THE BOOTSTRAP IN SUPERVISED LEARNING AND ITS APPLICATIONS IN GENOMICS/PROTEOMICS

A Dissertation

by

THANG VU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2011

Major Subject: Electrical Engineering

THE BOOTSTRAP IN SUPERVISED LEARNING AND ITS APPLICATIONS IN GENOMICS/PROTEOMICS

A Dissertation

by

THANG VU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Ulisses M. Braga-Neto
Committee Members,	Edward R. Dougherty
	Aniruddha Datta
	Alan R. Dabney
Head of Department,	Costas Georghiades

May 2011

Major Subject: Electrical Engineering

ABSTRACT

The Bootstrap in Supervised Learning and
its Applications in Genomics/Proteomics. (May 2011)

Thang Vu, B.S., Hanoi University of Technology;
M.S.E, The University of Michigan, Ann Arbor
Chair of Advisory Committee: Dr. Ulisses M. Braga-Neto

The small-sample size issue is a prevalent problem in Genomics and Proteomics today. Bootstrap, a resampling method which aims at increasing the efficiency of data usage, is considered to be an effort to overcome the problem of limited sample size. This dissertation studies the application of bootstrap to two problems of supervised learning with small sample data: estimation of the misclassification error of Gaussian discriminant analysis, and the bagging ensemble classification method.

Estimating the misclassification error of discriminant analysis is a classical problem in pattern recognition and has many important applications in biomedical research. Bootstrap error estimation has been shown empirically to be one of the best estimation methods in terms of root mean squared error. In the first part of this work, we conduct a detailed analytical study of bootstrap error estimation for the Linear Discriminant Analysis (LDA) classification rule under Gaussian populations. We derive the exact formulas of the first and the second moment of the zero bootstrap and the convex bootstrap estimators, as well as their cross moments with the resubstitution estimator and the true error. Based on these results, we obtain the exact formulas of the bias, the variance, and the root mean squared error of the deviation from the true error of these bootstrap estimators. This includes the moments of the popular .632 bootstrap estimator. Moreover, we obtain the optimal weight for unbiased and minimum-RMS convex bootstrap estimators. In the univariate case, all the expressions involve Gaussian distributions, whereas in the multivariate case, the results are written in terms of bivariate doubly non-central F distributions.

In the second part of this work, we conduct an extensive empirical investigation of bagging, which is an application of bootstrap to ensemble classification. We investigate the performance of bagging in the classification of small-sample gene-expression data and protein-abundance mass spectrometry data, as well as the accuracy of small-sample error estimation with this ensemble classification rule. We observed that, under t-test and RELIEF filter-based feature selection, bagging generally does a good job of improving the performance of unstable, overfitting classifiers, such as CART decision trees and neural networks, but that improvement was not sufficient to beat the performance of single stable, non-overfitting classifiers, such as diagonal and plain linear discriminant analysis, or 3-nearest neighbors. Furthermore, the ensemble method did not improve the performance of these stable classifiers significantly. We give an explicit definition of the out-of-bag estimator that is intended to remove estimator bias, by formulating carefully how the error count is normalized, and investigate the performance of error estimation for bagging of common classification rules, including LDA, 3NN, and CART, applied on both synthetic and real patient data, corresponding to the use of common error estimators such as resubstitution, leave-one-out, cross-validation, basic bootstrap, bootstrap 632, bootstrap 632 plus, bolstering, semi-bolstering, in addition to the out-of-bag estimator. The results from the numerical experiments indicated that the performance of the out-of-bag estimator is very similar to that of leave-one-out; in particular, the out-of-bag estimator is slightly pessimistically biased. The performance of the other estimators is consistent with their performance with the corresponding single classifiers, as reported in other studies. The results of this work are expected to provide helpful guidance to practitioners who are interested in applying the bootstrap in supervised learning applications.

To my family

ACKNOWLEDGMENTS

I have been supported by many people on the journey to the Ph.D degree. First, I thank my family for their love, support, and everything.

I thank my advisor, Dr. Ulisses Braga-Neto, for his direct guidance, generous financial support, and friendly discussions that we have had during the last few years. I thank Prof. Edward Dougherty for ”opening the door” into this challenging, yet exciting field for me when he first accepted me into the Genomics Signal Processing lab, as well as his support in the critical early stage of my PhD program at Texas A&M University. I thank Prof. Aniruddha Datta and Dr. Alan Dabney for their advice, help, and valuable time when serving as members of my PhD. committee.

I take this chance to thank my teachers at Hanoi-Amsterdam High School, in particular thay Tran Van Khai, thay Dang Tran Hung, and thay Nguyen Ham, for inspiring me to pursue this long journey of getting the doctoral degree. I thank Prof. Van Dinh De at Hanoi University of Technology for his help and encouragement for me to go to graduate school. I extend special thanks to Prof. Demosthenis Teneketzis and Dr. Petar Momcilovic, at the University of Michigan at Ann Arbor, for my first lectures of graduate school and for their kindness and support. I thank my Vietnamese friends in Ann Arbor and College Station for their friendships and the carefree fun moments, which helped me stay sane during difficult times. Last but not least, I thank the Vietnam Education Foundation for the fellowship for graduate study during my first two years in the US.

TABLE OF CONTENTS

CHAPTER Page

I INTRODUCTION 1

A. Introduction 1
B. Supervised Learning 2
C. Genomics and Proteomics 3
D. Case Examples of Applications 5
E. Small-sample Challenges and Resampling Technique 7
F. Contributions 10
G. Dissertation Outline 12

II REVIEW ON ERROR ESTIMATION 13

A. Preliminaries on Supervised Learning 13
B. Linear Discriminant Analysis 16
C. Classical Error Estimation Methods 18
1. Resubstitution Estimation 19
2. Cross-validation Estimation 20
3. Kernel-based Estimation 21
4. Specific Estimation for Linear Classifiers 21
D. Bootstrap Estimation Methods 22
1. Bootstrap in Classical Statistics 22
2. Bootstrap Error Estimation 25
3. Empirical Bootstrap Error Estimation 27
4. Complete Bootstrap Estimation in Small-sample Settings . 30
E. Applications of Misclassification Error Estimation 31

III BOOTSTRAP ERROR ESTIMATION - UNIVARIATE MODEL . . 33

A. The Bias, Variance, and RMS of Estimation Deviation 34
B. The Bootstrapped Linear Discriminant Analysis 34
1. First Moment 36
2. Second Moment 37
3. Cross Correlation 38
4. Cross Moment with Resubstitution Estimator 38
5. Cross Moment with True Error 39
6. The True Error 40

CHAPTER Page

7. The Zero Bootstrap Estimation 42
8. The Convex Bootstrap Estimation 42
C. Univariate Model 43
D. The C-bootstrap Linear Classifier 53
E. The Zero Bootstrap Error Estimation 57
F. The Convex Bootstrap Error Estimation 59
1. The Moments of the True Error 60
a. The First Moment 60
b. The Second Moment 60
2. The Moments of the Resubstitution Estimator 61
a. The First Moment 61
b. The Second Moment 62
c. The Correlation with the True Error 63
3. The Moments of the Convex Estimator 64
G. The .632 Bootstrap Error Estimation 66
H. The Optimal Bootstrap Error Estimation 66
I. The Unbiased Bootstrap Error Estimation 67

IV BOOTSTRAP ERROR ESTIMATION - MULTIVARIATE MODEL 70

A. Multivariate Model 71
1. Bootstrapped Linear Discriminant 71
2. Some Definitions 72
B. The C-bootstrap Linear Classifier 73
C. The Zero Bootstrap Error Estimation 80
D. The Convex Bootstrap Error Estimation 82
1. The Moments of the True Error 83
a. The First Moment 83
b. The Second Moment 83
2. The Moments of the Resubstitution Estimator 84
a. The First Moment 84
b. The Second Moment 85
c. The Cross Moment 87
3. The Moments of the Convex Estimator 89
E. The .632 Bootstrap Error Estimation 91
F. The Optimal Bootstrap Error Estimation 91
G. The Unbiased Bootstrap Error Estimation 92

V SMALL-SAMPLE PERFORMANCE OF BAGGING CLASSIFICATION RULES 95

CHAPTER Page

A. Introduction 96
B. Randomized Ensemble Classification Rules 97
C. Experimental Study 99
1. Methods 99
2. Data Sets 101
3. Results and Discussion 102
D. Conclusion 111

VI SMALL-SAMPLE ERROR ESTIMATION FOR BAGGING CLASSIFICATION RULES 112

A. Introduction 113
B. Error Estimation for Bagging Classification Rule 114
1. Classical Methods 114
2. Bolstered Error Estimation 115
3. Out-of-bag Error Estimation 115
C. Simulation Study 117
1. Methods 117
2. Simulation Based on Synthetic Data 118
3. Simulation Based on Patient Data 118
4. Results and Discussion 118
a. Synthetic Data 118
b. Patient Data 120
D. Conclusion 130

VII CONCLUSION 131

REFERENCES 132

APPENDIX A 154

APPENDIX B 163

VITA 181

LIST OF TABLES

TABLE Page

I Expected classification error of selected experiments with breast cancer gene expression data (full results available on the companion website). Bold-face type indicates the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test. 103

II Expected classification error of selected experiments with the lung cancer gene expression data (full results available on the companion website). Bold-face type indicates the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test. 104

III Expected classification error of selected experiments with prostate cancer protein abundance data (full results available on the companion website). Bold-face type indicates the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test. 105

IV Bias, variance (standard deviation), and RMS for different error estimators, with different base classification rules, for breast cancer gene expression data, dimensionality $p = 2$. 120

V Bias, variance (standard deviation), and RMS for different error estimators, with different base classification rules, for lung cancer gene expression data, dimensionality $p = 2$. 128

VI Bias, variance (standard deviation), and RMS for different error estimators, with different base classification rules, for prostate cancer mass-spectrometry data, dimensionality $p = 2$. 129

LIST OF FIGURES

FIGURE Page

1 Optimal weight w_u in the univariate case. The top figure displays w_u as a function of Bayes error for different sample sizes, whereas the bottom figure displays w_u as a function of the number of samples for different Bayes errors. 69

2 Optimal weight w_u in the multivariate case, $p = 2$. The top figure displays w_u as a function of Bayes error for different sample sizes, whereas the bottom figure displays w_u as a function of the number of samples for different Bayes errors. 94

3 Expected classification error as a function of number of component classifiers in the ensemble for selected experiments with the breast cancer gene expression data (full results available on the companion website). Error of single component classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test. 108

4 Expected classification error as a function of number of component classifiers in the ensemble for selected experiments with the lung cancer gene expression data (full results available on the companion website). Error of single component classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test. 109

5 Expected classification error as a function of number of component classifiers in the ensemble for selected experiments with the prostate cancer protein abundance data (full results available on the companion website). Error of single component classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test. 110

FIGURE Page

6 Comparison of out-of-bag and leave-one-out for different Gaussian models over the number of samples $p = 2$ (a) Sample mean, (b) Sample standard deviation. 121

7 Bias, variance (standard deviation), and RMS of as a function of the bayes error, for the synthetic data, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules: LDA, 3NN, and CART on the first, second, and third row, respectively. 122

8 Empirical deviation distribution (top row), box plots (middle row), and RMS as a function of sample size (bottom row), for synthetic Gaussian model with Bayes error = 0.05, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules. 123

9 Empirical deviation distribution (top row), box plots (middle row), and RMS as a function of sample size (bottom row), for synthetic Gaussian model with Bayes error = 0.15, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules. 124

10 Empirical deviation distribution (top row) and box plots (bottom row), for breast cancer gene-expression data, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules. 125

11 Empirical deviation distribution (top row) and box plots (bottom row), for lung cancer gene-expression data, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules. 126

12 Empirical deviation distribution (top row) and box plots (bottom row), for prostate cancer mass-spectrometry data, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules. 127

CHAPTER I

INTRODUCTION

This chapter provides a broad overview of the interface between the biomedical research and the quantitative methods, which is now generally known as Computational Biology, Systems Biology or Genomics Signal Processing. It also touches on the small sample problem, one of the major obstacles of the field. Despite of still being in the primary stage, Computational Biology has been showing very potential applications, some of which will also be highlighted in this chapter. Finally, some contributions of this dissertation are introduced and its organization is outlined.

A. Introduction

Quantitative methods are indispensable components of biomedical research in the 21st century. In the report ”Catalyzing Inquiry at the Interface of Computing and Biology,” by the National Institute of Health [1], Systems approach and the power of computation and engineering were considered as essential constituents of the life science research this century. In the same report, life science, in particular biology was characterized as ”empirical”, ”descriptive” and ”experimental”. With the recent advent of genome sequencing and high-throughput data, computation is now integrated into biological sciences as a crucial component. Computation means not only storage and visualization of but also the analysis and inference of biologically meaningful information from the data. While the former can be supported by information technology and software engineering, it is the statistical learning that fulfills the latter.

A tremendous amount of effort is invested to apply the statistical learning into biomed-

The journal model is *IEEE Transactions on Automatic Control.*

ical research and mining the high-throughput data. There are a plethora of research work on applying engineering and quantitative sciences into life science and medicine in the last ten years. It gave birth to new fields which are interdisciplinary between life sciences, physical sciences, and engineering. They are now generally known as Computational Biology, Systems Biology, Genomics Signal Processing, or Bioinformatics.

This field is still in its infancy [2]. Despite of commonplace critics among the medical research community about its reproducibility and reliability, the initial achievements of Computational Biology are promising and deserves appreciation. Some applications in cancer research will be mentioned in the next sections.

B. Supervised Learning

Supervised learning is an important quantitative method. It is one major type of statistical learning, in which a system is mathematically modeled and designed from the available information, which is usually in the form of numerical data sets. The supervised method is different from the unsupervised method in the way that, in the former we know *the label* of the data we have, instead of discovering them in the latter. More precisely, supervised learning is concerned with the problem of learning from the available information to be capable of predicting unknown information in the future. It has been also known under different names such as pattern recognition, machine learning, decision theory, pattern analysis, data mining or artificial intelligence. Besides Computational Biology, it has already demonstrated wide varieties of practical applications in diverse fields such as Image Analysis, Remote Sensing, Medical Image Diagnosis, Speech Recognition, Robotics etc for a long time [3].

Supervised learning problems can be sub-classified into three categories: *classification*, *error estimation*, and *feature selection*. The ultimate goal of a supervised learning problem is to design a reliable system which can accurately predict a future observation. In practice, the available information , i.e the number of data points is typically limited. This data set need to be used to *design* a predictor which is capable of predicting properties of future samples such as their labels *(classification)* or their values *(regression)*. Generally the more samples are used for designing, the more accurate the predictor. The prediction accuracy also depends significantly on the *classification rule* used to build it (*classification*). In order to know how correctly the classifier work, we need to evaluate by estimating its error rate. A good estimator requires, in principle, a large amount of independent information or data set (*error estimation*). Furthermore, there are a known relation between the accuracy of the classifier with the number of training samples and the number of characteristics or *features* based on which the classification is made. For a fixed number of training samples, it is not always the best to use as many features as possible. That leads to the problem of selecting the optimal subsets of features with the best discriminatory powers, which is known as the *feature selection* problem. The three problems have interacting relations. Given the restricted source of information, finding an optimal solution for these three closely interrelated problems is not an easy task.

Some more fundamental points of supervised learning are presented in Chapter II. The next section presents some observations and applications of supervised learning in biomedical research.

C. Genomics and Proteomics

The advent of biotechnology allow to measure simultaneously the activity of tens of thousands of biological entities in cellulars such as mRNA, protein, noncoding RNA, DNA methylation status of CpG sites, the numbers of copies of genes etc. For example, based on the hybridization technology, the *Human Genome U133 Plus 2.0 Array* by the Affymetrix

Corp. can measure the expression levels of about $47,400$ transcripts [4]. The *Agilent Human Genome CGH Microarray 244A* is able to quantify the genome copy number variations by performing Comparative Genome Hybridization (CGH) with about $236,381$ biological features [5]. The *Infinium HumanMethylation27 BeadChip* allows to investigate the methylation status of $27,578$ highly informative CpG sites located within the proximal promoter regions of transcription start sites of $14,475$ consensus coding sequencing (CCDS) in the NCBI Database (Genome Build 36) [6]. The Liquid-Chromatography Mass Spectrometry (LC-MS) and tandem Mass Spectrometry (LC-MS/MS) technology enable the quantitative assessment of protein expression level through the relation between the mass over charge ratio and the time of flight of the enzyme-digested peptides of the proteins.

These technologies obviously generate an enormous amount of data about the activity of the cellular biological systems. Even though there are some current technical issues of these technologies namely noise, image analysis, experimental design of microarray chips etc and the low resolution, low accuracy of the proteomics instruments etc, these high-throughput datasets can be considered as precious sources of information of the underlying cellular biological processes, which was generally unavailable to the life science research before. In order to *mine* the biological knowledge hidden in these numbers, the quantitative methods need to be applied, and more specifically statistical inference are regarded as a nature choice. One of the problems that the statistical inference of these data sets is presently facing is the small number of samples in comparison to the number of features. As mentioned before, most of the chips measure tens of thousand of biological features while the number of biological replicates, i.e the number of tissues are often as limited as hundreds in most of current biomedical research. The classical statistical inference has been established for the context of having large numbers of samples. In the settings of limited samples, these traditional inference methods work unsurprisingly differently. More small-sample studies need to be done to ensure the reliable and accurate outputs of these statistical inference algorithms in the contexts of Genomics and Proteomics.

