002] to the training data set to obtain the set of q' genes that exhibited the largest ratio of their between-group to within-group sums of squares and reduced the test data set to the same q' genes. We then constructed the competing classifiers from the training data set and calculated the proportion of incorrectly classified test observations to obtain an estimate for the conditional error rate. We calculated the conditional error rate estimates for 1000 random partitions of a microarray data set and estimated the expected error rate. We used version 2.15 of the open source statistical software `R` for all simulations in this paper. Also, we used the `R` package ggplot2 from Wickham [2009] to create the summary graphics. We remark that our Monte Carlo simulation procedure avoids the selection bias discussed in Ambroise and McLachlan [2002]. Here, we describe the four microarray data sets:

4.4.1 Alon Data Set

Alon et al. [1999] have examined the gene expression profiles measured with an Affymetrix oligonucleotide array for 40 tumor and 22 normal colon tissues for 6,500 human genes. We follow Alon et al. [1999] and restrict the data set to the 2,000 genes with the highest minimal intensity across the samples.

4.4.2 Chiaretti Data Set

Chiaretti et al. [2004] have presented a data set that contains the gene expression levels for 128 individuals with acute lymphoblastic leukemia (ALL). Following Xu, Brock, and Parrish [2009], we consider the $n_1 = 42$ observations labeled *NEG* and the $n_2 = 37$ observations labeled *BCR/ABL*, which have been obtained from Affymetrix human 95Av2 arrays. The robust multichip average (*RMA*) normalization method has been applied to all 12,625 gene expression levels.

4.4.3 Golub Data Set

Golub et al. [1999] have examined the gene expression levels for 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). Bone marrow samples from each patient were assayed using Affymetrix Hgu6800 chips. We used the merged version of the data set from the `golubEsets` package on Bioconductor.

4.4.4 Singh Data Set

Singh et al. [2002] have presented a data set that consists of 235 radical prostatectomy specimens from surgery patients between 1995 and 1997. Oligonucleotide microarrays containing probes for approximately 12,600 genes and expressed sequence tags were used. We consider 102 of the radical prostatectomy specimens that the authors reported as high-quality to obtain a data set consisting of 52 prostate tumor samples and 50 prostate non-tumor samples.

4.4.5 Simulation Results

In Figures 4.1 and 4.2 we display the estimated unconditional error rates of the competing classifiers for $n_k = 10$ and $n_k = 20$, $k = 1, 2$, respectively, as a function of q' to demonstrate the classifier performance as the dimension of the data increases. Similar to Pang et al. [2009] and Dudoit et al. [2002], we examined the classification performance of the competing classifiers for $q' = 50, 100, \ldots, 250$.

For the Alon data set, the *SimDiag* and *Pool-Diag* classifiers outperformed the competing classifiers for all values of n_k and q' considered. The improved mean and variance estimators employed with the *SDLDA* and *SmDLDA* classifiers, respectively, did not improve the classification performance of the *DLDA* classifier, so that the *DLDA*, *SDLDA*, and *SmDLDA* classifiers yielded similar estimated unconditional error rates. However, the simultaneous diagonalization employed with the *SimDiag* classifier yielded large improvements to the *DLDA* classifier, resulting in

estimated unconditional error rates up to 0.15 less than those of the *DLDA*, *SDLDA*, and *SmDLDA* classifiers.

For the Chiaretti data set, the *Pool-Diag* classifier yielded superior classification performance compared to the competing classifiers for $n_k = 10$, $k = 1, 2$. For $n_k = 20$, the *SimDiag* and *Pool-Diag* classifiers yielded similar estimated unconditional error rates and outperformed the competing diagonal classifiers. The *DLDA* and *SmDLDA* classifiers once again yielded similar estimated unconditional error rates. Although the *SDLDA* classifier attained superior classification performance compared to the *DLDA* and *SmDLDA* classifiers for $n_k = 10$, $k = 1, 2$, the shrinkage estimators employed with the *SDLDA* classifier degraded the classification performance of the *DLDA* classifier for $n_k = 20$.

The *DLDA*, *SDLDA*, and *SmDLDA* classifiers again yielded similar results for the Golub data set. These three classifiers exhibited slightly better classification performances than the *SimDiag* and *Pool-Diag* classifiers for small values of q'. However, as q' increased, the *Pool-Diag* classifier yielded estimated unconditional error rates that were superior to the competing classifiers for $n_k = 10$ and comparable to the competing classifiers for $n_k = 20$.

For the Singh data set, the *SimDiag* and *Pool-Diag* classifiers yielded similar estimated unconditional error rates for $q' \geq 100$. Moreover, the simultaneous diagonalization of the sample covariance matrices employed with the *SimDiag* and *Pool-Diag* classifiers yielded classification performance superior to that of the *DLDA*, *SDLDA*, and *SmDLDA* classifiers.

4.5 Discussion

We have considered a family of diagonal linear classifiers that have been shown to attain excellent classification performance for small-sample, high-dimensional microarray data. In particular, we have examined the *DLDA* classifier popularized by

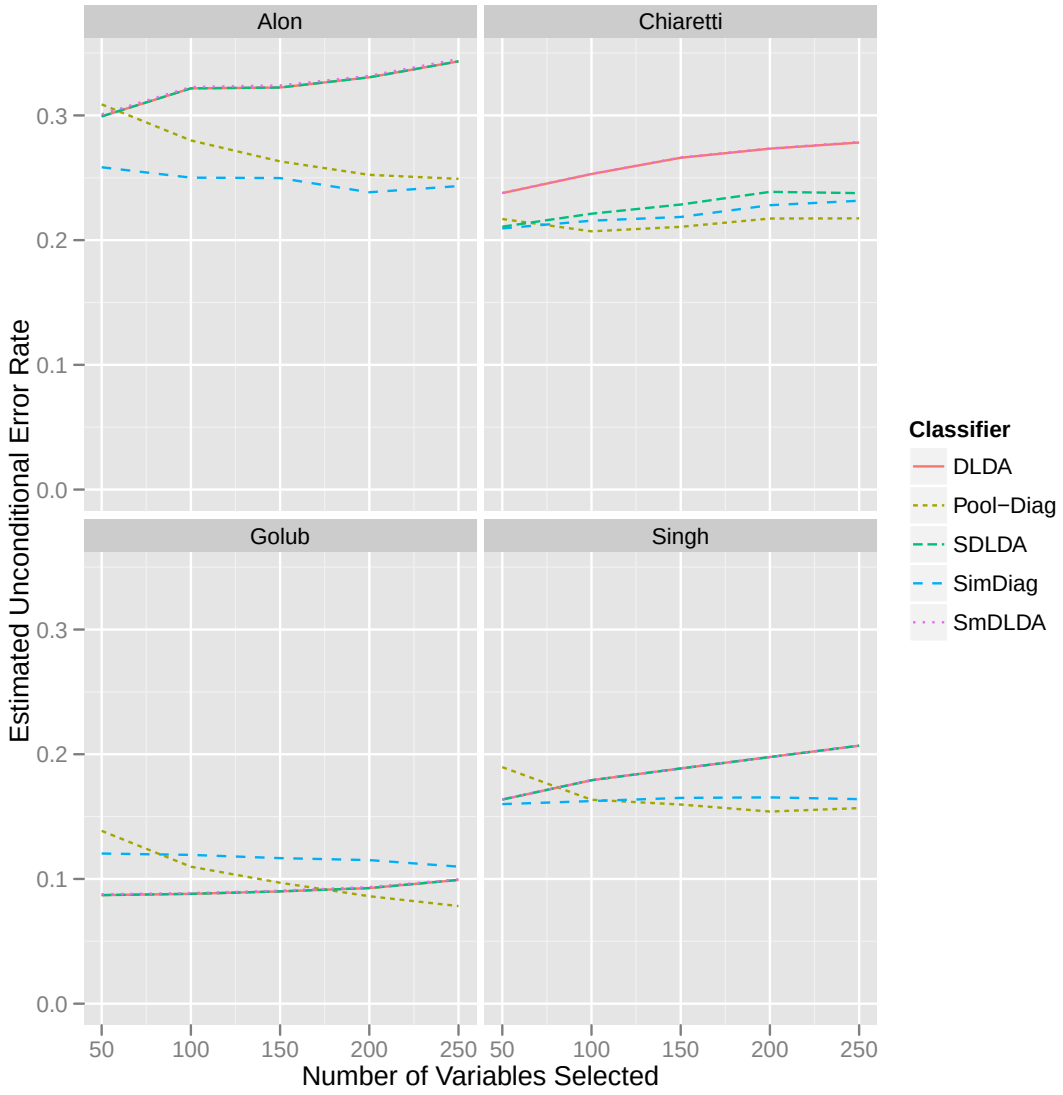

Figure 4.1: Estimated unconditional error rates as a function of the number of variable selected q' for $n_k = 10$, $k = 1, 2$.