D. Case Examples of Applications

The growth of advanced high-throughput technology and the sequencing technology of human genomes, together with some other factors, is a milestone of the new revolutionary period of medical science [7]. There have been a considerable amount of research efforts in applying the supervised learning using these genomics and proteomics data sets in solving biomedical research problems. The range of Computational Biology is very large, including the intersections of engineering, statistics and quantitative sciences with life sciences and medicine. As a result, Computational Biology has been appearing in a wide variety of biomedical research topics using different classes of quantitative research, which are applied for all kinds of high-throughput and sequencing data. Due to the constraint of this dissertation, we focus on applications of supervised and unsupervised learning in Genomics and Proteomics with the emphasis on the supervised method, which is more relevant to this dissertation. Following is a very brief highlight of Genomics-based and Proteomics-based applications in biomedical research generally, and in cancer research particularly.

Genomics and Proteomics have been integrated into studies of different cancers. Readers can obtain details about the genomics-based literature for each cancer type in many comprehensive reviews, namely for breast cancer in [8, 9, 10, 11, 12], for lung cancer in [13], for acute myeloid leukemia in [14], for melanoma in [15], for epithelial ovarian cancer in [16], for colorectal cancer in [17] etc.

For each type of cancer, statistical inference of high throughput data sets has been used to study some problems of oncology research. The applications of statistical learning can be mostly classified into three main categories: class discovery, class prediction, and class comparison. Besides these, it has also been employed in survival prediction, clinical trial design and biomarker discovery and validation.

First, class discovery has been playing roles in studing the biological mechanism of cancer to get more insights into oncogenesis [18, 19]. It is also used to either discover new cancer class or classify tumors to known classes, for which there have been no general approach [20]. It is now well known that tumors with similar phenotypes can be genetically very different. Understanding the pathogenesis of cancer subtypes is very important, because cancer in different subtypes can develop from different causes or cells of origin. As a result, a more suitable therapeutic approach for each specific subtype need to be used to provide better drug efficacy. For example in [21], Verhaak et al, using statistical analysis mostly based on hierarchical clustering - one of the most common unsupervised methods, identified clinically relevant subtypes of Glioblastoma Multiforme. They also found that Glioblastoma subtypes are reminiscent of distinct neural cell types and show different treatment efficacy. Another example is the works by Bhattacharjee et al in [22], in which distinct adenocarcinoma subclasses were revealed by mRNA expression profiling. Similar attempts in identifying molecularly cancer subtypes can be found in [23, 24, 25].

Second, genomics- and proteomics-based studies have been used to find significantly differently expressed genes or proteins, which are usually known as studies of class prediction and comparison. These differentially expressed genes or proteins can be considered as molecular biomarkers serving a wide varieties of applications namely diagnosis, prognosis, staging or selecting optimal personal therapy [26]. They can be used for early cancer detection. Also, they can be clinically relevant therapeutic biomarkers, based on which a better treatment plan is applied with expectedly better efficacy [27]. These tasks are to be archived by using supervised learning methods together with well designed clinical trials.

Moreover, the microarray-based clinical trial research has been emerging as a new and active area [28].They have been also used in studying of survival prediction [29, 30]. The identification of new targets and new drug in drug discover, application of individualized medicine based on pharmagenomic biomarkers is significantly assisted by the statistical inference [31].

Even though most of the findings of these research have not yet become part of the medical practice today [32, 33, 34, 35, 9, 13, 36, 27, 37, 38], and obviously more works needs to be done to realize them, they are unprecedented and deserve appreciation as the initial step in directing the biomedical research to a new direction. The main drawbacks of using microarray-based studies are reproducibility and validity [33]. The reasons for these two problems, besides technical issues, are related to the small-sample size and the computational models used. The basic points of the small-sample size problem is addressed in the next section.

E. Small-sample Challenges and Resampling Technique

While the challenge of small-sample sizes was long time ago raised in the research literature of statistical pattern recognition regarding the relation between the sample sizes and the optimal subsets of features used to design classification systems and the effect of that relation on the system performance in [39, 40], it becomes particularly prevalent today in the application of genomics and proteomics [41, 42, 43].

As a highly application-oriented field, statistical discriminant analysis received considerable attention on its issues regarding practical design and implementation [44, 39, 45, 46, 47]. The basic question was, for a fixed number of samples, what was the optimal subset of features that gives the best classifiers. The topic has been long known as the *peaking phenomenon* or the *curse of dimensionality*. Given a fixed limited dataset, designing a classification system should be conducted as a process involved all the three closely interrelated stages: feature selection, which involves picking the best subset of features to design the classifiers on; the classifier design, which is concerned with formulating a predictor; and very importantly error estimation which determines how accurate the designed classifier can be. As a general principle, the more data we have, the more accurate each stage. In many practical applications, all these three stages must be implemented using one dataset of limited samples.

So, the first difficulty of small-sample problems is naturally concerned with the design and validation of the system. Small training set makes classifiers unstable and variable [47]. Data is limited and has to be split to first design the system and then evaluate its performance, not to mention the process of feature selection. It is a trade-off because the fewer samples are used to design, generally the less accurate the classifier; the fewer samples are left out to test the classifier, the more unreliable the estimators are.

The problem of sample sizes is remarkably important to the practitioners who want to design a reliable and accurate system in practice. In principle, the small sample size can easily contaminate the design and evaluation of the systems. Because first, data sets of few samples fail to statistically represent the underlying distributions. Consequently, the classifiers designed on this sparse data often perform poorly when validated on the independent future observations. This fact can be explained clearly for parametric classification rules where small-sample estimates for parameters of the label-feature distributions are far from reliable and accurate. As a consequence, the parametrically designed classifiers are unstable and inaccurate [48]. The linear classifier with unknown covariances under Gaussianity assumption is a clear example, when the covariance matrix is to be estimated by the pooled sample covariance matrix. These matrices are even singular when the numbers of samples are smaller than the number of features; the classification design fails consequently.

Small samples results in severe model selection bias [49] and overfitting. Overfitting generally means while the classifiers perform very well on the training data, or even on the hold-out test data, which gives the apparent error or the hold-out error almost zero, they show disappointing performance on the validation samples. This behavior typically happens for classification rules of complexed structure, which normally require a large amount of data to work well [48]. Apparently, these overfitting classification rules fail to work well in the small-sample settings.

This small-sample problem is particularly prevalent in genomics and proteomics today [27, 37, 33, 42, 50, 49, 51, 52, 53]. While high-throughput biotechnology chips can be regarded as a breakthrough in the life sciences allowing activity measurement of tens of thousands of cellular entities at the same times, it also poses a challenge for those who want to statistically mine them by offering only a small number of replicates due to subjective constraints such as the tissue sources, time, and, cost etc. One reason which hinder the applications of molecular biomarker in cancer research, found by genomics-based and proteomics-based classifiers, into clinical practice is the lack of valid validation [35, 14, 27, 54, 13, 36, 32, 8, 33, 17, 34, 55, 56, 16, 9, 57, 58, 53, 59].

Cross-validation has been used an effort for validation [37]. It is good giving an almost unbiased estimate. Problem with cross-validation is its high variance, in particular for estimating the misclassification error of expression-based classifiers of small sample sizes. While giving an almost unbiased estimate, the wide variability of cross-validation can ruin its reliability.

Bootstrap can be regarded as a smooth version of cross-validation. It performs better than cross-validation by giving smaller variance; and so ultimately the superior performance in term of root-mean-squared error to most of other error estimation methods. Moreover, bootstrap resampling increases the efficiency of data usage when it can reuse the samples through the process of uniform resampling with replacement. That property of bootstrap has been applied in designing more accurate classifiers. The ensemble classification rules are considered as typical application of that idea. The ensemble classifiers combine the classification decisions of an ensemble of individual classifiers, which are designed either on bootstrap samples (bootstrap aggregation or bagging) or on different subspace of features (random subspace method). So basically data are reused in designing member classifiers of the ensemble; and the member classifers are clearly correlated. Empirical studies have generally shown better performance for ensemble classifier, in particular when the individual classifiers are diverse and weakly correlated with each other.

The major drawback of the resampling method is computation time when it needs a larger number of iterations, in comparison with cross-validation. This was a problem for the effort to implement bootstrap about twenty years ago. Nowadays with the revolutionary growth of information technology with strong personal computers and supercomputers, this is no longer a big problem.

As a conclusion, resampling is one approach to beat the problem of limited samples. This dissertation studies the applications of this method for the first two problems of supervised learning; bootstrap error estimation and ensemble classification rule.

F. Contributions

In the first part of this work, we conduct a detailed analytical study of bootstrap error estimation for the Linear Discriminant Analysis (LDA) classification rule under Gaussian populations. We derive the exact formulas of the first and the second moment of the zero bootstrap and the convex bootstrap estimators, as well as their cross moments with the resubstitution estimator and the true error. Based on these results, we obtain the exact formulas of the bias, the variance, and the root mean squared error of the deviation from the true error of these bootstrap estimators. This includes the moments of the popular .632 bootstrap estimator. Moreover, we obtain the optimal weight for unbiased and minimum-RMS convex bootstrap estimators. In the univariate case, all the expressions involve Gaussian distributions, whereas in the multivariate case, the results are written in terms of bivariate doubly non-central F distributions.

In the second part of this work, we conduct an extensive empirical investigation of bagging, which is an application of bootstrap to ensemble classification. We investigate the performance of bagging in the classification of small-sample gene-expression data and protein-abundance mass spectrometry data, as well as the accuracy of small-sample error estimation with this ensemble classification rule. We observed that, under t-test and RELIEF filter-based feature selection, bagging generally does a good job of improving the performance of unstable, overfitting classifiers, such as CART decision trees and neural networks, but that improvement was not sufficient to beat the performance of single stable, non-overfitting classifiers, such as diagonal and plain linear discriminant analysis, or 3-nearest neighbors. Furthermore, the ensemble method did not improve the performance of these stable classifiers significantly. We give an explicit definition of the out-of-bag estimator that is intended to remove estimator bias, by formulating carefully how the error count is normalized, and investigate the performance of error estimation for bagging of common classification rules, including LDA, 3NN, and CART, applied on both synthetic and real patient data, corresponding to the use of common error estimators such as resubstitution, leave-one-out, cross-validation, basic bootstrap, bootstrap 632, bootstrap 632 plus, bolstering, semi-bolstering, in addition to the out-of-bag estimator. The results from the numerical experiments indicated that the performance of the out-of-bag estimator is very similar to that of leave-one-out; in particular, the out-of-bag estimator is slightly pessimistically biased. The performance of the other estimators is consistent with their performance with the corresponding single classifiers, as reported in other studies. The results of this work are expected to provide helpful guidance to practitioners who are interested in applying the bootstrap in supervised learning applications.

G. Dissertation Outline

Concerning the coverage of individual chapters, Chapter I introduces briefly the supervised learning as well as its applications in biomedical research, in particularly Genomics and Proteomics. It also presents the main points of the small-sample problems and describes the resampling method, which is considered as an approach to resolve the problem of the limited samples. The first part of this dissertation is about the theoretical analysis of bootstrap error estimation for the linear classification rule. First, Chapter II provides the preliminaries on supervised learning and a review on error estimation, with emphasis on the bootstrap methods. Chapter III presents the theoretical analysis of some variants of bootstrap estimations for linear discriminant analysis under univariate Gaussian model, while Chapter IV provides the results for the multivariate Gaussian model. The second part of this dissertation begins with Chapter V, which reports the performance of a varieties of bagging classifiers in small-sample settings applied for some Genomics and Proteomics datasets. Chapter VI provides the results of an extensive empirical study on estimating errors of bagging classifiers. The last chapter, Chapter VII, presents some concluding remarks.

CHAPTER II

REVIEW ON ERROR ESTIMATION

This chapter first provides the preliminaries on supervised learning and the basic notations which are used throughout the dissertation. Then a review on error estimation problem is given with the emphasis on the bootstrap methods. Finally, we highlight the importance of error estimation via some applications in computational biology.

A. Preliminaries on Supervised Learning

There are excellent references on supervised learning. Here, we present the main points of supervised learning, which acts more as the introduction of the notation we will use, other than a review of the subject. Thorough material of the subject can be found in the works by Duda, Hart, and Stork [60], Devroye, Györfi, and Lugosi [61], Fukunaga [62], Mclachlan [63], Jain, Duin, and Mao [64], the panel on Discriminant Analysis, Classification, and Clustering of the committee on Applied and Theoretical Statistics of the Board on Mathematical Sciences, National Research Council [65] or elsewhere.

Almost all the supervised problems can be modeled mathematically as following. Suppose we need to classify an object, named X into one kind out of C categories. For simplicity, we assume there are only two categories: Π_0 and Π_1 ($C = 2$). For $C > 2$, all the concepts apply with slight modifications for which readers are referred to the previously mentioned references. The classification problem with two classes are often referred to as *binary* or *dichotomous*. The class of an object is also called its *label*. Let use Y to denote the label of object X. For binary classification, object X can either have *label* $Y = 0$ ($X \in \Pi_0$) or *label* $Y = 1$ ($X \in \Pi_1$). The object X is classified based on its *characteristics* or *features*. Features can be either numerical or categorical, or converted to either of these. Different characteristics makes up different *features*. The number of characteristics,

p, available about the object X is called *number of features* and the set of p features create the *feature space* of X.

$$\mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2 \times \cdots \times \mathscr{X}_p \tag{2.1}$$

From now on, we use X to denote the object and the its features interchangably. For supervised learning, we usually have a set of data with known labels for both classes. Let call the data set $S_n = S_0 \cup S_1$ where $S_0 = \{(X_1,Y_1), (X_2,Y_2), \cdots, (X_{n_0},Y_{n_0})\}$ and $S_1 = \{(X_{n_0+1},Y_{n_0+1}), \cdots, (X_{n_0+n_1},Y_{n_0+n_1})\}$, where $X_i \in \Pi_0$ or $Y_i = 0$, for all $1 \leq i \leq n_0$ and $X_i \in \Pi_1$ or $Y_i = 1$, for all $n_0 + 1 \leq i \leq n_0 + n_1$. The data collected often contain random noise. The relationship between the features and their labels, therefore, is statistically random. It is modeled as the joint distribution between label and features, often referred to simply as the *feature label distribution* or the *conditional distribution* or *underlying distribution* $F_{X,Y}(.)$. This distribution is often unknown. It is subjectively ideal to find a deterministic function $f(.)$ such that

$$\begin{aligned} f : \mathscr{X} &\longrightarrow \{0,1\}, \\ Y &= f(X). \end{aligned}$$

Given the random nature of the relationship between label and features mentioned above, it is generally impossible to find such a deterministic $f(.)$. The classification instead is implemented via the *discriminant function* $W(S_n,X)$ or the *classifier* $\psi(X)$ found by applying the *classification rule* Ψ on the *training sample* S_n.

$$\psi(X) = \begin{cases} 0 & \text{if } W(S_n,X) \geq c \\ 1 & \text{otherwise} \end{cases} \tag{2.2}$$

where c is a threshold found when designing the classifier.

There are varieties of classification rules, which can be classified into different types based on different classification criterion. They can be sample-based v.s. optimization-based. They are parametric v.s. non parametric. They are stable v.s weak. They can be individual or ensemble. More details can be found in the review paper by Jain [64].

One of the ultimate goals of designing the *classifier* $\psi(X)$ is to be able to accurately predict the unknown label of new observations X. The probability of incorrectly classifying X is

$$\varepsilon = P\{Y \neq \psi(X)\}\,. \tag{2.3}$$

More precisely,

$$\varepsilon = P\{\psi(X) = 1\,|\,Y = 0\}P\{Y = 0\} + P\{\psi(X) = 0\,|\,Y = 1\}P\{Y = 1\}\,, \tag{2.4}$$

or

$$\varepsilon = (1-\gamma)\varepsilon^0 + \gamma\varepsilon^1 \tag{2.5}$$

where $\varepsilon^0 = P\{\psi(X) = 1\,|\,Y = 0\}\,, \varepsilon^1 = P\{\psi(X) = 0\,|\,Y = 1\}$ and $\gamma = P\{Y = 1\}$ is the *class priori probability.*

The *Bayes rule* is the classification rule Ψ^* which produces the *Bayes classifier* $\psi^*(X)$ with the minimum misclassification error ε^*.

$$\varepsilon^* = \underbrace{min}_{\Psi}\varepsilon \tag{2.6}$$

In the literature of error estimation in classification, ε is often referred to as the *conditional error* to denote the given condition of training sample S_n, i.e the classifier ψ is still a function of S_n. It is often of interest to investigate ε over the distribution of S_n, i.e consider the expectation of ε (*conditional error*) and its other moments over the distribution of S_n. In this dissertation, ε is used to denote the *conditional error* and $E_{S_n}[\varepsilon] = \mathrm{E}[\varepsilon]$ for *unconditional error.*

B. Linear Discriminant Analysis

Because the major part of this work is concerned with the bootstrap error estimation method for linear classifiers under Gaussianity assumptions, details about this classification rule under this standard condition is provided in the following.

Linear Discriminant Analysis (LDA) employs Anderson's W discriminant [66], which is defined as follows:

$$W(X) = \left(X - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right)^T \Sigma^{-1} (\hat{\mu}_0 - \hat{\mu}_1) \tag{2.7}$$

where

$$\begin{aligned} \hat{\mu}_0 &= \frac{1}{n_0} \sum_{i=1}^{n_0} X_i, \\ \hat{\mu}_1 &= \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} X_i \end{aligned} \tag{2.8}$$

are the sample means of the sample sets S_0 and S_1, respectively. This defines the LDA classification rule, whereby the designed LDA classifier is defined by:

$$\psi(X) = \begin{cases} 1, & \text{if } W(X) < 0 \\ 0, & \text{if } W(X) \geq 0 \end{cases}, \tag{2.9}$$

that is, the sign of $W(X)$ determines the classification of X. Here we are assuming that the covariance matrix Σ is known.

For the case with the assumption of unknown covariance matrices, they are estimated by the pooled sample covariance matrix

$$S = \frac{1}{n_0 + n_1} \left(\sum_{i=1}^{n_0} (X_i - \hat{\mu}_0)(X_i - \hat{\mu}_0)^T + \sum_{i=n_0+1}^{n_0+n_1} (X_i - \hat{\mu}_1)(X_i - \hat{\mu}_1)^T \right). \tag{2.10}$$

The W becomes

$$W(X) = \left(X - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right)^T S^{-1} (\hat{\mu}_0 - \hat{\mu}_1). \tag{2.11}$$

In the univariate case, both (4.1) and (2.11) reduce to

$$\psi(X) = \begin{cases} 1, & \text{if } \left(X - \frac{\hat{\mu}_0 - \hat{\mu}_1}{2}\right)(\hat{\mu}_0 - \hat{\mu}_1) < 0 \\ 0, & \text{otherwise} \end{cases} . \tag{2.12}$$

Under the standard assumption of Gaussinity, i.e $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0 + 1, \ldots, n_0 + n_1$, the conditional true error ε has the closed form as follows: with assumption of known covariance,

$$\varepsilon = (1-\gamma)\Phi\left(-\frac{1}{2}\sqrt{(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1)}\right) + \gamma\Phi\left(\frac{1}{2}\sqrt{(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1)}\right); \tag{2.13}$$

and with assumption of unknown covariance,

$$\varepsilon = (1-\gamma)\Phi\left(\frac{-(\mu_0-\mu_1)^T S^{-1}(\mu_0-\mu_1)}{2\sqrt{(\mu_0-\mu_1)^T S^{-1}\Sigma S^{-1}(\mu_0-\mu_1)}}\right) + \gamma\Phi\left(\frac{(\mu_0-\mu_1)^T S^{-1}(\mu_0-\mu_1)}{2\sqrt{(\mu_0-\mu_1)^T S^{-1}\Sigma S^{-1}(\mu_0-\mu_1)}}\right). \tag{2.14}$$

Moreover, the unconditional error in the case of known covariance matrix is

$$\mathrm{E}[\varepsilon] = (1-\gamma)P\left(\frac{W_1}{W_2} < \frac{1-\rho_0}{1+\rho_0}\right) + \gamma P\left(\frac{W_3}{W_4} > \frac{1+\rho_0}{1-\rho_0}\right), \tag{2.15}$$

where W_1, W_2, W_3, and W_4 (W_1, W_2 are independent and so are W_3, W_4) are distributed as noncentral chi-square variables with p degrees of freedom with noncentrality parameters λ_1, λ_2, λ_3, and λ_4, respectively with

$$\begin{aligned} \lambda_1 &= \lambda_4 = \frac{n_0 n_1}{2(1+\rho_0)}\left(\frac{1}{\sqrt{n_0+n_1}} - \frac{1}{\sqrt{n_0+n_1+4n_0n_1}}\right)^2 \Delta^2, \\ \lambda_2 &= \lambda_3 = \frac{n_0 n_1}{2(1-\rho_0)}\left(\frac{1}{\sqrt{n_0+n_1}} + \frac{1}{\sqrt{n_0+n_1+4n_0n_1}}\right)^2 \Delta^2, \\ \rho_0 &= \frac{n_1 - n_0}{\sqrt{(n_0+n_1)(n_0+n_1+4n_0n_1)}}, \end{aligned} \tag{2.16}$$

where $\Delta^2 = (\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)$ is the Mahalanobis distance between the populations. The unconditional error $\mathrm{E}[\varepsilon]$ in the case of unknown covariance matrix has very complexed distributional properties involving the distribution of Hotelling's T^2 distributions [67]. In [68], Sitgreaves obtained a complicated closed form for $\mathrm{E}[\varepsilon_0]$ involving five infinite summations. There was a line of work on the topic of asymptotic expansion of the moments of the conditional error. Such typical studies includes [69, 67, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79]. The part of this dissertation on bootstrap error estimation for LDA is concerned only with the case of known covariance matrix.

LDA is a simple rule but has been shown to work quite competitively in small-sample settings. LDA can be trained/designed quickly, in comparison to other such as Classification and Regression Tree (CART), or Neural Networks, etc which takes much longer time to train. It also acts as the base rule for the nonlinear classification rule to be projected to higher dimensional space. For example, the nonlinear Support Vector Machine is usually projected onto higher-dimensional space, on which it is linear. The effectiveness of LDA when there are limited sample points was affirmed in the works by Raudys [39]. Moreover, in a study comparing the performance of ensemble classifiers with their corresponding individual classifiers [80] in the context of Genomics and Proteomics, LDA was shown to be consistently one of the best in term of accuracy and training time.

C. Classical Error Estimation Methods

In practice, pattern recognition systems are often designed based on a fixed set of available data; the accuracy of designed classifiers is evaluated based on the conditional error. While the unconditional error gives us a global view of the performance of the classification rule under a certain conditions and/or assumptions i.e. the average performance over all possible training sets, it is the conditional error that is useful in practice. Therefore, it is the conditional error that needs estimating. All of the following estimation methods, unless otherwise stated, are for estimating the conditional error.

Together with inventing estimation methods of the conditional error, evaluating their performance is a critical issue, as well. This important issue was mentioned in the seminal paper of Raudys [39]. Being functions of the data, the estimates themselves are statistics with distributional properties. The estimators $\hat{\varepsilon}$s are often evaluated based on the moments of their deviation from the true error ε. The common moments of interests often are the first (the bias $\mathrm{E}[\hat{\varepsilon}-\varepsilon]$), second (the variance $\mathrm{E}[(\hat{\varepsilon}-\mathrm{E}\hat{\varepsilon})^2]$), and the cross-moment ($\mathrm{E}[\hat{\varepsilon}\varepsilon]$), which are involved in forming the RMS error $\mathrm{E}[(\hat{\varepsilon}-\varepsilon)^2]$, the usual metric used to evaluate the behavior of estimators.

There have been some excellent reviews on estimating the misclassification rates including [81, 82, 83, 84, 85]. These papers provide thorough overviews of the statistical properties of the true error and its classical estimators including resubstitution, holdout, cross-validation and kernel-based estimators as well as their performance in simulation studies. We therefore in the first part of the followings mention concisely about these with focus on the most up-to-date progress. The emphasis of this review is largely on the bootstrap methods presented in the later part.

1. Resubstitution Estimation

Data in practice are often limited, and the training sample S_n has to be used for both designing the classifier ψ_n and estimate the true error ε. An obvious method to estimate ε is thus to use S_n itself as the test set. This is called the *resubstitution* estimator:

$$\hat{\varepsilon}_r = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \Psi_n(S_n)(X_i)| = \frac{1}{n}\left[\sum_{i=1}^{n_0}\mathrm{I}_{\psi(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1}\mathrm{I}_{\psi(X_i)=0}\right] \tag{2.17}$$

This method has been well known as often, although not always, optimistic, especially in small sample settings. Zollanvari, Braga-Neto, and Dougherty provided some theoretical results of distributional properties of resubstitution estimator for linear discriminant analysis under Gaussianity assumption with known covariance matrix. The key part of their results includes the exact formula for bias, variance, and the root-mean-square of the deviation of the resubstition estimator from the true error in the univariate case; and the asymptotic exact approximation in the multivariate case. More details can be found in [86].

2. Cross-validation Estimation

In *k-fold cross-validation*, S_n is partitioned into *k folds* $S_{(i)}$, for $i = 1, \ldots, k$ (for simplicity, we assume that k divides n), each fold is left out of the design process and used as a testing set, and the estimate is the overall proportion of error committed on all folds [60]:

$$\hat{\varepsilon}_{\text{cvk}} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n/k} |Y_j^{(i)} - \Psi_n(S_n \backslash S_{(i)})(X_j^{(i)})|, \tag{2.18}$$

where $(X_j^{(i)}, Y_j^{(i)})$ is a sample in the i-th fold. The process may be repeated, where several cross-validated estimates are computed, using different partitions of the data into folds, and the results averaged. In *leave-one-out estimation*, a single observation is left out each time, which corresponds to n-fold cross-validation. The leave-one-out estimator is nearly unbiased as an estimator of $\mathrm{E}[\varepsilon]$. Zollanvari et al presented some theoretical distributional properties of this special case of cross-validation, again for linear discriminant analysis under Gaussianity assumption with known covariance matrix in [86].

On one hand, this has been one of the most widely used methods thanks to its almost unbiased property. On the other hands, it is also known for the high variance. This behavior of cross-validation has been explicitly identified in a number of extensive empirical studies. Braga-Neto and Dougherty [43] did a substantial simulation study of error estimation in the small-sample settings of Genomics application and illustrated clearly the high-variability of cross-validation estimators in that scenario.

3. Kernel-based Estimation

Most of the misclassification estimation methods are counting-based. There are some efforts to introduce the kernel-based methods, which are also known as smooth estimation. This is basically the continuity-corrected version of the regular counting methods, in order to reduce variance. These works [87, 88, 89, 90, 43] presented promising performances of kernel-based estimators, in term of RMS error. While it is arguably competitive with the best methods, the major problem of smooth estimators is to choose the best kernel with its optimal kernel bandwidth.

4. Specific Estimation for Linear Classifiers

Under the assumption of Gaussian distribution with equal covariance matrix, the error rate of LDA has the closed form as in (2.13) and (2.14). Based on these formulas, the following estimators were proposed by different authors and summarized in [63]:

- D method: The Mahananobis distance between the two classes is estimated by plugging in the sample means and sample covariance matrix.

$$D = \sqrt{(\hat{\mu}_0 - \hat{\mu}_1)^T S^{-1} (\hat{\mu}_0 - \hat{\mu}_1)} \tag{2.19}$$

- DS method: This is a modified version of the D method, in which the estimator of Mahananobis distance is scaled with a weight to make it unbiased.

$$D_{ds} = \sqrt{\frac{n-p-1}{n}} \sqrt{(\hat{\mu}_0 - \hat{\mu}_1)^T S^{-1} (\hat{\mu}_0 - \hat{\mu}_1)}, \tag{2.20}$$

  where $\hat{\mu}_0$, $\hat{\mu}_1$, and S are defined as in (2.8) and (2.10), respectively.

- O method: This is based on Okamoto's asymptotic expansion of ε^0 and ε^1 [69] with

Δ replaced by D.

- OS method: This is the unbiased version of the O method where D is obtained as in the DS method.

- $\bar{U}$ method treats the discriminant function as a Gaussian random variable and estimates ε^0 and ε^1 separately by using cross-validation in combination with the above normal-based approach.

D. Bootstrap Estimation Methods

The bootstrap method originated from Quenouille [91], Tukey [92] and Hartigan [93, 94, 95]. Efron first officially proposed it as a general statistics method in [96]. Bootstrap was then further developed in [97, 98, 99, 100, 101, 102]. It has been used in a very wide range of applications, namely engineering, social science [103], economics [104], biology [105], and in particular statistics. This section gives a brief review on the application of the methods in discriminant analysis and statistics, which is, in one way or another, related to this dissertation. Details about implementation of bootstrap in other fields can be found in the mentioned references and many others elsewhere.

1. Bootstrap in Classical Statistics

In [96], Efron presented a general principle for the bootstrap method, which was then applied for multiple kinds of statistic including the error rate of prediction rules. That principle can be briefly described as follows. First, consider the one-sample situation. Suppose we have a random sample of size n observed from a completely unknown distribution F.

$$X_i = x_i, X_i \sim F, \ \forall i = 1, 2, \ldots, n. \tag{2.21}$$

Denote $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ and $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ the random sample and its observed realization. Suppose $R(\mathbf{X}, F)$ is an statistic of interest, which we wish to estimate based on the observation $\mathbf{x}$. Then the bootstrap estimate R^* of $R(\mathbf{X}, F)$ can be constructed as follows:

- Construct the empirical distribution $\hat{F}$ by putting mass 1/n at each point $x_i, \forall i = (1, 2 \ldots, n)$.
- Draw a random sample of size n from $\hat{F}$,say

  $$X_i^* = x_i^*, X_i^* \sim \hat{F},\ \forall i = 1, 2, \ldots, n. \tag{2.22}$$

  In simple words, this process means resampling $\mathbf{X}$ uniformly with replacement n times.
- Approximate $R(\mathbf{X}, F)$ by $R^*(\mathbf{X}^*, \hat{F})$

In practice, when it is not possible to get a closed form of $R^*(\mathbf{X}^*, \hat{F})$, it is often estimated by the sampling estimators of $R^*(\mathbf{X}^*, \hat{F})$by implementing the Monte Carlo approximation i.e repeating the above process multiples of times.

The two-sample situation can be expanded using the same principle. For example, in the case of binary classification, we have the training sample S_i from the class Π_i, $i = 0, 1$. A *bootstrap sample* S^* can be defined in two ways: S^* may contain n samples drawn uniformly, with replacement, from S (full bootstrap sampling); or the process may be applied to S_0 and S_1 independently, producing bootstrap samples S_0^* and S_1^*, and one lets $S^* = S_0^* \cup S_1^*$ (stratified bootstrap sampling). The development that we present in this paper is valid in either case; but the latter case is sometimes preferred due to smaller computational complexity when applying the complete bootstrap method in the next section, and also due to the fact that it is consistent with the stratified sampling of the data into S_0 and S_1, the sampling setting that is assumed here.

In either the full or stratified bootstrap sampling case, some of the samples in S may appear multiple times in S^*, whereas others may not appear at all. Let C be a vector of size n, where the i-th component counts the number of appearances in S^* of the i-th sample in S. In addition, we consider the partition $C = C_0 \cup C_1$, where C_0 (resp. C_1) is the vector containing the first n_0 (resp. last n_1) components of C. We call C a *bootstrap vector*. For given S, the vector C (or, equivalently, C_0 and C_1) uniquely determines the bootstrap sample S^*. In the full bootstrap sampling case, C has a multinomial distribution with parameters $(n, 1/n, \ldots, 1/n)$, that is,

$$P(C = (i_1, \ldots, i_n)) = \frac{1}{n^n} \frac{n!}{i_1! \cdots i_n!}, \quad i_1 + \cdots + i_n = n, \tag{2.23}$$

whereas in the stratified bootstrap sampling case, the distribution of C is a product of two multinomial distributions with parameters $(n_0, 1/n_0, \ldots, 1/n_0)$ and $(n_1, 1/n_1, \ldots, 1/n_1)$,

$$P(C = (i_1, \ldots, i_n)) = \frac{1}{n_0^{n_0} n_1^{n_1}} \frac{n_0! n_1!}{i_1! \cdots i_n!}, \quad i_1 + \cdots + i_{n_0} = n_0,\, i_{n_0+1} + \cdots i_n = n_1\,. \tag{2.24}$$

Bootstrap has been used extensively in estimating a number of standard statistics such as mean, median, confidence interval, and particular variance. There are mathematically rigorous works on the asymptotic behaviors of bootstrap, i.e when the number of samples goes to infinity [106, 107, 108].

Even though there are some controversial opinions about whether or not bootstrap is valid in every scenarios, bootstrap has been still receiving positive feedback on its wide applicability. There are other versions of resampling scheme. More can be found in the statistics literature.

While the asymptotic behavior of bootstrap has been studied, small sample properties are not well understood, in particular when it comes to estimating the error rates of classification rules. Young [109] calls for more practical research on bootstrap, i.e., for the case of finite samples. According to Chernick, Murthy and Nealy [110], “Although large sample properties of bootstrap have been studied, little is known about its small sample behavior". Shao and Tu [111] state: "Fixed sample (especially small sample) properties are also important. Unfortunately, the bootstrap estimators are usually complicated, so that we can only assess their fixed sample properties by empirical simulations carried out under some special circumstances."

2. Bootstrap Error Estimation

There has been a considerable amount of research on bootstrap error estimation methods and they have been shown to usually outperform the traditional methods of resubstitution and cross-validation, in terms of root mean square (RMS) error. [99, 102, 100, 112, 113, 114, 84, 115, 116, 117, 118, 110, 119, 120, 121, 43, 122].

Bootstrap was originally used to estimate the optimistic bias of the resubstitution error from the the true error [96, 99].

$$R(\mathbf{X}, F) = w = \varepsilon - \hat{\varepsilon}_r \tag{2.25}$$

The bootstrap estimate of w, $\hat{w}^b$ is

$$\hat{w}_b = E_* \left[\sum_i \left(\frac{1}{n} - P_i^* \right) I_{Y_i = \psi(S_n^*, X_i)} \right] \tag{2.26}$$

where $P_i^* = \frac{|\{X_j^* = x_i\}|}{n} = \frac{C(i)}{n}$ $\forall i = (1, 2, \ldots, n)$.
The standard bootstrap was defined as

$$\hat{\varepsilon}_b = \hat{\varepsilon}_r + \hat{w}_b. \tag{2.27}$$

The actual proportion of times a data point (X_i, Y_i) appears in a bootstrap sample S_n^* can be written as $P_i^* = \frac{1}{n} \sum_{j=1}^n I_{(X_j^*, Y_j^*) = (X_i, Y_i)}$, where $I_S = 1$ if the statement S is true, zero otherwise. The basic bootstrap is given by (note that S_n is fixed here):

$$\hat{\varepsilon}_0 = \frac{\sum_{b=1}^{B} \sum_{i=1}^{n} |Y_i - \Psi_n(S_n^{*b})(X_i)| \, I_{P_i^{*b}=0}}{\sum_{b=1}^{B} \sum_{i=1}^{n} I_{P_i^{*b}=0}}. \tag{2.28}$$

with the number of bootstrap sample B being between 25 and 200, as recommended in [99]. This is known as the *bootstrap zero estimator* [99].]