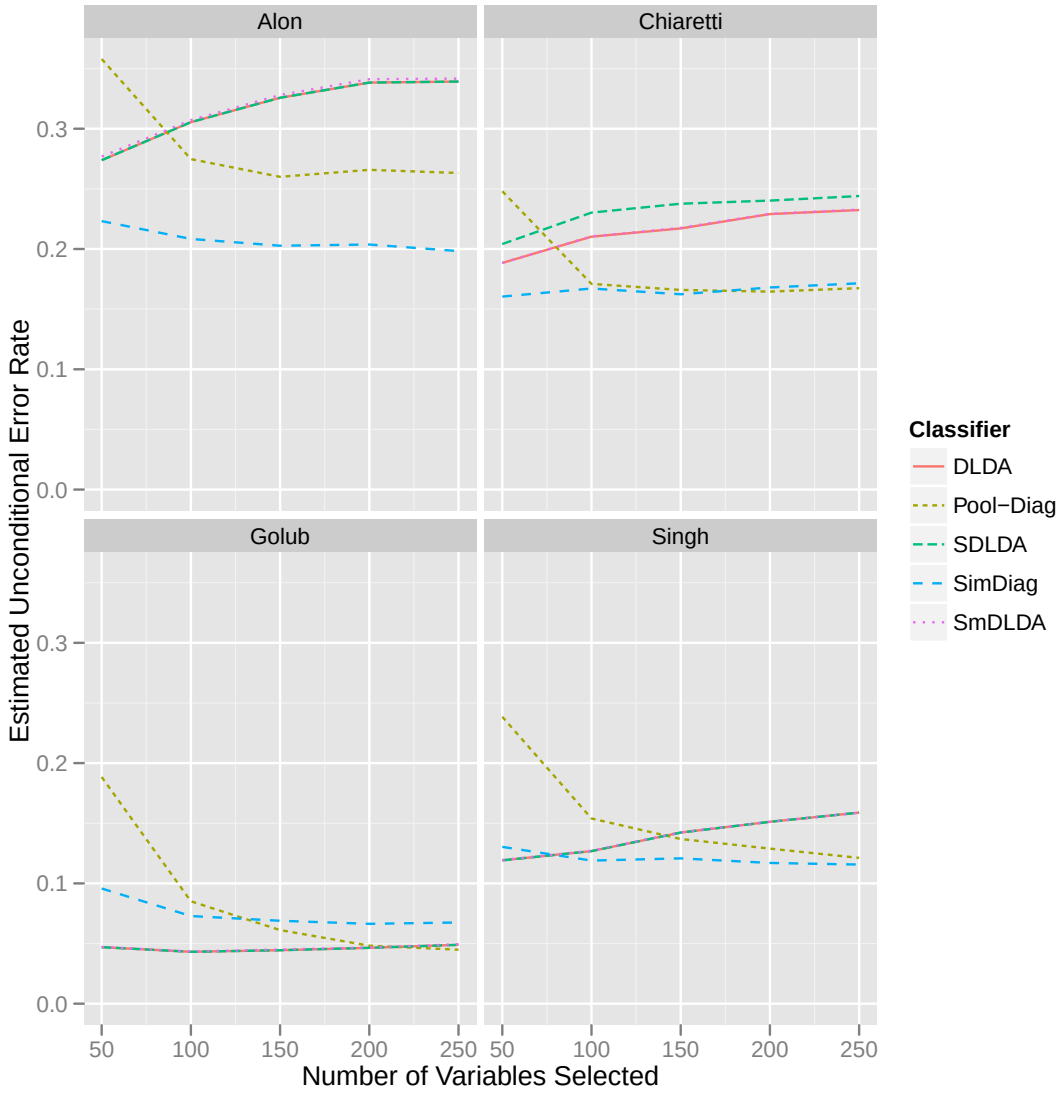

Figure 4.2: Estimated unconditional error rates as a function of the number of variable selected q' for $n_k = 20$, $k = 1, 2$.

Dudoit et al. [2002], the *SDLDA* classifier from Pang et al. [2009], and the *SmDLDA* classifier from Tong et al. [2012]. Although these classifiers have been shown to perform well in other classification studies, we have argued that the preservation of classificatory information present in the off-diagonal elements can improve classification accuracy. Thus, we have proposed two classifiers that simultaneously diagonalize the sample covariance matrices of each class with a common linear transformation prior to employing the *DLDA* classifier.

We have demonstrated that our proposed *SimDiag* and *Pool-Diag* can yield improved classification performance over the three competing classifiers. Moreover, we have shown that improving the diagonal assumption of the *DLDA* classifier can yield superior classification performance compared to the improved variance estimation of the *SDLDA* classifier and the improved mean estimation of the *SmDLDA* classifier. We have also found that the classification performance of the *SmDLDA* classifier seldom differs from that of the *DLDA* classifier. Thus, the improved mean estimator of the *SmDLDA* classifier does not yield a better classifier than the *DLDA* classifier for the data sets considered here. Furthermore, the *SDLDA* classifier yielded similar estimated unconditional error rates to the *DLDA* and *SmDLDA* classifiers for the Alon, Chiaretti, and Singh data sets. For the Chiaretti data set, the *SDLDA* classifier yielded superior classification performance to the *DLDA* and *SmDLDA* classifiers for small training sample sizes. However, for the same data set, the *DLDA* and *SmDLDA* classifiers had classification performance superior to the *SDLDA* classifier for the larger training sample sizes considered.

CHAPTER FIVE

Discriminant Analysis with Simultaneous Diagonalization of Covariance Matrices

5.1 Introduction

Pang et al. [2009] have claimed that classification of patients into known classes is one of the most important statistical problems in cancer genomics. The task is important yet arduous because high-throughput microarrays yield gene expression data sets with p gene expression levels and N observations, where typically $p \gg N$. Modern gene expression data sets often have between 10,000 and 50,000 probes or probe sets obtained from a relatively small number of patients (e.g., 25-50 patients per group). Furthermore, the classification performance of standard supervised learning methods, such as the LDA classifier, degrades due to the *curse of dimensionality* [Bellman, 1961]. As a result, many researchers emphasize simple, parsimonious models to avoid the estimation of a large number of parameters relative to the training sample size, N. In fact, seemingly naive models can often perform well due to their simplicity. For example, with a naive assumption that features are uncorrelated within each class, Dudoit et al. [2002] have used a modification of the LDA classifier where the off-diagonal elements of the class covariance matrices are assumed to be zero. We refer to this supervised learning method as the *DLDA* classifier.

Dudoit et al. [2002], Pang et al. [2009], and Tong et al. [2012] have shown that the *DLDA* classifier performs well with high-dimensional gene expression data. Furthermore, Bickel and Levina [2004] have shown that the *DLDA* classifier is asymptotically superior to the LDA classifier for two multivariate normal populations with equal covariance matrices. Also, Pang et al. [2009] have proposed the *SDLDA* classifier and have shown that it can outperform the *DLDA*, support vector machines, and a k-nearest neighbors classifiers in many small-sample, high-dimensional situations.

Moreover, Tong et al. [2012] have asserted that their proposed *SmDLDA* classifier is superior to the *DLDA* classifier in terms of classification performance because of an improved mean estimator for high-dimensional data.

Although the *DLDA* classifier and its variants can perform well when $p \gg N$, we contend classificatory information is present in the off-diagonal elements of the covariance matrices, which if preserved, can yield gains in classification performance with the *DLDA* classifier. In this paper, we aim to preserve much of the classificatory information present in the covariances between genes expressions while taking advantage of the reduction in the number of estimated model parameters and the computational efficiency of the *DLDA* classifier. We propose alternative classifiers that can yield superior classification performance to the *DLDA* classifier and its variants from Pang et al. [2009] and Tong et al. [2012]. Specifically, prior to applying the *DLDA* classifier, we nearly diagonalize the sample covariance matrices of each class by constructing a linear projection matrix that transforms the feature space so that gene expression levels are nearly uncorrelated within each class.

Our proposed approach is motivated by the independent components analysis and blind source separation literatures, where a large number of methods have been proposed to construct a linear transformation that nearly simultaneously diagonalizes a set of K square matrices. We propose a classifier that utilizes a whitening transform [Duda, Hart, and Stork, 2001] to diagonalize a pooled sample covariance matrix estimator to improve the diagonal covariance matrix assumption. We also employ two simultaneous diagonalization algorithms from Asfari [2006] and Souloumiac [2009] to simultaneously diagonalize the sample covariance matrices from each class to improve the naive diagonal assumption prior to applying the *DLDA* classifier. In short, we first reduce the naiveté of the diagonal covariance matrix assumption before employing the *DLDA* classifier to improve its classification performance.

We have organized the remainder of the paper as follows. In Section 2, we present and discuss the details of the *DLDA* classifier as well as the methods from Pang et al. [2009] and Tong et al. [2012]. In Section 3, we discuss the simultaneous diagonalization of covariance matrices and our proposed methods. We then compare the classification performance of our proposed methods with the *DLDA*, *SDLDA*, and *SmDLDA* classifiers with two well-known microarray data sets in Section 4 and conclude with some brief comments in Section 5.

5.2 Discriminant Analysis

In discriminant analysis, also known as supervised learning, we attempt to classify an unlabeled p-dimensional observation $\boldsymbol{x} = (x_1, \ldots, x_p)'$ into one of K known groups or classes, where we assume that $\boldsymbol{x}$ belongs to class k with *a priori* probability π_k for $k = 1, 2, \ldots, K$, with $\sum_{k=1}^{K} \pi_k = 1$. We assume that the *a priori* probabilities π_k are equal for $k = 1, \ldots, K$. Let $\mathbb{R}_{m\times n}$ and $\mathbb{R}^{>}_{p\times p}$ denote the matrix space of all $m \times n$ matrices over the real field $\mathbb{R}$, the cone of $p \times p$ positive definite real matrices, respectively. Also, let $\boldsymbol{A}'$ and $\boldsymbol{A}^{+}$ denote the transpose and the Moore-Penrose pseudoinverse of the matrix $\boldsymbol{A} \in \mathbb{R}_{m\times n}$, respectively. We denote by $\boldsymbol{B} \circ \boldsymbol{D}$ the element-wise Hadamard product of $\boldsymbol{B}, \boldsymbol{D} \in \mathbb{R}_{m\times m}$ [Harville, 2008]. Additionally, we assume that we have drawn n_k independently and identically distributed random vectors from the kth class

$$\boldsymbol{x}_{k,1} \ldots, \boldsymbol{x}_{k,n_k} \overset{IID}{\sim} N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}_k \in \mathbb{R}_{p\times 1}$ and covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{>}_{p\times p}$. Specifically, we say that $\boldsymbol{x}_{k,i}$ is the p-dimensional collection of gene expressions for the ith sample from class k. We construct a supervised classifier (hereafter, *classifier*) from the $N = n_1 + n_2 + \ldots + n_K$ training observations to predict the class membership of $\boldsymbol{x}$. Typically, with gene expression data, the groups are the diseases or disease subtypes

under consideration. We estimate the unknown parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ with their maximum likelihood estimators (MLEs), $\widehat{\boldsymbol{\mu}}_k = (\widehat{\mu}_{k1}, \ldots, \widehat{\mu}_{kp})'$ and $\widehat{\boldsymbol{\Sigma}}_k$, respectively, where

$$\widehat{\boldsymbol{\mu}}_k = n_k^{-1} \sum_{i=1}^{n_k} \boldsymbol{x}_{k,i} \tag{5.1}$$

and

$$\widehat{\boldsymbol{\Sigma}}_k = n_k^{-1} \sum_{i=1}^{n_k} (\boldsymbol{x}_{k,i} - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_{k,i} - \widehat{\boldsymbol{\mu}}_k)'. \tag{5.2}$$

We further assume that $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} \in \mathbb{R}^{>}_{p \times p}$ (k = 1, . . . , K). The pooled sample covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{k=1}^{K} n_k \widehat{\boldsymbol{\Sigma}}_k \tag{5.3}$$

is the MLE of $\boldsymbol{\Sigma}$. The sample *LDA* classifier is defined as follows: assign an unlabeled observation $\boldsymbol{x}$ to class k using

$$\widehat{D}(\boldsymbol{x}) = \arg\min_k (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k)' \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k), \tag{5.4}$$

5.2.1 Diagonal Linear Discriminant Analysis

For $p > n$, the pooled sample covariance matrix estimator in (5.3) is singular, and, hence, equation (5.4) is ill-posed and incalculable. Often, we employ some combination of variable selection, dimension reduction, and covariance matrix regularization to ensure that (5.3) is positive definite [Ramey and Young, 2011]. By assuming that $\boldsymbol{\Sigma}$ is a diagonal matrix, we have another effective approach for stabilizing (5.4). In particular, we assume that the gene expressions within class k are uncorrelated and, therefore, that the off-diagonal elements of $\boldsymbol{\Sigma}_k$ are zero, $k = 1, \ldots, K$. That is, we assume that $\boldsymbol{\Sigma}_k = \text{diag}(\sigma_{k1}^2, \sigma_{k2}^2, \ldots, \sigma_{kp}^2)$, where σ_{kj}^2 is the true marginal variance of the jth gene expression ($j = 1, \ldots, p$) of the kth class, and, thus, $\widehat{\boldsymbol{\Sigma}}_k = \text{diag}(\widehat{\sigma}_{k1}^2, \widehat{\sigma}_{k2}^2, \ldots, \widehat{\sigma}_{kp}^2)$. Returning to our assumption of equal covariance matrices, we have that the diagonal pooled sample covariance matrix is

$\widehat{\mathbf{\Sigma}} = \text{diag}(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \ldots, \widehat{\sigma}_p^2)$, where $\widehat{\sigma}_j^2 = N^{-1} \sum_{k=1}^{K} n_k \widehat{\sigma}_{kj}^2$. Hence, with the diagonal covariance matrix assumption, (5.4) reduces to the *DLDA* classifier

$$\widehat{D}^{DLDA}(\boldsymbol{x}) = \sum_{j=1}^{p} (x_j - \widehat{\mu}_{kj})^2 / \widehat{\sigma}_j^2. \tag{5.5}$$

Our assumption of equal, diagonal covariance matrices reduces the number of covariance parameters to estimate from $Kp(p+1)/2$ to p. Furthermore, the inverse of $\widehat{\mathbf{\Sigma}}$ exists and (5.5) is easily and quickly calculated as a dot product.