Bootstrap 632 is a variant of bootstrap which tries to correct the bias of the basic bootstrap estimator by performing an average with the resubstitution estimator [99]:

$$\hat{\varepsilon}_{b632} = (1 - 0.632)\hat{\varepsilon}_r + 0.632\hat{\varepsilon}_0 \tag{2.29}$$

Bootstrap 632 plus is another modified version of bootstrap, proposed in [102], which is intended for highly-overfitting classification rules. Bootstrap 632 attempts to adaptively find the weights in (3.9) that offset the effects of overfitting. The weights depend on the *relative overfitting rate* R and *no-information error rate* α. In dichotomous classification, R and α are estimated from $\hat{p}_1$, the proportion of observed samples belonging to class 1 and $\hat{q}_1$, the proportion of classifier outputs belonging to class 1. The relations are as follows

$$\begin{aligned}
\hat{\alpha} &= \hat{p}_1(1 - \hat{q}_1) + \hat{q}_1(1 - \hat{p}_1), \\
\hat{R} &= \frac{\hat{\varepsilon}_0 - \hat{\varepsilon}_r}{\hat{\alpha} - \hat{\varepsilon}_r}, \\
\hat{w_{b632+}} &= \frac{.632}{1 - .368\hat{R}}, \\
\hat{\varepsilon}_{b632+} &= (1 - \hat{w_{b632+}})\hat{\varepsilon}_r + \hat{w_{b632+}}\hat{\varepsilon}_0 .
\end{aligned} \tag{2.30}$$

In [99], Efron also proposed a set of variants of resampling schemes including double, randomized, and randomized double bootstrap, which are corresponding to the variants of bootstrap estimators.

3. Empirical Bootstrap Error Estimation

In this section, we highlight some of the most substantial papers on the topic of empirical bootstrap error estimation since Efron proposed the idea until recently.

In [99], Efron expanded the resampling scheme idea for predicting the error rate of a prediction rule. Besides formulating the problem, he ran a simulation study to compare the five variants of bootstrap with the synthetic data of Gaussian distribution of 2 and 5 dimensions, sample size 14 and 20. In [102], Efron presented an improved version of the bootstrap .632 estimator called the bootstrap .632+, which is specifically designed for dischotomous classification when the classification rules is highly overfitting. The overfitting property makes the apparent error almost zero, which eliminates the ability of balancing between optimistic and pessimistic biases in the bootstrap .632 estimation. As a result, a more appropriate convex scalar is needed to find and the bootstrap .632+ estimator is expected to find give a better balanced combination in term of unbiasedness.

In [117], Chatterjee and Chatterjee presented a comparison empirical study of bootstrap and other estimation methods including parametric substitution, resubstitution, split-sample, and jackknife for linear classifiers. Their results on the synthetic data of univariate gaussian model with three sample size 10, 20, and 50 and three real datasets of small, medium, and large sample size gave complementary remarks on bootstrap methods.

In [113, 110], Chernick, Murthy, and Nealy studied bootstrap in the context of small-samples for classification problem of two and three classes ($n = 12, 20$, and 29 for two- and five-dimensional Gaussian vectors. By using two other resampling procedures other than the original one by Efron [96], they proposed two more variants of bootstrap named MC estimator and convex bootstrap, corresponding to their new resampling methods. Their first new resampling procedure was based on the observation in another work by the same authors [110] that while the asymptotic probability of a sample point that will not be included in the bootstrap sample is approximately .368, this probability is much smaller for small n. For example for $n = 14$, the odds are 0.354. So in the MC estimator, the individual bootstrap samples were controlled to contain a certain proportion of the training set. The other new resampling of Chernick et al was to construct the new samples by taking convex combinations of the original data. Based on those, they compared seven estimators including the apparent, leave-one-out, zero bootstrap, .632 bootstrap, standard bootstrap, MC and convex estimators for linear discriminant analysis.

Jain, Dubes, and Chen reported in their paper [118] favorable results of bootstraps in term of the estimated confidence intervals with respect to the other estimation methods with 1-NN, quadratic, and Fisher classification rules on simulation and three real data sets.

In [119], Raudys suggested that the well-known decrease in bias of the standard bootstrap was due to the negative correlation between the apparent error and the bias w. He stated that this correlation increased as the sample size got smaller or the classification problem became more difficult, i.e the asymptotic error was larger, which was supported by his theoretical establishment under asymptotic settings. Raudys also presented complimentary simulation results for linear and Parzen-windown classification rule under Gaussian and mixed Gaussian models.

In [114], a study of the effects of finite sample sizes on the performance of classifiers by Fukunaga and Hayes, statistical properties of the bootstrap was analyzed. They provided a general framework for theoretical analysis of the standard bootstrap in the form of ”manageable” expressions for linear and quadratic classifiers under Gaussian assumptions.

The dominance of the bootstrap estimation was again confirmed in a review of advances in estimating the misclassification rate in 1987 by McLachlan [84]. The bootstrap technique and its variants from the seminal paper of Efron [96, 99] were considered as the main factor which trgiggered a series of works leading to improved estimators of error rates by appropriate bias correction and small vatriance.

Molinaro, Simon, and Pfeiffer published an extensive comparison study on the resampling error estimation methods in [123]. Different estimation methods including twofold, fivefold, tenfold, leave-one-out, split one-third, split one-half, .632+ were implemented on the microarray and mass spectrometry proteomics data. They ran the simulations for a number of classification rules such as diagonal linear discriminant, linear, CART and nearest-neighbor classification rules. The .632+ was reports as the best methods when the signal-to-noise ratios are moderate or weak. Moreover, the differences between resampling methods were observed to decrease as the sample sizes increase.

In [124], Fu and Carroll presented a study of combining the two competing resampling methods, bootstrap and cross-validation on microarray data. In their methods, a cross-validation estimation was implemented on each bootstrap sample and the final estimate was the sample mean of a number of cross-validation estimates. The simulation results using that simple combination idea was reported to be promising for small sample sizes and applicable for both parametric and nonparametric classification rules.

In [120, 121, 43], the authors provided substantial experimental studies on the performance of error estimation methods for different classification rules when the sample sizes are limited. Based on the root-mean-squared errors obtained on both synthetic models and microarray data, bootstrap error estimation were confirmed to be among the most competitive methods.

In [112], Sima and Dougherty presented a study, in which the bias of the bootstrap estimators were to be removed by finding the optimal convex scalar.

There also other works on the resampling methods for non-normality situations [125, 126].

This review on the empirical bootstrap is by no means exhaustive. It mostly focuses on featuring some of the most typical works in the applications of resampling error estimation methods. More references on the topic can be found in the papers [122, 127, 128, 129, 130, 131, 132].

4. Complete Bootstrap Estimation in Small-sample Settings

In practice, bootstrap estimators are often obtained by Monte Carlo approximation, meaning the resampling process is iterated a number of times, each times will yield an estimate. The bootstrap estimator is the sample mean of these estimates. Choosing the optimal number of iteration has been a topic in the research of bootstrap methods.

Since each possible bootstrap sample S^* from the training data S is associated in one-to-one correspondence with a unique bootstrap vector C, we may write $S^* = T_C(S)$, for some C. Note that the original sample set itself is included: if $C = (1,\ldots,1) \overset{\text{def}}{=} \mathbf{1}_n$, then $S^* = T_{\mathbf{1}_n}(S) = S$, since each original sample point appears once in the bootstrap sample. Note however that the number of distinct bootstrap samples, i.e., values for C, is equal to $\binom{2n-1}{n}$ and $\binom{2n_0-1}{n_0}\binom{2n_1-1}{n_1}$ in the full and stratified bootstrap sampling cases, respectively; even for small n_0, n_1, and n, these are very large numbers. For example, in the full bootstrap sampling case, the total number of possible bootstrap samples of size $n = 20$ is larger than 6.8×10^{10}.

Given the fact that the total number of distinct bootstrap samples C grows exponentially fast when n increases, it is almost impossible to compute the exact bootstrap as n is moderate or large. In stead, a Monte Carlo approximation is often implemented as the second method proposed by Efron [96]. For small sample case, which is prevalent in many genomics and proteomics application, complete bootstrap becomes feasible and is of practical interest.

The complete bootstrap method, which goes through all the distinct bootstrap samples and is assumed here, was argued to be competitive and practical for small samples by Fisher and Hall [133], and to be sometimes even computationally cheaper than the more common Monte-Carlo bootstrap by Diaconis and Holmes [134]. Other papers have studied the properties of the complete bootstrap method in small sample cases [135, 136, 137].

The first part of this dissertation including chapter III and IV is dedicated to theoretical analysis of complete bootstrap error estimation of linear discriminant analysis under standard Gaussian assumption. The analysis is concerned with establishing the moments of the bootstrap estimation , and as a result, the bias, variance, and root mean square of deviation from the true error, which are the usual metrics to globally evaluate estimation methods.

E. Applications of Misclassification Error Estimation

Error estimation plays a very important roles in every statistical inference problem. Whenever it comes to evaluation of the statistical inference algorithm, estimating the error rate needs to be implemented. This fact is explicitly demonstrated in the practical applications of statistical learning, in particular supervised learning, in biomedical research.

First, in the area of genomics-based and proteomics-based class prediction and comparison, the outputs are predictors such as genes, peptides etc which expectedly have discriminatory power to classify different disease states, i.e normal v.s diseased, or different cancer subtypes. How accurately these predictors can work is a crucial question in the process of biomarker discovery and validation. It is the evaluation process that examines the validity of the discovered biomarkers. The efficacy of these biomarkers when they are integrated in future practical routine of diagnosis and prognosis of cancer and other diseases entirely depends on the reliability and accuracy of the validation procedures. One of the drawbacks, which hinder the realization of the quantitatively found biomarkers into the clinical practice routine is the failure of validation process required by the FDA [37, 38].

Second, in the problem of class discovery such as classifying cancers into subtypes or biclustering in functional Genomics, evaluation of the statistical algorithms used to discover classes is even more important given the fact that we do not know the ground truth under the dataset but are trying to discover it. This interprets literally as the problem of error estimation for unsupervised learning or clustering, which is generally harder than the supervised learning. In future medicine, distinct subtypes of diseases which originate from different causes are to be handled with different treatments with hopefully better efficacy. Failure to correctly distinguish cancer subtypes can result consequences such as treatment cost and efficacy.

Another typical example of the important role of error estimation lies in inferring gene regulatory networks [138]. Studying the biological pathways, in particular the regulatory mechanism of the genomes is crucial in accelerating the understanding the molecular mechanism of cancer and other diseases. Based on the high-throughput data, gene regulatory networks are attempted to be inferred using different models. Interested readers can find more about this topic in the review paper [139, 140]. Obviously, in order to ascertain our knowledge of cancers and diseases from these findings, it is foremost to confirm the validity of the discovered gene networks.

In addition, estimation of the accuracy prediction of transcriptional binding factors deserved more attention regarding its wide applicability in understanding the regulatory mechanism of the genome [141].

This chapter covers the main points of supervised learning and provides a review of error estimation as well as some highlights of its applications. The next chapter is devoted to the analysis of bootstrap estimation methods.

CHAPTER III

BOOTSTRAP ERROR ESTIMATION - UNIVARIATE MODEL *

This chapter presents the theoretical analysis of complete bootstrap error estimation for linear discriminant analysis under univariate Gaussian model. The variances of the label feature distribution are assumed to be known. The analysis is concerned with some bootstrap estimators including zero, .632, and convex bootstrap estimation. The results include the first moments, the second moments, the correlation of these bootstrap estimators with the true error and the resubstitution estimator. As a result, we obtain the exact formulas for the bias, variance, and the root mean square of the estimation deviations from the true error, which are the usual metrics for evaluation of estimation methods. Also, we propose unbiased bootstrap estimation by zeroing the deviation bias and optimal bootstrap estimation by minimizing the root mean square of the deviation. All the formulas are involved with multivariate Gaussian random variables, up to dimension 4. Given the increasing difficulty of complete bootstrap computation as the number of samples increases, an efficient algorithm is introduced to compute the complete enumeration for up to moderate sample sizes. Finally, some figures of the optimal convex scalar for the unbiased bootstrap estimation are provided for different number of samples under various Gaussian models.

* Part of this chapter is reprinted with permission from ”Unbiased Bootstrap Error Estimation for Linear Discriminant Analysis.” by T. T. Vu, U. M. Braga-Neto, and E. R. Dougherty, 2010. submitted, copyright 2010 of *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

A. The Bias, Variance, and RMS of Estimation Deviation

Let $\hat{\varepsilon}$ be an estimator for the true error ε, then the bias, variance, and RMS of estimation deviation are defined as followings:

$$\text{Bias}[\hat{\varepsilon}] = \text{E}[\hat{\varepsilon} - \varepsilon], \tag{3.1}$$

$$\text{Var}_d[\hat{\varepsilon}] = \text{Var}[\hat{\varepsilon} - \varepsilon] = \text{Var}[\hat{\varepsilon}] - 2\text{Cov}[\hat{\varepsilon}\varepsilon] + \text{Var}[\varepsilon], \tag{3.2}$$

$$\text{RMS}[\hat{\varepsilon}] = \sqrt{\text{E}[(\varepsilon - \hat{\varepsilon})^2]} = \sqrt{\text{E}[\varepsilon^2] - 2\text{E}[\varepsilon\hat{\varepsilon}] + \text{E}[\hat{\varepsilon}^2]}. \tag{3.3}$$

It is simple to check that

$$\text{RMS}^2[\hat{\varepsilon}] = \text{Bias}^2[\hat{\varepsilon}] + \text{Var}_d[\hat{\varepsilon}]. \tag{3.4}$$

While the bias represents that average centrality of the estimator around the true error, the deviation variance measures the dispersion of the estimator from the true error. The optimal estimator is the uniformly unbiased minimum variance one. There is a trade-off between the bias and the variance. So, the ultimate metric to evaluate an estimator is RMS, which combines bias and variance. From (3.3), we can see that to compute RMS$[\hat{\varepsilon}]$, we need to know the second moments of the true error and the estimator, as well as the correlation between them. The main sections of this chapter present theorems to compute the moments of some bootstrap estimators and their correlation with the true error, and so ultimately allows us to obtain the RMS of these bootstrap estimators using the relation (3.3).

B. The Bootstrapped Linear Discriminant Analysis

Let S^* denote the *bootstrap sample* uniformly taken with replacement from S with the same size like S and the corresponding weight vector C. All the probability formulas derived herein assume C is given. Otherwise, it is explicitly stated. For brevity, we will omit the conditional notation of C. Let $\psi_C = \Psi(S^*)$ be the classifier designed on S^* using the same classification rule Ψ.

The *classification error rate* ε_C of classifier ψ_C

$$\begin{aligned} \varepsilon_C &= (1-\gamma)\,\mathrm{P}(\psi_C(X)=1 \mid X \in \Pi_0) + \gamma \mathrm{P}(\psi_C(X)=0 \mid X \in \Pi_1) \\ &\stackrel{\text{def}}{=} (1-\gamma)\,\varepsilon_C^0 + \gamma \varepsilon_C^1 . \end{aligned} \tag{3.5}$$

We can define a "test-set" error estimator $\hat{\varepsilon}_C$ for ε_C as the average error committed by the bootstrap classifier ψ_C on the data left out of the bootstrap sample:

$$\hat{\varepsilon}_C = \frac{1}{\sum_{i=1}^{n} \mathrm{I}_{C(i)=0}} \left[\sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0}\, \mathrm{I}_{\psi_C(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\, \mathrm{I}_{\psi_C(X_i)=0} \right] \tag{3.6}$$

where $C(i)$ denotes the i-th component of vector C.

With our assumption of complete bootstrap, the zero bootstrap error estimator is defined as the expected value of $\hat{\varepsilon}_C$ over the bootstrap sampling mechanism, i.e., over the distribution of C:

$$\hat{\varepsilon}_0 = \mathrm{E}[\hat{\varepsilon}_C \,|\, S] = \sum_C \hat{\varepsilon}_C \mathrm{P}(C). \tag{3.7}$$

It can be seen that the zero bootstrap error estimator defined as in (2.28) is a Monte Carlo approximation version of (3.7).

The more popular variants of bootstrap estimation are *.632 bootstrap estimator* and *convex bootstrap estimator*. The *.632+ bootstrap estimator* is a special case of the latter for the dichotomous classification problem, in which the convex scalar w is found adaptively with the "relative overfitting rate" (See (2.30).

$$\hat{\varepsilon}_{\mathrm{b}632} = (1-0.632)\,\hat{\varepsilon}_r + 0.632\,\hat{\varepsilon}_0\,, \tag{3.8}$$

$$\hat{\varepsilon}_{\mathrm{b}632+} = (1-\hat{w}_{b632+})\,\hat{\varepsilon}_r + \hat{w}_{b632+}\,\hat{\varepsilon}_0\,, \tag{3.9}$$

More generally, we have the convex bootstrap estimate:

$$\hat{\varepsilon}_{\mathrm{w}} = (1-w)\,\hat{\varepsilon}_r + w\,\hat{\varepsilon}_0\,. \tag{3.10}$$

In the followings, we establish some useful relations of the error rate $\hat{\varepsilon}_C$ to compute its moments. Define the following notations:

$$m_0(C) = \sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0}, \quad m_1(C) = \sum_{i=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}, \quad m(C) = m_0(C) + m_1(C), \tag{3.11}$$

$$s_0(C) = \frac{1}{n_0^2}\sum_{i=1}^{n_0} C^2(i), \quad s_1(C) = \frac{1}{n_1^2}\sum_{i=n_0+1}^{n_0+n_1} C^2(i), \quad s(C) = s_0(C) + s_1(C), \tag{3.12}$$

$$r_0(C_1,C_2) = \frac{1}{n_0^2}\sum_{i=1}^{n_0} C_1(i)C_2(i), \quad r_1(C_1,C_2) = \frac{1}{n_1^2}\sum_{i=n_0+1}^{n_0+n_1} C_1(i)C_2(i). \tag{3.13}$$

It is clear that $r_i(C,C) = s_i(C)$ for $i \in \{0,1\}$. While these numbers m, s, and r are functions of C, we will omit the notations Cs throughout the work for brevity in some of the results, unless keeping them is necessary to differentiate different bootstrap vectors Cs. Also, suppose $X^* \in \Pi_0$, and $X^{**} \in \Pi_1$ are two samples independent of S_n.