5.2.2 Shrinkage-based Diagonal Linear Discriminant Analysis

To improve the estimation of the diagonal covariance matrices for $p \gg N$, Pang et al. [2009] have proposed the *SDLDA* classifier based on a family of shrinkage-based estimators from Tong and Wang [2007]. Pang et al. [2009] have demonstrated that their proposed classifier is superior to a support vector machine classifier and a k-nearest neighbors classifier in terms of classification performance on several small-sample simulated and real microarray data sets. Following Pang et al. [2009], we write the shrinkage-based estimator for σ_j^{2t} as

$$\tilde{\sigma}_j^{2t}(\alpha) = \left\{ h_{\nu,p}(t) \widehat{\sigma}_{pool}^{2t} \right\}^{\alpha} \left\{ h_{\nu,1}(t) \widehat{\sigma}_j^{2t} \right\}^{1-\alpha}, \tag{5.6}$$

where $t \in \mathbb{R}$, $\alpha \in [0,1]$, $\nu = N - K$, $\widehat{\sigma}_{pool}^{2t} = \prod_{j=1}^{p} \left(\widehat{\sigma}_j^2\right)^{t/p}$ is a pooled variance estimator,

$$h_{\nu,p}(t) = (\nu/2)^t \left(\frac{\Gamma(\nu/2)}{\Gamma(\nu/2 + t/2)} \right)^p, \tag{5.7}$$

and $\Gamma(\cdot)$ is the gamma function.

The term (5.7) is a bias-correction term such that $h_{\nu,1}(t)\widehat{\sigma}_j^{2t}$ is an unbiased estimator for σ_j^{2t}, and $h_{\nu,p}(t)\widehat{\sigma}_{pool}^{2t}$ is an unbiased estimator for σ^{2t} when $\sigma_j^2 = \sigma^2$, $j = 1, \ldots p$. The shrinkage parameter α controls the amount of shrinkage from the individual variance estimator toward the bias-corrected pooled estimator. Tong and Wang [2007] have shown that for fixed p, ν, and $t > -\nu/2$, there exists a unique

optimal α with respect to the risk function $R_{Stein}(\boldsymbol{\sigma}^{2t}, \tilde{\boldsymbol{\sigma}}^{2t})$ corresponding to the Stein loss function $L_{Stein}(\sigma^2, \tilde{\sigma}^2) = \tilde{\sigma}^2/\sigma^2 - \ln(\tilde{\sigma}^2/\sigma^2) - 1$, where $\boldsymbol{\sigma}^{2t} = (\sigma_1^{2t}, \ldots, \sigma_p^{2t})'$ and $\tilde{\boldsymbol{\sigma}}^{2t} = (\tilde{\sigma}_1^{2t}, \ldots, \tilde{\sigma}_p^{2t})'$. For brevity, we do not include the Stein risk function but note that it is an implicit function of α.

To estimate α, Pang et al. [2009] have proposed that α^* be chosen from a grid of candidate parameters such that

$$\alpha^* = \underset{\alpha \in [0,1]}{\arg\min}\, R_{Stein}(\boldsymbol{\sigma}^{2t}, \tilde{\boldsymbol{\sigma}}^{2t}). \tag{5.8}$$

Even for large p, the empirical minimization of the Stein risk function requires little computation time. Using (5.8) with $t = -1$, Pang et al. [2009] have substituted (5.6) for $\widehat{\sigma}_j^{-2}$ into (5.5) to obtain the *SDLDA* classifier

$$\widehat{D}^{SDLDA}(\boldsymbol{x}) = \underset{k}{\arg\min} \sum_{j=1}^{p} (x_j - \widehat{\mu}_{kj})^2 \tilde{\sigma}_j^{-2}(\alpha^*).$$

In our simulations in Section 5.4, we estimate α using a grid of 21 equidistant candidate values between 0 and 1, inclusively.

5.2.3 The DLDA Classifier with Improved Mean Estimation

While Pang et al. [2009] have sought to improve the *DLDA* classifier with improved estimators of the marginal variances, Tong et al. [2012] have argued that the MLEs for $\boldsymbol{\mu}_k$, $k = 1, \ldots, K$, are unreliable within the *DLDA* classifier for $p \gg N$. Instead, Tong et al. [2012] have considered improved mean estimators with a shrinkage parameter optimized under quadratic loss as $p \to \infty$. Specifically, Tong et al. [2012] have remarked that James-Stein mean estimators of the form

$$\widehat{\boldsymbol{\mu}}_k^{(JS)} = \left(1 - \frac{(p-2)/n_k}{||\widehat{\boldsymbol{\mu}}_k||_{\widehat{\boldsymbol{\Sigma}}}^2}\right) \widehat{\boldsymbol{\mu}}_k,$$

where $||\boldsymbol{z}||_{\boldsymbol{A}}^2 = \boldsymbol{z}'\boldsymbol{A}^{-1}\boldsymbol{z}$ for $\boldsymbol{z} \in \mathbb{R}_{p\times 1}$ and $\boldsymbol{A} \in \mathbb{R}_{p\times p}^{>}$, cannot be used when $p > N$ because $\widehat{\boldsymbol{\Sigma}}$ is singular. Hence, Tong et al. [2012] have utilized a similar family of

estimators, where $\boldsymbol{\Sigma}$ is assumed to be diagonal and positive definite. Under this assumption, Tong et al. [2012] have proposed the estimator

$$\widehat{\boldsymbol{\mu}}_k(\widehat{r}_k) = \left(1 - \frac{\widehat{r}_k}{||\widehat{\boldsymbol{\mu}}_k||^2_{\widehat{\boldsymbol{\Sigma}} \circ \boldsymbol{I}_p}}\right) \widehat{\boldsymbol{\mu}}_k, \tag{5.9}$$

where $\widehat{r}_k = (n_k - 1)(p-2)/\{n_k(n_k - 3)\}$ is a shrinkage estimator optimized under a quadratic loss function with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma} \circ \boldsymbol{I}_p$ as $p \to \infty$. Substituting (5.9) for $\widehat{\boldsymbol{\mu}}_k$ in (5.5), we have the *SmDLDA* classifier

$$\widehat{D}^{SmDLDA}(\boldsymbol{x}) = \{\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k(\widehat{r}_k)\}' \left(\widehat{\boldsymbol{\Sigma}} \circ \boldsymbol{I}_p\right)^{-1} \{\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k(\widehat{r}_k)\}. \tag{5.10}$$

5.3 Nearly Diagonal Discriminant Analysis

As we have discussed, the *DLDA*, *SDLDA*, and *SmDLDA* classifiers have been developed for small-sample, high-dimensional microarray data sets with the naive assumption of diagonal covariance matrices. However, we contend that if we can preserve classificatory information present in the off-diagonal elements of the covariance matrices before applying the *DLDA* classifier, then we can improve its classification performance. Thus, we remark that a reasonable approach is to apply a common linear transformation to each population such that the off-diagonal elements of the resulting covariance matrices are nearly zero. We argue that this approach preserves classificatory information in the off-diagonal elements of the covariance matrices. More precisely, let $\boldsymbol{B} \in \mathbb{R}_{q \times p}$, and notice that $\boldsymbol{B}\boldsymbol{x} \sim N_q(\boldsymbol{B}\boldsymbol{\mu}_k, \boldsymbol{B}\boldsymbol{\Sigma}_k\boldsymbol{B}' \circ \boldsymbol{I}_q)$ for all $k = 1, \ldots, K$. Rather than restricting the covariance matrices of the original populations to be diagonal, we desire that the covariance matrices of the transformed populations are diagonal. Hence, we wish to determine a $\boldsymbol{B} \in \mathbb{R}_{q \times p}$ that improves our goal of diagonal covariance matrices.