1. First Moment

From (3.6), we have:

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_C] &= \mathrm{E}\left\{ \frac{1}{m}\left[\sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0}\mathrm{I}_{\psi_C(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\mathrm{I}_{\psi_C(X_i)=0} \right] \right\} \\
&= \frac{1}{m}\sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0}\mathrm{E}[\mathrm{I}_{\psi_C(X_i)=1}] + \frac{1}{m}\sum_{i=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\mathrm{E}[\mathrm{I}_{\psi_C(X_i)=0}] \\
&= \frac{1}{m}\sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0}\mathrm{P}\{\psi_C(X_i)=1\} + \frac{1}{m}\sum_{i=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\mathrm{P}\{\psi_C(X_i)=0\}
\end{aligned}$$

So,

$$\mathrm{E}[\hat{\varepsilon}_C] = \frac{m_0}{m}\mathrm{P}\{\psi_C(X^*)=1\} + \frac{m_1}{m}\mathrm{P}\{\psi_C(X^{**})=0\} \tag{3.14}$$

2. Second Moment

From (3.6), we have:

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_C^2] &= \mathrm{E}\left\{ \frac{1}{m^2} \left[\sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0} \mathrm{I}_{\psi_C(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0} \mathrm{I}_{\psi_C(X_i)=0} \right]^2 \right\} \\
&= \mathrm{E}\left\{ \frac{1}{m^2} \left[\sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0} \mathrm{I}_{\psi_C(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0} \mathrm{I}_{\psi_C(X_i)=0} + \right.\right. \\
&\quad + \sum_{i=1}^{n_0} \sum_{j \neq i}^{n_0} \mathrm{I}_{C(i)=0,\, C(j)=0} \mathrm{I}_{\psi_C(X_i)=1,\, \psi_C(X_j)=1} + \\
&\quad + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j \neq i}^{n_0+n_1} \mathrm{I}_{C(i)=0,\, C(j)=0} \mathrm{I}_{\psi_C(X_i)=0} \mathrm{I}_{\psi_C(X_j)=0} + \\
&\quad + \sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0,\, C(j)=0} \mathrm{I}_{\psi_C(X_i)=1,\, \psi_C(X_j)=0} + \\
&\quad \left.\left. + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j=1}^{n_0} \mathrm{I}_{C(i)=0,\, C(j)=0} \mathrm{I}_{\psi_C(X_i)=0,\, \psi_C(X_j)=1} \right] \right\}
\end{aligned}$$

So,

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_C^2] &= \frac{m_0}{m^2} \mathrm{P}\{\psi_C(X^*) = 1\} + \frac{m_1}{m^2} \mathrm{P}\{\psi_C(X^{**}) = 0\} + \\
&\quad + \frac{1}{m^2} \left[\sum_{i=1}^{n_0} \sum_{j \neq i}^{n_0} \mathrm{I}_{C(i)=0, C(j)=0} \mathrm{P}\{\psi_C(X_i) = 1, \psi_C(X_j) = 1\} + \right. \\
&\quad + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j \neq i}^{n_0+n_1} \mathrm{I}_{C(i)=0, C(j)=0} \mathrm{P}\{\psi_C(X_i) = 0, \psi_C(X_j) = 0\} + \\
&\quad + \sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0, C(j)=0} \mathrm{P}\{\psi_C(X_i) = 1, \psi_C(X_j) = 0\} + \\
&\quad \left. + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j=1}^{n_0} \mathrm{I}_{C(i)=0, C(j)=0} \mathrm{P}\{\psi_C(X_i) = 0, \psi_C(X_j) = 1\} \right].
\end{aligned} \tag{3.15}$$

3. Cross Correlation

From (3.6), we have for $C_1 \neq C_2$:

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_{C_1}\hat{\varepsilon}_{C_2}] &= \mathrm{E}\left\{ \frac{1}{m(C_1)} \left[\sum_{i=1}^{n_0} \mathrm{I}_{C_1(i)=0} \mathrm{I}_{\psi_{C_1}(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} \mathrm{I}_{C_1(i)=0} \mathrm{I}_{\psi_{C_1}(X_i)=0} \right] \times \right. \\
&\quad \left. \times \frac{1}{m(C_2)} \left[\sum_{j=1}^{n_0} \mathrm{I}_{C_2(j)=0} \mathrm{I}_{\psi_{C_2}(X_j)=1} + \sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C_2(j)=0} \mathrm{I}_{\psi_{C_2}(X_j)=0} \right] \right\} \\
&= \mathrm{E}\left\{ \frac{1}{m(C_1)m(C_2)} \left[\sum_{i=1}^{n_0} \sum_{j=1}^{n_0} \mathrm{I}_{C_1(i)=0} \mathrm{I}_{\psi_{C_1}(X_i)=1} \mathrm{I}_{C_2(j)=0} \mathrm{I}_{\psi_{C_2}(X_j)=1} + \right. \right. \\
&\quad + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C_1(i)=0} \mathrm{I}_{\psi_{C_1}(X_i)=0} \mathrm{I}_{C_2(j)=0} \mathrm{I}_{\psi_{C_2}(X_j)=0} + \\
&\quad + \sum_{i=1}^{n_0} \sum_{j=n_0}^{n_0+n_1} \mathrm{I}_{C_1(i)=0} \mathrm{I}_{\psi_{C_1}(X_i)=1} \mathrm{I}_{C_2(j)=0} \mathrm{I}_{\psi_{C_2}(X_j)=0} + \\
&\quad \left. \left. + \sum_{i=n_0}^{n_0+n_1} \sum_{j=1}^{n_0} \mathrm{I}_{C_1(i)=0} \mathrm{I}_{\psi_{C_1}(X_i)=0} \mathrm{I}_{C_2(j)=0} \mathrm{I}_{\psi_{C_2}(X_j)=1} \right] \right\}
\end{aligned}$$

So, the correlation between ”hold-out” errors of any two distinct C-bootstrap linear classifiers is

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_{C_1}\hat{\varepsilon}_{C_2}] = \frac{1}{m(C_1)m(C_2)} & \left[\sum_{i=1}^{n_0} \sum_{j=1}^{n_0} \mathrm{I}_{C_1(i)=0, C_2(j)=0} \mathrm{P}\{\psi_{C_1}(X_i)=1, \psi_{C_2}(X_j)=1\} + \right. \\
& + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C_1(i)=0, C_2(j)=0} \mathrm{P}\{\psi_{C_1}(X_i)=0, \psi_{C_2}(X_j)=0\} + \\
& + \sum_{i=1}^{n_0} \sum_{j=n_0}^{n_0+n_1} \mathrm{I}_{C_1(i)=0, C_2(j)=0} \mathrm{P}\{\psi_{C_1}(X_i)=1, \psi_{C_2}(X_j)=0\} + \\
& \left. + \sum_{i=n_0}^{n_0+n_1} \sum_{j=1}^{n_0} \mathrm{I}_{C_1(i)=0, C_2(j)=0} \mathrm{P}\{\psi_{C_1}(X_i)=0, \psi_{C_2}(X_j)=1\} \right].
\end{aligned} \tag{3.16}$$

4. Cross Moment with Resubstitution Estimator

We are interested in the correlation between the ”hold-out” error $\hat{\varepsilon}_C$ of the C-bootstrapped classifier $\psi_C(X)$ and the resubstitution estimator $\hat{\varepsilon}_r$ of the original classifier $\psi(X)$. From

(2.17) and (3.6), we have

$$
\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_C \hat{\varepsilon}_r] = \mathrm{E}\Bigg\{ & \frac{1}{m}\left[\sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0}\,\mathrm{I}_{\psi_C(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\,\mathrm{I}_{\psi_C(X_i)=0}\right] \times \\
& \times \frac{1}{n}\left[\sum_{j=1}^{n_0} \mathrm{I}_{\psi(X_j)=1} + \sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{\psi(X_j)=0}\right]\Bigg\} \\
= \mathrm{E}\Bigg\{ & \frac{1}{nm}\Bigg[\sum_{i=1}^{n_0}\sum_{j=1}^{n_0} \mathrm{I}_{C(i)=0}\,\mathrm{I}_{\psi_C(X_i)=1}\mathrm{I}_{\psi(X_j)=1} + \sum_{i=n_0+1}^{n_0+n_1}\sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\,\mathrm{I}_{\psi_C(X_i)=0}\mathrm{I}_{\psi(X_j)=0} + \\
& + \sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\,\mathrm{I}_{\psi_C(X_i)=1}\mathrm{I}_{\psi(X_j)=0} + \sum_{i=n_0+1}^{n_0+n_1}\sum_{j=1}^{n_0} \mathrm{I}_{C(i)=0}\,\mathrm{I}_{\psi_C(X_i)=0}\mathrm{I}_{\psi(X_j)=1}\Bigg]\Bigg\}
\end{aligned}
$$

So,

$$
\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_C \hat{\varepsilon}_r] = \frac{1}{nm}\Bigg[& \sum_{i=1}^{n_0}\sum_{j=1}^{n_0} \mathrm{I}_{C(i)=0}\,\mathrm{P}\{\psi_C(X_i)=1, \psi(X_j)=1\} + \\
& + \sum_{i=n_0+1}^{n_0+n_1}\sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\,\mathrm{P}\{\psi_C(X_i)=0, \psi(X_j)=0\} + \\
& + \sum_{i=1}^{n_0}\sum_{j=n_0}^{n_0+n_1} \mathrm{I}_{C(i)=0}\,\mathrm{P}\{\psi_C(X_i)=1, \psi(X_j)=0\} + \\
& + \sum_{i=n_0}^{n_0+n_1}\sum_{j=1}^{n_0} \mathrm{I}_{C(i)=0}\,\mathrm{P}\{\psi_C(X_i)=0, \psi(X_j)=1\}\Bigg].
\end{aligned}
\tag{3.17}
$$

5. Cross Moment with True Error

It is useful to know the correlation between $\hat{\varepsilon}_C$ of the C-bootstrapped classifier and the true error ε in the next sections. From (2.5) and (3.6), we have

$$
\begin{aligned}
\mathrm{E}[\varepsilon\hat{\varepsilon}_C] &= \mathrm{E}\left\{\left((1-\gamma)\,\varepsilon^0 + \gamma\varepsilon^1\right)\frac{1}{m}\left[\sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0}\,\mathrm{I}_{\psi_C(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\,\mathrm{I}_{\psi_C(X_i)=0}\right]\right\} \\
&= \frac{1-\gamma}{m}\sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0}\,\mathrm{E}\left[\varepsilon^0 \mathrm{I}_{\psi_C(X_i)=1}\right] + \frac{1-\gamma}{m}\sum_{i=n_0}^{n_0+n_1} \mathrm{I}_{C(i)=0}\,\mathrm{E}\left[\varepsilon^0 \mathrm{I}_{\psi_C(X_i)=0}\right] + \\
&\quad + \frac{\gamma}{m}\sum_{i=1}^{n_0} \mathrm{I}_{C(i)=0}\,\mathrm{E}\left[\varepsilon^1 \mathrm{I}_{\psi_C(X_i)=1}\right] + \frac{\gamma}{m}\sum_{i=n_0}^{n_0+n_1} \mathrm{I}_{C(i)=0}\,\mathrm{E}\left[\varepsilon^1 \mathrm{I}_{\psi_C(X_i)=0}\right].
\end{aligned}
$$

$$\begin{aligned}\mathrm{E}[\varepsilon\hat{\varepsilon}_C] = \frac{m_0(1-\gamma)}{m}\mathrm{E}\left[\varepsilon^0 \mathrm{I}_{\psi_C(X_1)=1}\right] + \frac{m_1(1-\gamma)}{m}\mathrm{E}\left[\varepsilon^0 \mathrm{I}_{\psi_C(X_{n_0+1})=0}\right] + \\ + \frac{m_0\gamma}{m}\mathrm{E}\left[\varepsilon^1 \mathrm{I}_{\psi_C(X_1)=1}\right] + \frac{m_1\gamma}{m}\mathrm{E}\left[\varepsilon^1 \mathrm{I}_{\psi_C(X_{n_0+1})=0}\right], \text{ with } C(X_1) = C(X_{n_0+1}) = 0.\end{aligned} \tag{3.18}$$

We have

$$\begin{aligned}\mathrm{E}[\varepsilon^0 \mathrm{I}_{\psi_C(X_1)=1}] &= \mathrm{E}[\mathrm{P}\{\psi(X) = 1 | X \in \Pi_0, S_n\}\mathrm{I}_{\psi_C(X_1)=1}] \\ &= \mathrm{E}[\mathrm{E}(\mathrm{I}_{\psi(X)=1} | X \in \Pi_0, S_n)\mathrm{I}_{\psi_C(X_1)=1}] \\ &= \mathrm{E}[\mathrm{E}(\mathrm{I}_{\psi(X)=1}\,\mathrm{I}_{\psi_C(X_1)=1} | X \in \Pi_0, S_n)] \\ &= \mathrm{E}[\mathrm{I}_{\psi(X)=1}\,\mathrm{I}_{\psi_C(X_1)=1}] \\ &= \mathrm{P}\{\psi(X^*) = 1, \psi_C(X_1) = 1\}.\end{aligned}$$

Similarly for $\mathrm{E}[\varepsilon^1 \mathrm{I}_{\psi_C(X_1)=1}]$, $\mathrm{E}[\varepsilon^0 \mathrm{I}_{\psi_C(X_{n_0+1}=0}]$, $\mathrm{E}[\varepsilon^1 \mathrm{I}_{\psi_C(X_{n_0+1}=0}]$. So,

$$\begin{aligned}\mathrm{E}[\varepsilon\hat{\varepsilon}_C] = \frac{m_0(1-\gamma)}{m}\mathrm{P}\{\psi(X^*) = 1, \psi_C(X_1) = 1\} + \frac{m_0\gamma}{m}\mathrm{P}\{\psi(X^{**}) = 0, \psi_C(X_1) = 1\} + \\ + \frac{m_1(1-\gamma)}{m}\mathrm{P}\{\psi(X^*) = 1, \psi_C(X_{n_0+1}) = 0\} + \frac{m_1\gamma}{m}\mathrm{P}\{\psi(X^{**}) = 0, \psi_C(X_{n_0+1}) = 0\}.\end{aligned} \tag{3.19}$$

with $C(X_1) = C(X_{n_0+1}) = 0$.

6. The True Error

The first two moments of the true error, $\mathrm{E}[\varepsilon]$ and $\mathrm{E}[\varepsilon^2]$ (also of the resubstitution estimator $\mathrm{E}[\varepsilon_r]$ and $\mathrm{E}[\varepsilon_r^2]$), were expressed in the forms involved with probabilities of discriminant functions W in [86]. Based on that, Zollanvari et al [86] then derived the exact formulas for the univariate case and obtained approximation ones for the multivariate case. Because we need the second moment of the true error, $\mathrm{E}[\varepsilon^2]$, to compute the root mean square of the bootstrap estimators, we rewrite Zollanvari's formulas and his univariate results in our notation in this chapter, and present the exact results of the true error and the resubstitution estimator for the multivariate case in Chapter IV.

- The first moment of the the true error

  From (2.5),

$$\begin{aligned}\mathrm{E}[\varepsilon] &= \mathrm{E}[(1-\gamma)\varepsilon_0+\gamma\varepsilon_1]\\ &= \mathrm{E}\left[(1-\gamma)P\{\psi(X^*)=1\,|S_n\}+(1-\gamma)P\{\psi(X^{**})=0\,|S_n\}\right]\\ &= (1-\gamma)P\{\psi(X^*)=1\}+(1-\gamma)P\{\psi(X^{**})=0\}\end{aligned} \tag{3.20}$$

- The second moment of the the true error

  From (2.5),

$$\begin{aligned}\mathrm{E}[\varepsilon^2] &= \mathrm{E}[((1-\gamma)\varepsilon_0+\gamma\varepsilon_1)^2]\\ &= (1-\gamma)^2\mathrm{E}[\varepsilon_0\varepsilon_0]+2\gamma(1-\gamma)\mathrm{E}[\varepsilon_0\varepsilon_1]+\gamma^2\mathrm{E}[\varepsilon_1\varepsilon_1]\end{aligned}$$

  Also,

$$\begin{aligned}\mathrm{E}[\varepsilon_0\varepsilon_0] &= \mathrm{E}[\mathrm{P}\{\psi(X^*)=1\,|S_n\}\mathrm{P}\{\psi(X^{*'})=1\,|S_n\}],\\ &\quad \text{where } X^* \text{ and } X^{*'}\in\Pi_0 \text{ are independent with each other and of } S_n\\ &= \mathrm{E}[\mathrm{P}\{\psi(X^*)=1,\,\psi(X^{*'})=1\,|S_n\}\\ &= \mathrm{P}\{\psi(X^*)=1,\,\psi(X^{*'})=1\}\end{aligned}$$

  Similarly for $\mathrm{E}[\varepsilon_1\varepsilon_1]$ and $\mathrm{E}[\varepsilon_0\varepsilon_1]$. So,

$$\begin{aligned}\mathrm{E}[\varepsilon^2] = (1-\gamma)^2\mathrm{P}\{\psi(X^*)=1,\,\psi(X^{*'})=1\}+2\gamma(1-\gamma)\mathrm{P}\{\psi(X^*)=1,\,\psi(X^{**})=0\}+\\ +\gamma^2\mathrm{P}\{\psi(X^{**})=0,\,\psi(X^{**'})=0\}\end{aligned} \tag{3.21}$$

7. The Zero Bootstrap Estimation

$$\hat{\varepsilon}_0 = \mathrm{E}[\hat{\varepsilon}_C | S] = \sum_C \hat{\varepsilon}_C \mathrm{P}(C). \tag{3.22}$$

- The first moment of zero bootstrap estimator

$$\mathrm{E}\left[\hat{\varepsilon}_0\right] = \sum_C \mathrm{P}(C)\mathrm{E}[\hat{\varepsilon}_C] \tag{3.23}$$

- The second moment of zero bootstrap

$$\begin{aligned} \mathrm{E}\left[\hat{\varepsilon}_0^2\right] &= \mathrm{E}\Big[\sum_C \mathrm{P}(C)\hat{\varepsilon}_C\Big]^2 \\ &= \sum_C \mathrm{P}^2(C)\mathrm{E}\left[\hat{\varepsilon}_C^2\right] + 2\sum_{C_1 \neq C_2} \mathrm{P}(C_1)\mathrm{P}(C_2)\mathrm{E}\left[\hat{\varepsilon}_{C_1}\hat{\varepsilon}_{C_2}\right] \end{aligned} \tag{3.24}$$

- The correlation of zero bootstrap estimator with the resubstitution estimator

$$\begin{aligned} \mathrm{E}\left[\hat{\varepsilon}_r\hat{\varepsilon}_0\right] &= \mathrm{E}\Big[\hat{\varepsilon}_r \sum_C \mathrm{P}(C)\hat{\varepsilon}_C\Big] \\ &= \sum_C \mathrm{P}(C)\mathrm{E}[\hat{\varepsilon}_r\hat{\varepsilon}_C] \end{aligned} \tag{3.25}$$

- The correlation of zero bootstrap estimator with the true error

$$\begin{aligned} \mathrm{E}\left[\varepsilon\hat{\varepsilon}_0\right] &= \mathrm{E}\Big[\varepsilon \sum_C \mathrm{P}(C)\hat{\varepsilon}_C\Big] \\ &= \sum_C \mathrm{P}(C)\mathrm{E}[\varepsilon\hat{\varepsilon}_C] \end{aligned} \tag{3.26}$$

8. The Convex Bootstrap Estimation

- The first moment of convex bootstrap estimator

$$\begin{aligned} \mathrm{E}\left[\hat{\varepsilon}_w\right] &= \mathrm{E}\Big[(1-w)\hat{\varepsilon}_r + w\hat{\varepsilon}_0\Big] \\ &= (1-w)\mathrm{E}[\hat{\varepsilon}_r] + w\sum_C \mathrm{P}(C)\mathrm{E}[\hat{\varepsilon}_C] \end{aligned} \tag{3.27}$$

- The second moment of convex bootstrap estimator

$$\begin{aligned} \mathrm{E}\left[\hat{\varepsilon}_w^2\right] &= \mathrm{E}\Big[(1-w)\hat{\varepsilon}_r + w\hat{\varepsilon}_0\Big]^2 \\ &= (1-w)^2\mathrm{E}[\hat{\varepsilon}_r^2] + w^2\mathrm{E}[\hat{\varepsilon}_0^2] + 2w(1-w)\mathrm{E}[\hat{\varepsilon}_r\hat{\varepsilon}_0] \end{aligned} \tag{3.28}$$

- The correlation of convex bootstrap estimator with the true error

$$\begin{aligned} \mathrm{E}\left[\varepsilon\hat{\varepsilon}_w\right] &= \mathrm{E}\Big[\varepsilon((1-w)\varepsilon_r + w\varepsilon_0)\Big] \\ &= (1-w)\mathrm{E}[\varepsilon\hat{\varepsilon}_r] + w\mathrm{E}[\varepsilon\hat{\varepsilon}_0] \\ &= (1-w)\mathrm{E}[\varepsilon\hat{\varepsilon}_r] + \sum_C \mathrm{P}(C)\mathrm{E}[\varepsilon\hat{\varepsilon}_C] \end{aligned} \tag{3.29}$$

All the expressions in this section are applicable for any conditional distributions including both univariate and multivariate models. Following are the results derived for univariate Gaussian model.

C. Univariate Model

Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0+1, \ldots, n_0+n_1$ be a set of $n = n_0+n_1$ i.i.d. observations. In this univariate case, the W statistic becomes greatly simplified, being a function only of the sample means, and the LDA classifier is given by

$$\psi(X) = \begin{cases} 1, & \text{if } \left(X - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)(\hat{\mu}_0 - \hat{\mu}_1) < 0 \\ 0, & \text{otherwise} \end{cases}, \tag{3.30}$$

The C-bootstrap LDA classifier designed on S^* corresponding to the bootstrap vector C is obtained by replacing μ_i by μ_i^C, $i = 0, 1$, in (3.30):

$$\psi_C(X) = \begin{cases} 1, & \text{if } \left(X - \frac{\hat{\mu}_0^C+\hat{\mu}_1^C}{2}\right)(\hat{\mu}_0^C - \hat{\mu}_1^C) < 0 \\ 0, & \text{otherwise} \end{cases}, \tag{3.31}$$

where

$$
\begin{aligned}
\hat{\mu}_0^C &= \frac{1}{n_0}\sum_{i=1}^{n_0} C(i)X_i \\
\hat{\mu}_1^C &= \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} C(i)X_i
\end{aligned}
\tag{3.32}
$$

are *C-bootstrap sample means*.

Define the following Gaussian vectors with $X_{u_1}, X_{u_2} \in \Pi_0$, $X_{v_1}, X_{v_2} \in \Pi_1$, i.e. $1 \leq u_1, u_2 \leq n_0, n_0+1 \leq v_1, v_2 \leq n_0+n_1$. In the following definitions for Fs, assume $C_i(u_i) = C_i(v_i) = 0$ for $i = 1, 2$:

$$F_{00}^{I}(u_1,u_2,C_1,C_2) = \left[X_{u_1} - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2}, \hat{\mu}_1^{C_1}-\hat{\mu}_0^{C_1}, X_{u_2} - \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2}, \hat{\mu}_1^{C_2}-\hat{\mu}_0^{C_2}\right]^T, \tag{3.33}$$

$$F_{00}^{II}(u_1,u_2,C_1,C_2) = \left[X_{u_1} - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2}, \hat{\mu}_1^{C_1}-\hat{\mu}_0^{C_1}, \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2} - X_{u_2}, \hat{\mu}_0^{C_2}-\hat{\mu}_1^{C_2}\right]^T, \tag{3.34}$$

$$F_{11}^{I}(v_1,v_2,C_1,C_2) = \left[X_{v_1} - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2}, \hat{\mu}_0^{C_1}-\hat{\mu}_1^{C_1}, X_{v_2} - \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2}, \hat{\mu}_0^{C_2}-\hat{\mu}_1^{C_2}\right]^T, \tag{3.35}$$

$$F_{11}^{II}(v_1,v_2,C_1,C_2) = \left[X_{v_1} - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2}, \hat{\mu}_0^{C_1}-\hat{\mu}_1^{C_1}, \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2} - X_{v_2}, \hat{\mu}_1^{C_2}-\hat{\mu}_0^{C_2}\right]^T, \tag{3.36}$$

$$F_{01}^{I}(u_1,v_2,C_1,C_2) = \left[X_{u_1} - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2}, \hat{\mu}_1^{C_1}-\hat{\mu}_0^{C_1}, X_{v_2} - \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2}, \hat{\mu}_0^{C_2}-\hat{\mu}_1^{C_2}\right]^T, \tag{3.37}$$

$$F_{01}^{II}(u_1,v_2,C_1,C_2) = \left[X_{u_1} - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2}, \hat{\mu}_1^{C_1}-\hat{\mu}_0^{C_1}, \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2} - X_{v_2}, \hat{\mu}_1^{C_2}-\hat{\mu}_0^{C_2}\right]^T, \tag{3.38}$$

Basic algebra gives us the mean vectors and the covariance matrices as following:

$$
\begin{aligned}
&\mathrm{E}[F_{00}^{I}] = \mathrm{E}[F_{01}^{II}] = \left[\frac{\mu}{2}, -\mu, \frac{\mu}{2}, -\mu\right]^T, &&\mathrm{E}[F_{00}^{II}] = \mathrm{E}[F_{01}^{I}] = \left[\frac{\mu}{2}, -\mu, \frac{-\mu}{2}, \mu\right]^T, \\
&\mathrm{E}[F_{11}^{I}] = \left[\frac{-\mu}{2}, \mu, \frac{-\mu}{2}, \mu\right]^T, &&\mathrm{E}[F_{11}^{II}] = \left[\frac{-\mu}{2}, \mu, \frac{\mu}{2}, -\mu\right]^T.
\end{aligned}
$$

where $\mu = \mu_0 - \mu_1$, and the covariance matrices are

$$\Sigma_{F^{I}_{00}(u_1,u_2,C_1,C_2)} = \begin{pmatrix} \left(1+\frac{s_0(C_1)}{4}\right)\sigma_0^2+\frac{s_1(C_1)}{4}\sigma_1^2 & \frac{s_0(C_1)\sigma_0^2}{2}-\frac{s_1(C_1)\sigma_1^2}{2} & \left(I_{\{u_1=u_2\}}+\frac{r_0}{4}-\frac{C_2(u_1)+C_1(u_2)}{2n_0}\right)\sigma_0^2+\frac{r_1}{4}\sigma_1^2 & \left(\frac{r_0}{2}-\frac{C_2(u_1)}{n_0}\right)\sigma_0^2-\frac{r_1}{2}\sigma_1^2 \\ . & s_0(C_1)\sigma_0^2+s_1(C_1)\sigma_1^2 & \left(\frac{r_0}{2}-\frac{C_1(u_2)}{n_0}\right)\sigma_0^2-\frac{r_1}{2}\sigma_1^2 & r_0\sigma_0^2+r_1\sigma_1^2 \\ . & . & \left(1+\frac{s_0(C_2)}{4}\right)\sigma_0^2+\frac{s_1(C_2)}{4}\sigma_1^2 & \frac{s_0(C_1)\sigma_0^2}{2}-\frac{s_1(C_1)\sigma_1^2}{2} \\ . & . & . & s_0(C_2)\sigma_0^2+s_1(C_2)\sigma_1^2 \end{pmatrix},$$

$$\Sigma_{F^{II}_{00}(u_1,u_2,C_1,C_2)} = \begin{pmatrix} \left(1+\frac{s_0(C_1)}{4}\right)\sigma_0^2+\frac{s_1(C_1)}{4}\sigma_1^2 & \frac{s_0(C_1)\sigma_0^2}{2}-\frac{s_1(C_1)\sigma_1^2}{2} & \left(\frac{C_2(u_1)+C_1(u_2)}{2n_0}-I_{\{u_1=u_2\}}-\frac{r_0}{4}\right)\sigma_0^2-\frac{r_1}{4}\sigma_1^2 & \left(\frac{C_2(u_1)}{n_0}-\frac{r_0}{2}\right)\sigma_0^2+\frac{r_1}{2}\sigma_1^2 \\ . & s_0(C_1)\sigma_0^2+s_1(C_1)\sigma_1^2 & \left(\frac{C_1(u_2)}{n_0}-\frac{r_0}{2}\right)\sigma_0^2+\frac{r_1}{2}\sigma_1^2 & -r_0\sigma_0^2-r_1\sigma_1^2 \\ . & . & \left(1+\frac{s_0(C_2)}{4}\right)\sigma_0^2+\frac{s_1(C_2)}{4}\sigma_1^2 & \frac{s_0(C_1)\sigma_0^2}{2}-\frac{s_1(C_1)\sigma_1^2}{2} \\ . & . & . & s_0(C_2)\sigma_0^2+s_1(C_2)\sigma_1^2 \end{pmatrix},$$

$$\Sigma_{F^{I}_{11}(v_1,v_2,C_1,C_2)} = \begin{pmatrix} \left(1+\frac{s_1(C_1)}{4}\right)\sigma_1^2+\frac{s_0(C_1)}{4}\sigma_0^2 & \frac{s_1(C_1)\sigma_1^2}{2}-\frac{s_0(C_1)\sigma_0^2}{2} & \left(I_{\{v_1=v_2\}}+\frac{r_1}{4}-\frac{C_2(v_1)+C_1(v_2)}{2n_1}\right)\sigma_1^2+\frac{r_0}{4}\sigma_0^2 & \left(\frac{r_1}{2}-\frac{C_2(v_1)}{n_1}\right)\sigma_1^2-\frac{r_0}{2}\sigma_0^2 \\ . & s_0(C_1)\sigma_0^2+s_1(C_1)\sigma_1^2 & -\frac{r_0}{2}\sigma_0^2+\left(\frac{r_1}{2}-\frac{C_1(v_2)}{n_1}\right)\sigma_1^2 & r_0\sigma_0^2+r_1\sigma_1^2 \\ . & . & \left(1+\frac{s_1(C_2)}{4}\right)\sigma_1^2+\frac{s_0(C_2)}{4}\sigma_0^2 & -\frac{s_0(C_1)\sigma_0^2}{2}+\frac{s_1(C_1)\sigma_1^2}{2} \\ . & . & . & s_0(C_2)\sigma_0^2+s_1(C_2)\sigma_1^2 \end{pmatrix},$$

$$\Sigma_{F_{11}^{II}(v_1,v_2,C_1,C_2)} = \begin{pmatrix} \left(1+\frac{s_1(C_1)}{4}\right)\sigma_1^2+\frac{s_0(C_1)}{4}\sigma_0^2 & \frac{s_1(C_1)\sigma_1^2}{2}-\frac{s_0(C_1)\sigma_0^2}{2} & \left(\frac{C_2(v_1)+C_1(v_2)}{2n_1}-I_{\{v_1=v_2\}}-\frac{r_1}{4}\right)\sigma_1^2-\frac{r_0}{4}\sigma_0^2 & \frac{r_0}{2}\sigma_0^2+\left(\frac{C_2(v_1)}{n_1}-\frac{r_1}{2}\right)\sigma_1^2 \\ . & s_0(C_1)\sigma_0^2+s_1(C_1)\sigma_1^2 & \left(\frac{C_1(v_2)}{n_1}-\frac{r_1}{2}\right)\sigma_1^2+\frac{r_0}{2}\sigma_0^2 & -r_0\sigma_0^2-r_1\sigma_1^2 \\ . & . & \left(1+\frac{s_1(C_2)}{4}\right)\sigma_1^2+\frac{s_0(C_2)}{4}\sigma_0^2 & -\frac{s_0(C_1)\sigma_0^2}{2}+\frac{s_1(C_1)\sigma_1^2}{2} \\ . & . & . & s_0(C_2)\sigma_0^2+s_1(C_2)\sigma_1^2 \end{pmatrix},$$

$$\Sigma_{F_{01}^{I}(u_1,v_2,C_1,C_2)} = \begin{pmatrix} \left(1+\frac{s_0(C_1)}{4}\right)\sigma_0^2+\frac{s_1(C_1)}{4}\sigma_1^2 & \frac{s_0(C_1)\sigma_0^2}{2}-\frac{s_1(C_1)\sigma_1^2}{2} & \left(\frac{r_1}{4}-\frac{C_2(u_1)}{2n_0}\right)\sigma_0^2+\left(\frac{r_1}{4}-\frac{C_1(v_2)}{2n_1}\right)\sigma_1^2 & \frac{r_1}{2}\sigma_1^2+\left(\frac{C_2(u_1)}{n_0}-\frac{r_0}{2}\right)\sigma_0^2 \\ . & s_0(C_1)\sigma_0^2+s_1(C_1)\sigma_1^2 & \frac{r_0}{2}\sigma_0^2+\left(\frac{C_1(v_2)}{n_1}-\frac{r_1}{2}\right)\sigma_1^2 & -r_0\sigma_0^2-r_1\sigma_1^2 \\ . & . & \frac{s_0(C_2)}{4}\sigma_0^2+\left(1+\frac{s_1(C_2)}{4}\right)\sigma_1^2 & -\frac{s_0(C_2)\sigma_0^2}{2}+\frac{s_1(C_2)\sigma_1^2}{2} \\ . & . & . & s_0(C_2)\sigma_0^2+s_1(C_2)\sigma_1^2 \end{pmatrix},$$

$$\Sigma_{F_{01}^{II}(u_1,v_2,C_1,C_2)} = \begin{pmatrix} \left(1+\frac{s_0(C_1)}{4}\right)\sigma_0^2+\frac{s_1(C_1)}{4}\sigma_1^2 & \frac{s_0(C_1)\sigma_0^2}{2}-\frac{s_1(C_1)\sigma_1^2}{2} & -\left(\frac{r_1}{4}-\frac{C_2(u_1)}{2n_0}\right)\sigma_0^2-\left(\frac{r_1}{4}-\frac{C_1(v_2)}{2n_1}\right)\sigma_1^2 & -\frac{r_1}{2}\sigma_1^2-\left(\frac{C_2(u_1)}{n_0}-\frac{r_=}{2}\right)\sigma_0^2 \\ . & s_0(C_1)\sigma_0^2+s_1(C_1)\sigma_1^2 & -\frac{r_0}{2}\sigma_0^2-\left(\frac{C_1(v_2)}{n_1}-\frac{r_1}{2}\right)\sigma_1^2 & r_0\sigma_0^2+r_1\sigma_1^2 \\ . & . & \frac{s_0(C_2)}{4}\sigma_0^2+\left(1+\frac{s_1(C_2)}{4}\right)\sigma_1^2 & -\frac{s_0(C_2)\sigma_0^2}{2}+\frac{s_1(C_2)\sigma_1^2}{2} \\ . & . & . & s_0(C_2)\sigma_0^2+s_1(C_2)\sigma_1^2 \end{pmatrix},$$