We say that the matrix $\boldsymbol{B}$ nearly decorrelates the kth class if, for a given $\epsilon > 0$, $||\boldsymbol{B}\boldsymbol{\Sigma}_k\boldsymbol{B}' - \boldsymbol{B}\boldsymbol{\Sigma}_k\boldsymbol{B}' \circ \boldsymbol{I}_q|| < \epsilon$ for $k = 1, \ldots K$, where $||\cdot||$ is a specified matrix norm. Intuitively, our proposed method should preserve the classificatory information in the off-diagonal elements if $||\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k \circ \boldsymbol{I}_p|| > ||\boldsymbol{B}\boldsymbol{\Sigma}_k\boldsymbol{B}' - \boldsymbol{B}\boldsymbol{\Sigma}_k\boldsymbol{B}' \circ \boldsymbol{I}_q||$ for

$k = 1, \ldots, K$. In the special case that $\boldsymbol{B}\boldsymbol{\Sigma}_k\boldsymbol{B}' = \boldsymbol{B}\boldsymbol{\Sigma}_k\boldsymbol{B}' \circ \boldsymbol{I}_q$, $k = 1, \ldots, K$, we say that the set of matrices $\{\boldsymbol{\Sigma}_k | k = 1, \ldots, K\}$ is simultaneously diagonalizable and that $\boldsymbol{B}$ is a *simultaneous diagonalizer*.

Harville [2008] and Anderson [2003] have discussed approaches to simultaneously diagonalize $K = 2$ square matrices under the assumption that at least one of the matrices is positive definite. Contrarily, we consider the general case of $K \geq 2$. Harville [2008] has stated that a necessary condition for the simultaneous diagonalization of K square matrices is pairwise commutativity, i.e., $\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_{k'} = \boldsymbol{\Sigma}_{k'}\boldsymbol{\Sigma}_k$ for all $1 \leq k, k' \leq K$. Furthermore, Harville [2008] has stated that pairwise commutativity is a sufficient condition for simultaneous diagonalization where $\boldsymbol{B}$ is orthogonal. Rather than requiring pairwise commutativity, our approach to select $\boldsymbol{B}$ is motivated by the independent components analysis and blind source separation literatures, where a large number of algorithms, often based on a Jacobi diagonalization method, have been proposed to generate a sequence of candidate simultaneous diagonalizers that successively yield a nearer simultaneous diagonalization of K square matrices.

In this paper, we propose three new classifiers. The first classifier is based on our proposed *Pool-Diag* simultaneous diagonalization method, and the two additional classifiers are based on the near simultaneous diagonalization algorithms of Souloumiac [2009] and Asfari [2006]. Because these latter two simultaneous diagonalization algorithms require a large number of computations for large p, we reduce the p-dimensional feature space to a q-dimensional subspace using the dimension reduction method of Young et al. [1987] prior to applying the simultaneous diagonalization algorithms. Below, we describe our proposed *Pool-Diag* classifier, the dimension reduction method from Young et al. [1987], and our proposed *Asfari* and *Souloumiac* classifiers.

5.3.1 The Pool-Diag Classifier

For $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, $k = 1, \ldots, K \geq 2$, consider the spectral decomposition of $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p) \in \mathbb{R}_{p\times p}$ is the matrix of eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{U} \in \mathbb{R}_{p\times p}$ is the matrix of the corresponding eigenvectors of $\boldsymbol{\Sigma}$, where $\lambda_1 \geq \ldots \geq \lambda_p > 0$ and $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_p$. Thus, the whitening transform $\boldsymbol{B} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}' = \boldsymbol{U}'\boldsymbol{\Lambda}^{-1/2}$ diagonalizes $\boldsymbol{\Sigma}$ such that $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}' = \boldsymbol{I}_p = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}' \circ \boldsymbol{I}_p$ [Duda, Hart, and Stork, 2001].

As discussed in Section 5.2, $\widehat{\boldsymbol{\Sigma}}^{-1}$ does not exist for $p > N$ because $\text{rank}(\widehat{\boldsymbol{\Sigma}}) = q < p$, where $\widehat{\boldsymbol{\Sigma}}$ is given in (5.2). Let $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}'$ be the spectral decomposition of $\widehat{\boldsymbol{\Sigma}}$, where $\widehat{\boldsymbol{\Lambda}} = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p) \in \mathbb{R}_{p\times p}$ is the matrix of eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{U}} \in \mathbb{R}_{p\times p}$ is the matrix of the corresponding eigenvectors of $\widehat{\boldsymbol{\Sigma}}$, where $\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}} = \boldsymbol{I}_p$. Recall that $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_q > 0$ and $\widehat{\lambda}_{q+1} = \ldots = \widehat{\lambda}_p = 0$. We partition $\widehat{\boldsymbol{U}} = [\widehat{\boldsymbol{U}}_q, \widehat{\boldsymbol{U}}_{p-q}]$, where the columns of $\widehat{\boldsymbol{U}}_q \in \mathbb{R}_{p\times q}$ correspond to the q nonzero eigenvalues of $\widehat{\boldsymbol{\Sigma}}$, and the columns of $\widehat{\boldsymbol{U}}_{p-q} \in \mathbb{R}_{p\times(p-q)}$ correspond to the latter $p-q$ eigenvalues. Furthermore, recall that $\widehat{\boldsymbol{\Sigma}}^{+} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}^{+}\widehat{\boldsymbol{U}}'$, where $\widehat{\boldsymbol{\Lambda}}^{+} = \text{diag}(\widehat{\lambda}^1, \ldots, \widehat{\lambda}^p)$ and

$$\widehat{\lambda}^j = \begin{cases} 1/\widehat{\lambda}_j, & \widehat{\lambda}_j > 0 \\ 0, & \widehat{\lambda}_j = 0 \end{cases}$$

for $j = 1, \ldots, p$. Now, writing $\boldsymbol{B}_{PD} = \widehat{\boldsymbol{U}}_q'\widehat{\boldsymbol{\Lambda}}^{+/2}$, we have that $\boldsymbol{B}_{PD}$ is a whitening, simultaneous diagonalizer in a q-dimensional feature subspace so that $\boldsymbol{B}_{PD}\widehat{\boldsymbol{\Sigma}}\boldsymbol{B}_{PD}' = \boldsymbol{I}_q$. Thus, by transforming the feature vectors from the K classes with $\boldsymbol{B}_{PD}$, we improve the diagonal, equal covariance matrix assumption. We denote the linearly transformed MLE of $\boldsymbol{\mu}_k$ by $\widehat{\boldsymbol{\mu}}_k^{PD} = \boldsymbol{B}_{PD}\widehat{\boldsymbol{\mu}}_K \in \mathbb{R}_{q\times 1}$, we have our proposed *Pool-Diag* classifier

$$\widehat{D}^{PD}(\boldsymbol{x}) = \arg\min_k (\boldsymbol{B}_{PD}\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k^{PD})'(\boldsymbol{B}_{PD}\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k^{PD}).$$

5.3.2 The M-Method for Low-Dimensional Projection

To reduce the runtime of the simultaneous diagonalization algorithms, our goal is to reduce the p-dimensional feature space to a q-dimensional subspace via a linear

transformation, $\boldsymbol{U}_q' \in \mathbb{R}_{q\times p}$. We obtain the matrix $\boldsymbol{U}_q'$ with the $\boldsymbol{M}$-method from Young et al. [1987].