In the following definitions for Gs, assume $C(u_1) = C(v_1) = 0$:

$$G_{00}^{I}(u_1,u_2,C) = \left[X_{u_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_1^C - \hat{\mu}_0^C, X_{u_2} - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}, \hat{\mu}_1 - \hat{\mu}_0\right]^T, \tag{3.39}$$

$$G_{00}^{II}(u_1,u_2,C) = \left[X_{u_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_1^C - \hat{\mu}_0^C, \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} - X_{u_2}, \hat{\mu}_0 - \hat{\mu}_1\right]^T, \tag{3.40}$$

$$G_{11}^{I}(v_1,v_2,C) = \left[X_{v_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_0^C - \hat{\mu}_1^C, X_{v_2} - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}, \hat{\mu}_0 - \hat{\mu}_1\right]^T, \tag{3.41}$$

$$G_{11}^{II}(v_1,v_2,C) = \left[X_{v_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_0^C - \hat{\mu}_1^C, \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} - X_{v_2}, \hat{\mu}_1 - \hat{\mu}_0\right]^T, \tag{3.42}$$

$$G_{01}^{I}(u_1,v_2,C) = \left[X_{u_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_1^C - \hat{\mu}_0^C, X_{v_2} - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}, \hat{\mu}_0 - \hat{\mu}_1\right]^T, \tag{3.43}$$

$$G_{01}^{II}(u_1,v_2,C) = \left[X_{u_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_1^C - \hat{\mu}_0^C, \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} - X_{v_2}, \hat{\mu}_1 - \hat{\mu}_0\right]^T, \tag{3.44}$$

$$G_{10}^{I}(v_1,u_2,C) = \left[X_{v_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_0^C - \hat{\mu}_1^C, X_{u_2} - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}, \hat{\mu}_1 - \hat{\mu}_0\right]^T, \tag{3.45}$$

$$G_{10}^{II}(v_1,u_2,C) = \left[X_{v_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_0^C - \hat{\mu}_1^C, \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} - X_{u_2}, \hat{\mu}_0 - \hat{\mu}_1\right]^T, \tag{3.46}$$

$$\mathrm{E}[G_{00}^{I}] = \mathrm{E}[G_{01}^{II}] = \left[\frac{\mu}{2}, -\mu, \frac{\mu}{2}, -\mu\right]^T, \qquad \mathrm{E}[G_{00}^{II}] = \mathrm{E}[G_{01}^{I}] = \left[\frac{\mu}{2}, -\mu, \frac{-\mu}{2}, \mu\right]^T,$$
$$\mathrm{E}[G_{11}^{I}] = \mathrm{E}[G_{10}^{II}] = \left[\frac{-\mu}{2}, \mu, \frac{-\mu}{2}, \mu\right]^T, \qquad \mathrm{E}[G_{11}^{II}] = \mathrm{E}[G_{10}^{I}] = \left[\frac{-\mu}{2}, \mu, \frac{\mu}{2}, -\mu\right]^T,$$

and

$$\Sigma_{G_{00}^{I}(u_1,u_2,C)} =$$
$$= \begin{pmatrix} \left(1 + \frac{s_0}{4}\right)\sigma_0^2 + \frac{s_1}{4}\sigma_1^2 & \frac{s_0\sigma_0^2}{2} - \frac{s_1\sigma_1^2}{2} & \left(I_{u_1=u_2} + \frac{1-2C(u_2)}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & s_0\sigma_0^2 + s_1\sigma_1^2 & \frac{1-2C(u_2)}{2n_0}\sigma_0^2 - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & \left(1 - \frac{3}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix},$$

$$\Sigma_{G^{II}_{00}(u_1,u_2,C)} = \\ = \begin{pmatrix} \left(1+\frac{s_0}{4}\right)\sigma_0^2+\frac{s_1}{4}\sigma_1^2 & \frac{s_0\sigma_0^2}{2}-\frac{s_1\sigma_1^2}{2} & \left(\frac{2C(u_2)-1}{4n_0}-I_{u_1=u_2}\right)\sigma_0^2-\frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0}+\frac{\sigma_1^2}{2n_1} \\ . & s_0\sigma_0^2+s_1\sigma_1^2 & \frac{2C(u_2)-1}{2n_0}\sigma_0^2+\frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0}-\frac{\sigma_1^2}{n_1} \\ . & . & \left(1-\frac{3}{4n_0}\right)\sigma_0^2+\frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0}-\frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1} \end{pmatrix},$$

$$\Sigma_{G^{I}_{11}(v_1,v_2,C)} = \\ = \begin{pmatrix} \frac{s_0}{4}\sigma_0^2+\left(1+\frac{s_1}{4}\right)\sigma_1^2 & \frac{s_1\sigma_1^2}{2}-\frac{s_0\sigma_0^2}{2} & \left(I_{v_1=v_2}+\frac{1-2C(v_2)}{4n_1}\right)\sigma_1^2+\frac{\sigma_0^2}{4n_0} & \frac{\sigma_1^2}{2n_1}+\frac{\sigma_0^2}{2n_0} \\ . & s_1\sigma_1^2+s_0\sigma_0^2 & \frac{1-2C(v_2)}{2n_1}\sigma_1^2-\frac{\sigma_0^2}{2n_0} & \frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0} \\ . & . & \left(1-\frac{3}{4n_1}\right)\sigma_1^2+\frac{\sigma_0^2}{4n_0} & -\frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} \\ . & . & . & \frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0} \end{pmatrix},$$

$$\Sigma_{G^{II}_{11}(v_1,v_2,C)} = \\ = \begin{pmatrix} \frac{s_0}{4}\sigma_0^2+\left(1+\frac{s_1}{4}\right)\sigma_1^2 & \frac{s_1\sigma_1^2}{2}-\frac{s_0\sigma_0^2}{2} & \left(\frac{2C(v_2)-1}{4n_1}-I_{v_1=v_2}\right)\sigma_1^2-\frac{\sigma_0^2}{4n_0} & -\frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} \\ . & s_1\sigma_1^2+s_0\sigma_0^2 & \frac{2C(v_2)-1}{2n_1}\sigma_1^2+\frac{\sigma_0^2}{2n_0} & -\frac{\sigma_1^2}{n_1}-\frac{\sigma_0^2}{n_0} \\ . & . & \left(1-\frac{3}{4n_1}\right)\sigma_1^2+\frac{\sigma_0^2}{4n_0} & -\frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} \\ . & . & . & \frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0} \end{pmatrix},$$

$$\Sigma_{G^{I}_{01}(u_1,v_2,C)} = \begin{pmatrix} \left(1+\frac{s_0}{4}\right)\sigma_0^2+\frac{s_1}{4}\sigma_1^2 & \frac{s_0\sigma_0^2}{2}-\frac{s_1\sigma_1^2}{2} & \frac{1-2C(v_2)}{4n_1}\sigma_1^2-\frac{\sigma_0^2}{4n_0} & -\frac{\sigma_0^2}{2n_0}-\frac{\sigma_1^2}{2n_1} \\ . & s_0\sigma_0^2+s_1\sigma_1^2 & \frac{2C(v_2)-1}{2n_1}\sigma_1^2+\frac{\sigma_0^2}{2n_0} & -\frac{\sigma_1^2}{n_1}-\frac{\sigma_0^2}{n_0} \\ . & . & \left(1-\frac{3}{4n_1}\right)\sigma_1^2+\frac{\sigma_0^2}{4n_0} & -\frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} \\ . & . & . & \frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1} \end{pmatrix},$$

$$\Sigma_{G^{II}_{01}(u_1,v_2,C)} = \begin{pmatrix} \left(1+\frac{s_0}{4}\right)\sigma_0^2 + \frac{s_1}{4}\sigma_1^2 & \frac{s_0\sigma_0^2}{2} - \frac{s_1\sigma_1^2}{2} & \frac{\sigma_0^2}{4n_0} + \frac{2C(v_2)-1}{4n_1}\sigma_1^2 & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & s_0\sigma_0^2 + s_1\sigma_1^2 & \frac{1-2C(v_2)}{2n_1}\sigma_1^2 - \frac{\sigma_0^2}{2n_0} & \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} \\ . & . & \left(1-\frac{3}{4n_1}\right)\sigma_1^2 + \frac{\sigma_0^2}{4n_0} & -\frac{\sigma_1^2}{2n_1} - \frac{\sigma_0^2}{2n_0} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix},$$

$$\Sigma_{G^{I}_{10}(v_1,u_2,C)} = \begin{pmatrix} \frac{s_0}{4}\sigma_0^2 + \left(1+\frac{s_1}{4}\right)\sigma_1^2 & \frac{s_1\sigma_1^2}{2} - \frac{s_0\sigma_0^2}{2} & \frac{1-2C(u_2)}{4n_0}\sigma_0^2 - \frac{\sigma_1^2}{4n_1} & \frac{\sigma_1^2}{2n_1} + \frac{\sigma_0^2}{2n_0} \\ . & s_1\sigma_1^2 + s_0\sigma_0^2 & \frac{2C(u_2)-1}{2n_0}\sigma_0^2 + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0} \\ . & . & \left(1-\frac{3}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix},$$

$$\Sigma_{G^{II}_{10}(v_1,u_2,C)} = \begin{pmatrix} \frac{s_0}{4}\sigma_0^2 + \left(1+\frac{s_1}{4}\right)\sigma_1^2 & \frac{s_1\sigma_1^2}{2} - \frac{s_0\sigma_0^2}{2} & \frac{2C(u_2)-1}{4n_0}\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_1^2}{2n_1} - \frac{\sigma_0^2}{2n_0} \\ . & s_1\sigma_1^2 + s_0\sigma_0^2 & \frac{1-2C(u_2)}{2n_0}\sigma_0^2 - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} \\ . & . & \left(1-\frac{3}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}.$$

In the following definitions for Ks, assume $C(u_1) = C(v_1) = 0$, again $X^* \in \Pi_0$, $X^{**} \in \Pi_1$ are independent of S_n:

$$K^I_{00}(C) = \left[X_{u_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_1^C - \hat{\mu}_0^C, X^* - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}, \hat{\mu}_1 - \hat{\mu}_0\right]^T, \tag{3.47}$$

$$K^{II}_{00}(C) = \left[X_{u_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_1^C - \hat{\mu}_0^C, \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} - X^*, \hat{\mu}_0 - \hat{\mu}_1\right]^T, \tag{3.48}$$

$$K^I_{11}(C) = \left[X_{v_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_0^C - \hat{\mu}_1^C, X^{**} - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}, \hat{\mu}_0 - \hat{\mu}_1\right]^T, \tag{3.49}$$

$$K^{II}_{11}(C) = \left[X_{v_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_0^C - \hat{\mu}_1^C, \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} - X^{**}, \hat{\mu}_1 - \hat{\mu}_0\right]^T, \tag{3.50}$$

$$K_{01}^{I}(C) = \left[X_{u_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_1^C - \hat{\mu}_0^C, X^{**} - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}, \hat{\mu}_0 - \hat{\mu}_1\right]^T, \tag{3.51}$$

$$K_{01}^{II}(C) = \left[X_{u_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_1^C - \hat{\mu}_0^C, \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} - X^{**}, \hat{\mu}_1 - \hat{\mu}_0\right]^T, \tag{3.52}$$

$$K_{10}^{I}(C) = \left[X_{v_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_0^C - \hat{\mu}_1^C, X^{*} - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}, \hat{\mu}_1 - \hat{\mu}_0\right]^T, \tag{3.53}$$

$$K_{10}^{II}(C) = \left[X_{v_1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}, \hat{\mu}_0^C - \hat{\mu}_1^C, \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} - X^{*}, \hat{\mu}_0 - \hat{\mu}_1\right]^T, \tag{3.54}$$

$$\begin{aligned}
\mathrm{E}[K_{00}^{I}(C)] = \mathrm{E}[K_{01}^{II}(C)] &= \left[\frac{\mu}{2}, -\mu, \frac{\mu}{2}, -\mu\right]^T, \\
\mathrm{E}[K_{00}^{II}(C)] = \mathrm{E}[K_{01}^{I}(C)] &= \left[\frac{\mu}{2}, -\mu, \frac{-\mu}{2}, \mu\right]^T, \\
\mathrm{E}[K_{11}^{I}(C)] = \mathrm{E}[K_{10}^{II}(C)] &= \left[\frac{-\mu}{2}, \mu, \frac{-\mu}{2}, \mu\right]^T, \\
\mathrm{E}[K_{11}^{II}(C)] = \mathrm{E}[K_{10}^{I}(C)] &= \left[\frac{-\mu}{2}, \mu, \frac{\mu}{2}, -\mu\right]^T,
\end{aligned}$$

$$\Sigma_{K_{00}^{I}(C)} = \begin{pmatrix} \left(1 + \frac{s_0}{4}\right)\sigma_0^2 + \frac{s_1}{4}\sigma_1^2 & \frac{s_0\sigma_0^2}{2} - \frac{s_1\sigma_1^2}{2} & \frac{\sigma_1^2}{4n_1} - \frac{\sigma_0^2}{4n_0} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & s_0\sigma_0^2 + s_1\sigma_1^2 & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & \left(1 + \frac{1}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix},$$

$$\Sigma_{K_{00}^{II}(C)} = \begin{pmatrix} \left(1 + \frac{s_0}{4}\right)\sigma_0^2 + \frac{s_1}{4}\sigma_1^2 & \frac{s_0\sigma_0^2}{2} - \frac{s_1\sigma_1^2}{2} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & \frac{\sigma_1^2}{2n_1} + \frac{\sigma_0^2}{2n_0} \\ . & s_0\sigma_0^2 + s_1\sigma_1^2 & \frac{\sigma_1^2}{2n_1} - \frac{\sigma_0^2}{2n_0} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & \left(1 + \frac{1}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix},$$

$$\Sigma_{K_{11}^{I}(C)}=\begin{pmatrix} \left(1+\frac{s_1}{4}\right)\sigma_1^2+\frac{s_0}{4}\sigma_0^2 & \frac{s_1\sigma_1^2}{2}-\frac{s_0\sigma_0^2}{2} & \frac{\sigma_0^2}{4n_0}-\frac{\sigma_1^2}{4n_1} & -\frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} \\ . & s_1\sigma_1^2+s_0\sigma_0^2 & \frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} & \frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0} \\ . & . & \left(1+\frac{1}{4n_1}\right)\sigma_1^2+\frac{\sigma_0^2}{4n_0} & \frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} \\ . & . & . & \frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0} \end{pmatrix},$$

$$\Sigma_{K_{11}^{II}(C)}=\begin{pmatrix} \left(1+\frac{s_1}{4}\right)\sigma_1^2+\frac{s_0}{4}\sigma_0^2 & \frac{s_1\sigma_1^2}{2}-\frac{s_0\sigma_0^2}{2} & \frac{\sigma_1^2}{4n_1}-\frac{\sigma_0^2}{4n_0} & \frac{\sigma_0^2}{2n_0}+\frac{\sigma_1^2}{2n_1} \\ . & s_1\sigma_1^2+s_0\sigma_0^2 & \frac{\sigma_0^2}{2n_0}-\frac{\sigma_1^2}{2n_1} & -\frac{\sigma_1^2}{n_1}-\frac{\sigma_0^2}{n_0} \\ . & . & \left(1+\frac{1}{4n_1}\right)\sigma_1^2+\frac{\sigma_0^2}{4n_0} & \frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} \\ . & . & . & \frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0} \end{pmatrix},$$

$$\Sigma_{K_{01}^{I}(C)}=\begin{pmatrix} \left(1+\frac{s_0}{4}\right)\sigma_0^2+\frac{s_1}{4}\sigma_1^2 & \frac{s_0\sigma_0^2}{2}-\frac{s_1\sigma_1^2}{2} & \frac{\sigma_1^2}{4n_1}-\frac{\sigma_0^2}{4n_0} & -\frac{\sigma_0^2}{2n_0}-\frac{\sigma_1^2}{2n_1} \\ . & s_1\sigma_1^2+s_0\sigma_0^2 & \frac{\sigma_0^2}{2n_0}-\frac{\sigma_1^2}{2n_1} & -\frac{\sigma_1^2}{n_1}-\frac{\sigma_0^2}{n_0} \\ . & . & \left(1+\frac{1}{4n_1}\right)\sigma_1^2+\frac{\sigma_0^2}{4n_0} & \frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} \\ . & . & . & \frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0} \end{pmatrix},$$

$$\Sigma_{K_{01}^{II}(C)}=\begin{pmatrix} \left(1+\frac{s_0}{4}\right)\sigma_0^2+\frac{s_1}{4}\sigma_1^2 & \frac{s_0\sigma_0^2}{2}-\frac{s_1\sigma_1^2}{2} & \frac{\sigma_0^2}{4n_0}-\frac{\sigma_1^2}{4n_1} & \frac{\sigma_1^2}{2n_1}+\frac{\sigma_0^2}{2n_0} \\ . & s_1\sigma_1^2+s_0\sigma_0^2 & \frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} & \frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0} \\ . & . & \left(1+\frac{1}{4n_1}\right)\sigma_1^2+\frac{\sigma_0^2}{4n_0} & \frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} \\ . & . & . & \frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0} \end{pmatrix},$$

$$\Sigma_{K_{10}^{I}(C)}=\begin{pmatrix} \left(1+\frac{s_1}{4}\right)\sigma_1^2+\frac{s_0}{4}\sigma_0^2 & \frac{s_1\sigma_1^2}{2}-\frac{s_0\sigma_0^2}{2n_0} & \frac{\sigma_0^2}{4n_0}-\frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0}+\frac{\sigma_1^2}{2n_1} \\ . & s_1\sigma_1^2+s_0\sigma_0^2 & \frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0} & -\frac{\sigma_1^2}{n_1}-\frac{\sigma_0^2}{n_0} \\ . & . & \left(1+\frac{1}{4n_0}\right)\sigma_0^2+\frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0}-\frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1} \end{pmatrix},$$

$$\Sigma_{K_{10}^{II}(C)} = \begin{pmatrix} \left(1+\frac{s_1}{4}\right)\sigma_1^2 + \frac{s_0}{4}\sigma_0^2 & \frac{s_1\sigma_1^2}{2} - \frac{s_0\sigma_0^2}{2n_0} & \frac{\sigma_1^2}{4n_1} - \frac{\sigma_0^2}{4n_0} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & s_1\sigma_1^2 + s_0\sigma_0^2 & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} \\ . & . & \left(1+\frac{1}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}.$$

$$H_{00} = \left[X_{u_1} - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}, \hat{\mu}_1 - \hat{\mu}_0, X_{u_2} - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right]^T, \tag{3.55}$$

$$H_{11} = \left[X_{v_1} - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}, \hat{\mu}_0 - \hat{\mu}_1, X_{v_2} - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right]^T, \tag{3.56}$$

$$H_{01} = \left[X_{u_1} - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}, \hat{\mu}_1 - \hat{\mu}_0, \frac{\hat{\mu}_0+\hat{\mu}_1}{2} - X_{v_1}\right]^T, \tag{3.57}$$

$$J_{00} = \left[X_{u_1} - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}, \hat{\mu}_1 - \hat{\mu}_0, X^* - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right]^T, \tag{3.58}$$

$$J_{11} = \left[X_{v_1} - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}, \hat{\mu}_0 - \hat{\mu}_1, X^{**} - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right]^T, \tag{3.59}$$

$$J_{01} = \left[X_{u_1} - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}, \hat{\mu}_1 - \hat{\mu}_0, \frac{\hat{\mu}_0+\hat{\mu}_1}{2} - X^{**}\right]^T, \tag{3.60}$$

$$J_{10} = \left[X_{v_1} - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}, \hat{\mu}_0 - \hat{\mu}_1, X^* - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right]^T, \tag{3.61}$$