Let $\boldsymbol{M} \in \mathbb{R}_{p\times(K-1)(p+1)}$ be defined as

$$\boldsymbol{M} = [\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 | \boldsymbol{\mu}_3 - \boldsymbol{\mu}_1 | \dots | \boldsymbol{\mu}_K - \boldsymbol{\mu}_1 | \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1 | \boldsymbol{\Sigma}_3 - \boldsymbol{\Sigma}_1 | \dots | \boldsymbol{\Sigma}_K - \boldsymbol{\Sigma}_1], \quad (5.11)$$

with $\text{rank}(\boldsymbol{M}) = m < p$. We assume that $\boldsymbol{\Sigma}_k \neq \boldsymbol{\Sigma}_1$ for at least one value of k, where $2 \leq k \leq K$. Next, using the full-rank decomposition of $\boldsymbol{M}$ [Harville, 2008], we write $\boldsymbol{M} = \boldsymbol{F}_m \boldsymbol{G}_m$, where $\boldsymbol{F}_m \in \mathbb{R}_{p\times m}$ and $\boldsymbol{G}_m \in \mathbb{R}_{m\times(K-1)(p+1)}$. Young et al. [1987] have shown that the Bayes rule based on $\boldsymbol{F}_m^+ \boldsymbol{x} \in \mathbb{R}_{m\times 1}$ attains the same Bayes error rate as the Bayes rule based on $\boldsymbol{x}$. Moreover, Young et al. [1987] have shown that m is the minimum reduced dimension for which the Bayes error rate is not increased for the Bayes rule based on $\boldsymbol{F}_m^+ \boldsymbol{x}$. However, we often wish to find a value smaller than m such that most of the classificatory information is preserved. Young et al. [1987] have proposed that one utilize the matrix $\boldsymbol{M}_q \in \mathbb{R}_{p\times(K-1)(p+1)}$, which best approximates $\boldsymbol{M}$ with respect to the Frobenius norm. That is, let $\boldsymbol{M}_q = \boldsymbol{U}_q \boldsymbol{\Lambda}_q \boldsymbol{V}_q'$ be the singular value decomposition of $\boldsymbol{M}_q$, where $\boldsymbol{U}_q \in \mathbb{R}_{p\times p}$, $\boldsymbol{\Lambda}_q \in \mathbb{R}_{p\times p}$ is the diagonal matrix of singular values of $\boldsymbol{M}_q$, and $\boldsymbol{V}_q \in \mathbb{R}_{(K-1)(p+1)\times p}$, such that the columns of each of $\boldsymbol{U}_q$ and $\boldsymbol{V}_q$ are orthonormal. Young et al. [1987] have shown $\mathcal{C}(\boldsymbol{U}_q') = \mathcal{C}(\boldsymbol{F}_q^+)$, where $\mathcal{C}(\boldsymbol{A})$ denotes the column space of $\boldsymbol{A} \in \mathbb{R}_{m\times n}$. Hence, for $q < m$, we lose only a small amount of classificatory information with the Bayes rule based on the dimension-reduced data $\boldsymbol{U}_q' \boldsymbol{x}$.

In practice we estimate the unknown $\boldsymbol{M}_q$ with its MLE, $\widehat{\boldsymbol{M}_q}$, by substituting $\widehat{\boldsymbol{\mu}_k}$ given in (5.1) for $\boldsymbol{\mu}_k$ and $\widehat{\boldsymbol{\Sigma}_k}$ given in (5.2) for $\boldsymbol{\Sigma}_k$ in (5.11) for $k = 1, \dots, K$. Furthermore, for $k = 1, \dots, K$, we denote the dimension-reduced sample covariance matrix by

$$\widehat{\boldsymbol{\Sigma}}_k^{(q)} = \boldsymbol{U}_q' \widehat{\boldsymbol{\Sigma}}_k \boldsymbol{U}_q, \quad (5.12)$$

where $\widehat{\boldsymbol{\Sigma}}_k$ is given in (5.2). We choose q to be the 95th percentile of the singular values of the matrix $\widehat{\boldsymbol{M}}_q$ analogous to a typical approach employed with principal component analysis [Izenman, 2008]. Similarly, we have

$$\widehat{\boldsymbol{\mu}}_k^{(q)} = \boldsymbol{U}_q' \widehat{\boldsymbol{\mu}}_k, \tag{5.13}$$

which is the linearly transformed MLE for $\boldsymbol{\mu}_k$. For additional insight into the $\boldsymbol{M}$-method, see Young et al. [1987] and McLachlan [1992].

5.3.3 The Asfari Classifier

Asfari [2006] has presented a nonorthogonal extension to the joint approximate diagonalization eigenmatrices (*JADE*) algorithm of Cardoso and Souloumiac [1996]. The Asfari [2006] algorithm, known as the *LUJ1D* algorithm, is based on the LU and QR matrix decompositions [Harville, 2008] and solves a least-squares optimization problem with the objective function

$$L(\boldsymbol{B}) = \sum_{k=1}^{K} ||\boldsymbol{B}\boldsymbol{A}_k\boldsymbol{B}' - \boldsymbol{B}\boldsymbol{A}_k\boldsymbol{B}' \circ \boldsymbol{I}_p||_F^2, \tag{5.14}$$

where $||\boldsymbol{A}||_F^2$ denotes the squared-Frobenius norm of $\boldsymbol{A} \in \mathbb{R}_{q\times q}$. We remark that (5.14) is the sum of the squared off-diagonal elements of $\boldsymbol{B}\boldsymbol{A}_k\boldsymbol{B}'$ for $k = 1, \ldots, K$. We write the solution to (5.14) as $\boldsymbol{B}_A \in \mathbb{R}_{q\times q}$.

As we have discussed above, we apply the *Asfari* simultaneous diagonalization algorithm to the dimension-reduced data after employing the M-method of Young et al. [1987]. That is, substituting (5.12) for $\boldsymbol{A}_k$ ($k = 1, \ldots, K$) in (5.14), we obtain the near simultaneous diagonalizer $\boldsymbol{B}_A$. After employing the near simultaneous diagonalizer $\boldsymbol{B}_A$, we transform (5.13) to obtain $\widehat{\boldsymbol{\mu}}_k^A = \boldsymbol{B}_A\widehat{\boldsymbol{\mu}}_k^{(q)}$ for $k = 1, \ldots, K$. Therefore, we write our proposed *Asfari* classifier as

$$\widehat{D}^{Asfari}(\boldsymbol{x}) = \arg\min_k (\boldsymbol{B}_A\boldsymbol{U}_q'\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k^A)' \left(\boldsymbol{B}_A\widehat{\boldsymbol{\Sigma}}_k^{(q)}\boldsymbol{B}_A'\right)^{-1} (\boldsymbol{B}_A\boldsymbol{U}_q'\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k^A).$$

5.3.4 The Jedi Classifier

Souloumiac [2009] has developed the Jacobi-like *J-Di* algorithm that also extends the *JADE* algorithm to a nonorthogonal simultaneous diagonalizer by iteratively constructing successive Givens and hyperbolic rotations. Following the `jointDiag` R package, we refer to the *J-Di* algorithm as the *Jedi* method. For brevity we do not include the objective function or the *Jedi* algorithm but instead refer the reader to its derivation and discussion by Souloumiac [2009].

Similar to the *Asfari* classifier, the *Jedi* simultaneous diagonalization algorithm can be applied to (5.12) to obtain the near simultaneous diagonalizer $\boldsymbol{B}_J$. We transform (5.13) to obtain $\widehat{\boldsymbol{\mu}}_k^J = \boldsymbol{B}_J \widehat{\boldsymbol{\mu}}_k^{(q)}$ for $k = 1, \ldots, K$. Therefore, we write our proposed *Jedi* classifier as

$$\widehat{D}^{Jedi}(\boldsymbol{x}) = \arg\min_k (\boldsymbol{B}_J \boldsymbol{U}_q' \boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k^J)' \left(\boldsymbol{B}_J \widehat{\boldsymbol{\Sigma}}_k^{(q)} \boldsymbol{B}_J' \right)^{-1} (\boldsymbol{B}_J \boldsymbol{U}_q' \boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k^J).$$

5.4 Monte Carlo Simulations

We contrasted the classification performance of our proposed *Pool-Diag*, *Jedi*, and *Asfari* classifiers with the *DLDA*, *SDLDA*, and *SmDLDA* classifiers using the two well-known microarray data sets from Golub et al. [1999] and Yeoh et al. [2002]. First, we applied the variable selection approach from Dudoit et al. [2002] to the data set to obtain the set of q genes that exhibited the largest ratio of their between-group to within-group sums of squares. To compute an error rate for classifier comparison, we followed a similar approach to that of Dudoit et al. [2002] and randomly partitioned a data set into a training data set, consisting of four-fifths of the observations, and a test data set comprised of the remaining one-fifth of the observations. We stratified our random partition across each class to preserve the training sample sizes. For each of the random partitions, we calculated the proportion of incorrectly classified test observations to obtain an estimate for the conditional error rate. We calculated the conditional error rate estimates for 500

random partitions of the data and estimated the expected error rate. We used version 2.15 of the open source statistical software `R` for all simulations in this paper.

5.4.1 Golub Leukemia Data Set

Golub et al. [1999] have examined the gene expression levels for 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). Bone marrow samples from each patient were assayed using Affymetrix Hgu6800 chips. We remark that ALL results from both T-cell lymphocytes and B-cell lymphocytes. Of the 47 ALL samples, 9 are T-lineage ALL samples, and the other 38 are B-lineage ALL samples. We used the merged version of the data set that is available in the `golubEsets` package on Bioconductor.

5.4.2 St. Jude Leukemia Data Set

Yeoh et al. [2002] have obtained the diagnostic bone marrow samples from 248 pediatric ALL patients who were determined to have one and only one of the six known pediatric ALL prognostic subtypes, which include T-cell lineage ALL (T-ALL), E2A-PBX1, TEL-AML1, MLL rearrangements, BCR-ABL, and hyperdiploid karyotypes with more than 50 chromosomes (HK50). The 248 patients included 43 T-ALL, 27 E2A-PBX1, 79 TEL-AML1, 15 BCR-ABL, 20 MLL, and 64 HK50 patients. We obtained the St. Jude data set from the `stjudem` package on Bioconductor. Because the data set contains data from other patients as well as a baseline set of subjects who are determined to have no form of acute leukemia, we reduced the St. Jude data set to include only the six ALL subtypes.

5.4.3 Simulation Results

In Table 5.1 and 5.2 we see the estimated unconditional error rates of the considered classifiers for the Golub and St. Jude data sets, respectively. For the Golub data set, the *DLDA*, *SDLDA*, and *SmDLDA* classifiers each yielded the same

estimated unconditional error rate for the specified values of q. Effectively, the variance shrinkage of the *SDLDA* classifier and the alternative mean estimator of the *SmDLDA* classifier did not improve the classification performance of the *DLDA* classifier. In this case, the *Pool-Diag* classifier did not improve the classification performance of the *DLDA* classifier. However, by improving the diagonal covariance matrix assumption, the *Asfari* and *Jedi* classifiers had superior average classification performance over the *DLDA* classifier. Moreover, the *Jedi* classifier yielded the largest improvement to the *DLDA* classifier and resulted in the minimum observed average conditional error rate.

For the St. Jude data set, the *DLDA* and *SmDLDA* classifiers resulted in the same average conditional error rate. Also, the *SDLDA* classifier resulted in much larger average conditional error rates, especially for the larger values of q. Although the *Jedi* classifier did not perform as well in terms of classification, the *Asfari* classifier was evidently comparable to the *DLDA* and *SDLDA* classifiers. The *Pool-Diag* classifier achieved the minimal average conditional error rate. Hence, we improved the diagonal covariance matrix assumption of the *DLDA* classifier, and, thus, improved classification performance.

5.5 Conclusion

We have considered a family of diagonal linear classifiers that have been shown to have excellent classification performance for small-sample, high-dimensional microarray data. In particular, we have examined the *DLDA* classifier popularized by Dudoit et al. [2002], the *SDLDA* classifier from Pang et al. [2009], and the *SmDLDA* classifier from Tong et al. [2012]. Although these classifiers have been shown to perform well in other classification studies, we have argued that the preservation of classificatory information present in the off-diagonal elements can improve classification accuracy. Thus, we have proposed three approaches that simultaneously

diagonalize or nearly diagonalize the sample covariance matrices of each class and have demonstrated that our new diagonal classifiers can yield improved classification performance over a variety of recently proposed diagonal classifiers.

Rather than employing an alternative mean or variance estimator with the *DLDA* classifier, we have demonstrated that a better approach is to improve the diagonal assumption of the *DLDA* classifier. Specifically, we have shown that simultaneously diagonalizing the Golub data set before employing the *DLDA* classifier produced better classification performance than the *DLDA*, *SDLDA*, and *SmDLDA* classifiers. Furthermore, we have shown that by simply diagonalizing the pooled sample covariance matrix, we achieved better classification accuracy than the *DLDA*, *SDLDA*, and *SmDLDA* classifiers did.

Additionally, we have found that the *SDLDA* classifier can yield extremely large error rates. After further investigation, we have determined that the pooled variance estimator utilized in the *SDLDA* classifier can yield near-zero variance estimates that can cause numerical instabilities when test observations are classified. In this case the *SDLDA* classifier will often classify each test observation into a single class, resulting in an unreasonably large error rate, as we have observed with the St. Jude data set. Also, in our preliminary studies, we found that the alternative mean estimator of the *SmDLDA* classifier often resulted in better estimation of the population mean than the MLE did. However, we have also found that the classification performance of the *SmDLDA* classifier seldom differs from that of the *DLDA* classifier. Thus, the improved mean estimator of the *SmDLDA* classifier does not appear to yield a better classifier than the *DLDA* classifier for the data sets considered here.

Table 5.1: Estimated unconditional error rates with approximate standard errors in parentheses for the considered classifiers with the Golub data set.

	Classifier	q=50	q=100	q=250	q=500
1	DLDA	0.037 (0.008)	0.047 (0.009)	0.051 (0.010)	0.033 (0.008)
2	SmDLDA	0.037 (0.008)	0.047 (0.009)	0.051 (0.010)	0.033 (0.008)
3	SDLDA	0.037 (0.008)	0.047 (0.009)	0.051 (0.010)	0.033 (0.008)
4	Pool-Diag	0.048 (0.010)	0.052 (0.010)	0.052 (0.010)	0.044 (0.009)
5	Jedi	0.025 (0.007)	0.044 (0.009)	0.031 (0.008)	0.052 (0.010)
6	Asfari	0.031 (0.008)	0.041 (0.009)	0.048 (0.010)	0.043 (0.009)

Table 5.2: Estimated unconditional error rates with approximate standard errors in parentheses for the considered classifiers with the St. Jude data set.

	Classifier	q=50	q=100	q=250	q=500
1	DLDA	0.062 (0.011)	0.040 (0.009)	0.032 (0.008)	0.024 (0.007)
2	SmDLDA	0.062 (0.011)	0.040 (0.009)	0.032 (0.008)	0.024 (0.007)
3	SDLDA	0.659 (0.021)	0.587 (0.022)	0.941 (0.011)	0.941 (0.011)
4	Pool-Diag	0.056 (0.010)	0.027 (0.007)	0.015 (0.005)	0.016 (0.006)
5	Jedi	0.182 (0.017)	0.123 (0.015)	0.142 (0.016)	0.151 (0.016)
6	Asfari	0.075 (0.012)	0.043 (0.009)	0.040 (0.009)	0.030 (0.008)

BIBLIOGRAPHY

Aeberhard, S., Coomans, D., and Vel, O. D. (1993), "Improvements to the classification performance of RDA," *Journal of Chemometrics*, 7, 99–115.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999), "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, 96, 6745–6750.

Ambroise, C. and McLachlan, G. J. (2002), "Selection bias in gene extraction on the basis of microarray gene-expression data." *Proceedings of the National Academy of Sciences of the United States of America*, 99, 6562–6566.

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley-Interscience, 3rd ed.

Asfari, B. (2006), "Simple LU and QR based Non-Orthogonal Matrix Joint Diagonalization," in *Proceedings of ICA2006, Springer LNCS series*, pp. 1–7.

Bellman, R. (1961), *Adaptive Control Processes: A Guided Tour*, Princeton University Press.

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*.

Breiman, L. (1996), "Bagging predictors," *Machine Learning*, 24, 123–140.

Brock, G., Datta, S., Datta, S., and Pihur, V. (2008), "clValid: An R Package for Cluster Validation," *Journal of Statistical Software*, 25.

Cardoso, J.-F. and Souloumiac, A. (1996), "Jacobi angles for simultaneous diagonalization," *SIAM Journal on Matrix Analysis and Applications*, 17, 161–164.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004), "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," *Blood*, 103, 2771–2778.

Datta, S. and Datta, S. (2006), "Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes," *BMC Bioinformatics*, 7, 397–397.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001), *Pattern Classification*, Wiley-Interscience, New York, 2nd ed.

Dudoit, S. and Fridlyand, J. (2003), "Bagging to improve the accuracy of a clustering procedure," *Bioinformatics*, 19, 1090–1099.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002), "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *Journal of the American Statistical Association*, 97, 77–87.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

Efron, B. and Tibshirani, R. (1994), *An Introduction to the Bootstrap*, Chapman and Hall/CRC.

— (1997), "Improvements on cross-validation: The. 632+ bootstrap method," *Journal of the American Statistical Association*, 548–560.

Fisher, L. and Van Ness, J. W. (1971), "Admissible Clustering Procedures," *Biometrika*, 58, 91–104.

Fitzmaurice, G. M., Krzanowski, W. J., and Hand, D. J. (1991), "A Monte Carlo study of the 632 bootstrap estimator of error rate," *Journal of Classification*, 8, 239–250.

Fligner, M. A., Verducci, J. S., and Blower, P. E. (2002), "A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings," *Technometrics*, 44, 110–119.

Foster, I. (1995), *Designing and Building Parallel Programs: Concepts and Tools for Parallel Software Engineering*, Addison Wesley.

Fraley, C. and Raftery, A. (2006), "MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering," Tech. rep., Department of Statistics.

Fraley, C. and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

Friedman, J. H. (1989), "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, 84, 165–175.

Fu, W. J., Carroll, R. J., and Wang, S. (2005), "Estimating misclassification error with small samples via bootstrap cross-validation." *Bioinformatics*, 21, 1979–1986.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, Academic Press Inc., 2nd ed.

Glele Kakaï, R. L. and Palm, R. (2009), "Empirical comparison of error-rate estimators in logistic discriminant analysis," *Journal of Statistical Computation and Simulation*, 79, 111–120.

Golub, G. H. and van Loan, C. F. (1996), *Matrix Computations*, The Johns Hopkins University Press, 3rd ed.

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

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

Hand, D. J. (1997), *Construction and Assessment of Classification Rules*, Chichester, West Sussex, England: Wiley Series in Probability and Statistics.

Handl, J., Knowles, J., and Kell, D. B. (2005), "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, 21, 3201–3212.

Hartigan, J. A. and Wong, M. A. (1979), "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28, 100–108.

Harville, D. A. (2008), *Matrix Algebra From a Statistician's Perspective*, New York: Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2008), *The Elements of Statistical Learning*, Data Mining, Inference, and Prediction, New York, NY: Springer New York, 2nd ed.

Hennig, C. (2007), "Cluster-wise assessment of cluster stability," *Computational Statistics and Data Analysis*, 52, 258–271.

— (2008), "Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods," *Journal of Multivariate Analysis*, 99, 1154–1176.

— (2010), *fpc: Flexible Procedures for Clustering*.

Izenman, A. J. (2008), *Modern Multivariate Statistical Techniques*, Springer Texts in Statistics, New York: Springer.

Jaccard, P. (1912), "The distribution of the flora in the alpine zone." *New Phytologist*, 11, 37–50.

Jackson, D., Somers, K., and Harvey, H. (1989), "Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence?" *American Naturalist*, 436–453.

Jain, A. (2010), "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, 31, 651–666.

Jain, A. K., Dubes, R. C., and Chen, C.-C. (1987), "Bootstrap Techniques for Error Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 628–633.

Jakobsson, M. and Rosenberg, N. A. (2007), "CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure." *Bioinformatics*, 23, 1801–1806.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001), "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nature Medicine*, 7, 673–679.

Lütkepohl, H. (1996), *Handbook of Matrices*, Chichester, West Sussex, England: John Wiley and Sons Ltd.

Marco, V. R., Young, D. M., and Turner, D. W. (1987), "The Euclidean distance classifier: an alternative to the linear discriminant function," *Communications in Statistics - Simulation and Computation*, 16, 485–505.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, New York: John Wiley and Sons Inc.

McLachlan, G. J., Do, K.-A., and Ambroise, C. (2004), *Analyzing Microarray Gene Expression Data*, Hoboken, N.J.: Wiley-Interscience.

Mkhadri, A. (1995), "Shrinkage parameter for the modified linear discriminant analysis," *Pattern Recognition Letters*, 16, 267–275.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005), "Prediction error estimation: a comparison of resampling methods." *Bioinformatics*, 21, 3301–3307.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003), "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, 52, 91–118.

Pang, H., Tong, T., and Zhao, H. (2009), "Shrinkage-based Diagonal Discriminant Analysis and Its Applications in High-Dimensional Data," *Biometrics*, 65, 1021–1029.

Ramey, J. A. and Young, P. D. (2011), "A comparison of regularization methods applied to the linear discriminant function with high-dimensional microarray data," *Journal of Statistical Computation and Simulation*, 1–16.

Richardson, S. and Green, P. J. (1997), "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society. Series B (Methodological)*, 59, 731–792.

Schiavo, R. A. and Hand, D. J. (2000), "Ten More Years of Error Rate Research," *International Statistical Review*, 68, 295–310.

Seber, G. A. F. (2004), *Multivariate Observations*, Wiley Series in Probability and Statistics, Wiley-Interscience.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002), "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, 1, 203–209.

Souloumiac, A. (2009), "Nonorthogonal joint diagonalization by combining Givens and hyperbolic rotations," *IEEE Transactions on Signal Processing*, 57, 2222–2231.

Stephens, M. (2000), "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 795–809.

Tibshirani, R. and Walther, G. (2005), "Cluster validation by prediction strength," *Journal of Computational and Graphical Statistics*, 14, 511–528.

Tibshirani, R., Walther, G., and Hastie, T. (2001), "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 411–423.

Tong, T., Chen, L., and Zhao, H. (2012), "Improved mean estimation and its application to diagonal discriminant analysis," *Bioinformatics*, 28, 531–537.

Tong, T. and Wang, Y. (2007), "Optimal shrinkage estimation of variances with applications to microarray data analysis," *Journal of the American Statistical Association*, 102, 113–122.

Toussaint, G. T. (1974), "Bibliography on estimation of misclassification," *IEEE Transactions on Information Theory*, IT-20, 472–479.

Wehberg, S. and Schumacher, M. (2004), "A comparison of nonparametric error rate estimation methods in classification problems," *Biometrical Journal*, 46, 35–47.

Wickham, H. (2009), *ggplot2*, Elegant Graphics for Data Analysis, New York: Springer, 2nd ed.

Xu, P., Brock, G. N., and Parrish, R. S. (2009), "Modified linear discriminant analysis approaches for classification of high-dimensional microarray data," *Computational Statistics and Data Analysis*, 53, 1674–1687.

Yao, W. (2012), "Model based labeling for mixture models," *Statistics and Computing*, 22, 337–347.

Ye, J. and Wang, T. (2006), "Regularized Discriminant Analysis for High Dimensional, Low Sample Size Data," in *The 12th ACM SIGKDD International Conference*, New York, New York, USA: ACM Press, p. 454.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.-H., Evans, W. E., Naeve, C., Wong, L., and Downing, J. R. (2002), "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling." *Cancer Cell*, 1, 133–143.

Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. (2001), "Validating clustering for gene expression data," *Bioinformatics*, 17, 309–318.

Young, D. M., Marco, V. R., and Odell, P. L. (1987), "Quadratic discrimination: some results on optimal low-dimensional representation," *Journal of Statistical Planning and Inference*, 17, 307–319.

Zhou, X. and Mao, K. Z. (2006), "The ties problem resulting from counting-based error estimators and its impact on gene selection algorithms." *Bioinformatics*, 22, 2507–2515.