Basic algebra gives us the mean vectors and the covariance matrices as following.

$$\mathrm{E}[H_{00}] = \left[\frac{\mu}{2}, -\mu, \frac{\mu}{2}\right] \qquad \mathrm{E}[H_{11}] = \left[\frac{-\mu}{2}, \mu, \frac{-\mu}{2}\right] \qquad \mathrm{E}[H_{01}] = \left[\frac{\mu}{2}, -\mu, \frac{-\mu}{2}\right],$$

$$\mathrm{E}[J_{00}] = \left[\frac{\mu}{2}, -\mu, \frac{\mu}{2}\right] \quad \mathrm{E}[J_{11}] = \left[\frac{-\mu}{2}, \mu, \frac{-\mu}{2}\right] \quad \mathrm{E}[J_{01}] = \left[\frac{\mu}{2}, -\mu, \frac{-\mu}{2}\right] \quad \mathrm{E}[J_{10}] = \left[\frac{-\mu}{2}, \mu, \frac{\mu}{2}\right].$$

$$\Sigma_{H_{00}} = \begin{pmatrix} \left(1-\frac{3}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_1^2}{4n_1} - \frac{3\sigma_0^2}{4n_0} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & \left(1-\frac{3}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} \end{pmatrix},$$

$$\Sigma_{H_{11}} = \begin{pmatrix} \left(1-\frac{3}{4n_1}\right)\sigma_1^2 + \frac{\sigma_0^2}{4n_0} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{3}{4n_1}\sigma_1^2 \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & \left(1-\frac{3}{4n_1}\right)\sigma_1^2 + \frac{\sigma_0^2}{4n_0} \end{pmatrix},$$

$$\Sigma_{H_{01}} = \begin{pmatrix} \left(1 - \frac{3}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & \left(1 - \frac{3}{4n_1}\right)\sigma_1^2 + \frac{\sigma_0^2}{4n_0} \end{pmatrix},$$

$$\Sigma_{J_{00}} = \begin{pmatrix} \left(1 - \frac{3}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_1^2}{4n_1} - \frac{\sigma_0^2}{4n_0} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & \left(1 + \frac{1}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} \end{pmatrix},$$

$$\Sigma_{J_{11}} = \begin{pmatrix} \left(1 - \frac{3}{4n_1}\right)\sigma_1^2 + \frac{\sigma_1^2}{4n_0} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_1^2}{2n_1} - \frac{\sigma_0^2}{2n_0} \\ . & . & \left(1 + \frac{1}{4n_1}\right)\sigma_1^2 + \frac{\sigma_1^2}{4n_0} \end{pmatrix},$$

$$\Sigma_{J_{01}} = \begin{pmatrix} \left(1 - \frac{3}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_1^2}{2n_1} - \frac{\sigma_0^2}{2n_0} \\ . & . & \left(1 + \frac{1}{4n_1}\right)\sigma_1^2 + \frac{\sigma_0^2}{4n_0} \end{pmatrix},$$

$$\Sigma_{J_{10}} = \begin{pmatrix} \left(1 - \frac{3}{4n_1}\right)\sigma_1^2 + \frac{\sigma_0^2}{4n_0} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_1^2}{2n_1} - \frac{\sigma_0^2}{2n_0} \\ . & . & \left(1 + \frac{1}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} \end{pmatrix},$$

D. The C-bootstrap Linear Classifier

This section presents the theorems to compute the first and the second moment of the hold-out error $\hat{\varepsilon}_C$ of the C-bootstrap linear classifier $\psi_C(X)$ and it cross moment with the resubstitution estimator and the true error of the original linear classifier $\psi(X)$.

Theorem 1 *Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0 + 1, \ldots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (3.31). Then we* *have:*

$$\begin{aligned}\mathrm{E}[\hat{\varepsilon}_C] = \frac{1}{m}\Big(&m_0(C)\mathrm{P}\{B_0(C) \geq 0\} + m_0\mathrm{P}\{B_0(C) < 0\} + \\ &+ m_1\mathrm{P}\{B_1(C) \geq 0\} + m_1\mathrm{P}\{B_1(C) < 0\}\Big),\end{aligned} \tag{3.62}$$

where $B_0(C)$ and $B_1(C)$ are bivariate Gaussian random vectors with the following means and covariance matrices:

$$\mathrm{E}\left[B_0(C)\right] = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \end{bmatrix}, \; \Sigma_{B_0(C)} = \begin{pmatrix} \left(1+\frac{s_0}{4}\right)\sigma_0^2 + \frac{s_1}{4}\sigma_1^2 & \frac{s_0}{2}\sigma_0^2 - \frac{s_1}{2}\sigma_1^2 \\ . & s_0\sigma_0^2 + s_1\sigma_1^2 \end{pmatrix}, \tag{3.63}$$

$$\mathrm{E}\left[B_1(C)\right] = \begin{bmatrix} \frac{-\mu}{2} \\ \mu \end{bmatrix}, \; \Sigma_{B_1(C)} = \begin{pmatrix} \frac{s_0}{4}\sigma_0^2 + \left(1+\frac{s_1}{4}\right)\sigma_1^2 & \frac{s_1}{2}\sigma_0^2 - \frac{s_0}{2}\sigma_1^2 \\ . & s_0\sigma_0^2 + s_1\sigma_1^2 \end{pmatrix}, \tag{3.64}$$

where $m, m_i, s, s_i, (i = 0, 1)$ are defined as in (3.11) and (3.12), respectively, and $\mu = \mu_0 - \mu_1$.

Proof: See appendix.

Theorem 2 *Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0 + 1, \ldots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (3.31). Then given a bootstrap weight vector C, we have:*

$$\begin{aligned}\mathrm{E}\left[\hat{\varepsilon}_C^2\right] = \frac{1}{m^2}\Big[&m_0\Big(\mathrm{P}\{B_0(C) \geq 0\} + \mathrm{P}\{B_0(C) < 0\}\Big) + m_1\Big(\mathrm{P}\{B_1(C) \geq 0\} + \mathrm{P}\{B_1(C) < 0\}\Big) + \\ &+ \left(m_0^2 - m_0\right)\Big(\mathrm{P}\{T_{00}(C) \geq 0\} + \mathrm{P}\{T_{00}(C) < 0\}\Big) + \\ &+ 2m_0m_1\Big(\mathrm{P}\{T_{01}(C) \geq 0\} + \mathrm{P}\{T_{01}(C) < 0\}\Big) + \\ &+ \left(m_1^2 - m_1\right)\Big(\mathrm{P}\{T_{11}(C) \geq 0\} + \mathrm{P}\{T_{11}(C) < 0\}\Big)\Big],\end{aligned} \tag{3.65}$$

where $B_0(C)$ and $B_1(C)$ are bivariate Gaussian random vectors defined as in (3.63) and (3.64), respectively. $T_{00}(C)$, $T_{11}(C)$, and $T_{01}(C)$ are trivariate Gaussian vectors with the *following means and covariance matrices:*

$$\mathrm{E}\left[T_{00}(C)\right] = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \end{bmatrix}, \; \Sigma_{T_{00}(C)} = \begin{pmatrix} \left(1+\frac{s_0}{4}\right)\sigma_0^2 + \frac{s_1}{4}\sigma_1^2 & \frac{s_0\sigma_0^2}{2} - \frac{s_1\sigma_1^2}{2} & \frac{s_0}{4}\sigma_0^2 + \frac{s_1}{4}\sigma_1^2 \\ . & s_0\sigma_0^2 + s_1\sigma_1^2 & \frac{s_0\sigma_0^2}{2} - \frac{s_1\sigma_1^2}{2} \\ . & . & \left(1+\frac{s_0}{4}\right)\sigma_0^2 + \frac{s_1}{4}\sigma_1^2 \end{pmatrix},$$

$$\mathrm{E}\left[T_{11}(C)\right] = \begin{bmatrix} \frac{-\mu}{2} \\ \mu \\ \frac{-\mu}{2} \end{bmatrix}, \; \Sigma_{T_{11}(C)} = \begin{pmatrix} \frac{s_0}{4}\sigma_0^2 + \left(1+\frac{s_1}{4}\right)\sigma_1^2 & -\frac{s_0\sigma_0^2}{2} + \frac{s_1\sigma_1^2}{2} & \frac{s_0}{4}\sigma_0^2 + \frac{s_1}{4}\sigma_1^2 \\ . & s_0\sigma_0^2 + s_1\sigma_1^2 & -\frac{s_0\sigma_0^2}{2} + \frac{s_1\sigma_1^2}{2} \\ . & . & \frac{s_0}{4}\sigma_0^2 + \left(1+\frac{s_1}{4}\right)\sigma_1^2 \end{pmatrix},$$

$$\mathrm{E}\left[T_{01}(C)\right] = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \end{bmatrix}, \; \Sigma_{T_{01}(C)} = \begin{pmatrix} \left(1+\frac{s_0}{4}\right)\sigma_0^2 + \frac{s_1}{4}\sigma_1^2 & \frac{s_0\sigma_0^2}{2} - \frac{s_1\sigma_1^2}{2} & -\frac{s_1}{4}\sigma_1^2 - \frac{s_0}{4}\sigma_0^2 \\ . & s_0\sigma_0^2 + s_1\sigma_1^2 & -\frac{s_0\sigma_1^2}{2} + \frac{s_1\sigma_1^2}{2} \\ . & . & \frac{s_0}{4}\sigma_0^2 + \left(1+\frac{s_1}{4}\right)\sigma_1^2 \end{pmatrix},$$

where s_0 and s_1 are defined as in (3.12), m as in (3.11), and $\mu = \mu_0 - \mu_1$.

Proof: See appendix.

Theorem 3 *Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0+1, \ldots, n_0+n_1$ be a set of $n = n_0+n_1$ i.i.d. observations used to derive the classifier in (3.31). Then given two distinct bootstrap vectors C_1 and C_2, we have:*

$$\begin{aligned} &\mathrm{E}\left[\hat{\varepsilon}_{C_1}\hat{\varepsilon}_{C_2}\right] = \\ &= \lambda(C_1,C_2)\Big(\sum_{i,j=1}^{n_0} \mathrm{I}_{C_1(i)=0,C_2(j)=0}\mathrm{F}_{00}(i,j,C_1,C_2) + \sum_{i,j=n_0+1}^{n_0+n_1} \mathrm{I}_{C_1(i)=0,C_2(j)=0}\mathrm{F}_{11}(i,j,C_1,C_2) + \\ &\quad + \sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C_1(i)=0,C_2(j)=0}\mathrm{F}_{01}(i,j,C_1,C_2) + \sum_{i=n_0+1}^{n_0+n_1}\sum_{j=1}^{n_0} \mathrm{I}_{C_1(i)=0,C_2(j)=0}\mathrm{F}_{01}(j,i,C_2,C_1)\Big), \end{aligned} \tag{3.66}$$

where $\lambda(C_1,C_2)=\frac{1}{m(C_1)m(C_2)}$, *and*

$$\begin{aligned}\mathrm{F}_{ab}(i,j,C_1,C_2) = \mathrm{P}\{F^{I}_{ab}(i,j,C_1,C_2)>0\}+\mathrm{P}\{F^{I}_{ab}(i,j,C_1,C_2)<0\}+\\ +\mathrm{P}\{F^{II}_{ab}(i,j,C_1,C_2)>0\}+\mathrm{P}\{F^{II}_{ab}(i,j,C_1,C_2)<0\},\end{aligned}$$

where $F^{I}_{ab}(i,j,C_1,C_2),F^{II}_{ab}(i,j,C_1,C_2)$, $a,b=0,1$ *are 4-dimensional Gaussian random vectors defined as in (3.33), (3.34), (3.35), (3.36), (3.37), and (3.38), respectively, and m as in (3.11).*

Proof: See appendix.

Theorem 4 *Let* $X_i \sim N(\mu_0,\sigma_0^2)$ *for* $i=1,\ldots,n_0$, *and* $X_i \sim N(\mu_1,\sigma_1^2)$ *for* $i=n_0+1,\ldots,n_0+n_1$ *be a set of* $n=n_0+n_1$ *i.i.d. observations used to derive the classifier in (3.31). Then given a bootstrap vector C, we have:*

$$\begin{aligned}\mathrm{E}\,[\hat{\varepsilon}_C\hat{\varepsilon}_r] = \frac{1}{nm}\Big(\sum_{i,j=1}^{n_0}\mathrm{I}_{C(i)=0}\mathrm{G}_{00}(i,j,C)+\sum_{i,j=n_0+1}^{n_0+n_1}\mathrm{I}_{C(i)=0}\mathrm{G}_{11}(i,j,C)+\\ +\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n_0+n_1}\mathrm{I}_{C(i)=0}\mathrm{G}_{01}(i,j,C)+\sum_{i=n_0+1}^{n_0+n_1}\sum_{j=1}^{n_0}\mathrm{I}_{C(i)=0}\mathrm{G}_{10}(i,j,C)\Big),\end{aligned} \tag{3.67}$$

where m is defined as in (3.11), and

$$\begin{aligned}\mathrm{G}_{ab} = \mathrm{P}\{G^{I}_{ab}(i,j,C)>0\}+\mathrm{P}\{G^{I}_{ab}(i,j,C)<0\}+\\ +\mathrm{P}\{G^{II}_{ab}(i,j,C)>0\}+\mathrm{P}\{G^{II}_{ab}(i,j,C)<0\},\end{aligned}$$

where $G^{I}_{ab}(i,j,C)$, $G^{II}_{ab}(i,j,C)$, a, $b=0$, 1 *are 4-dimensional Gaussian random vectors defined as in (3.39), (3.40), (3.41), (3.42), (3.43), (3.44), (3.45), and (3.46), respectively.*

Proof: See appendix.

Theorem 5 *Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \dots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0+1, \dots, n_0+n_1$ be a set of $n = n_0+n_1$ i.i.d. observations used to derive the classifier in (3.31). Then given a bootstrap vector C, we have:*

$$\mathrm{E}\left[\hat{\varepsilon}_C \varepsilon\right] = \frac{m_0(1-\gamma)}{m}\mathrm{K}_{00}(C) + \frac{m_1\gamma}{m}\mathrm{K}_{11}(C) + \frac{m_0\gamma}{m}\mathrm{K}_{01}(C) + \frac{m_1(1-\gamma)}{m}\mathrm{K}_{10}(C), \tag{3.68}$$

where m, m_0, and m_1 are defined as in (3.11), and

$$\mathrm{K}_{ab}(C) = \mathrm{P}\{K_{ab}^{I}(C) < 0\} + \mathrm{P}\{K_{ab}^{I}(C) > 0\} + \mathrm{P}\{K_{ab}^{II}(C) < 0\} + \mathrm{P}\{K_{ab}^{II}(C) > 0\},$$

where $K_{ab}^{I}(C)$, $K_{ab}^{II}(C)$, $a, b = 0, 1$ are 4-dimensional Gaussian random vectors defined as in (3.47), (3.48), (3.49), (3.50), (3.51), (3.52), (3.53), and (3.54), respectively.

Proof: See appendix.

E. The Zero Bootstrap Error Estimation

The followings present the theorems to compute the first and second moments of zero bootstrap estimator and its correlation with the true error as well as the resubstitution estimator.

Theorem 6 *Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \dots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0+1, \dots, n_0+n_1$ be a set of $n = n_0+n_1$ i.i.d. observations used to derive the classifier in (3.31). Then we have:*

$$\begin{aligned}\mathrm{E}\left[\hat{\varepsilon}_0\right] = \sum_C \frac{\mathrm{P}(C)}{m(C)}\Big(&m_0(C)\mathrm{P}\{B_0(C) \geq 0\} + m_0(C)\mathrm{P}\{B_0(C) < 0\} + \\ &+ m_1(C)\mathrm{P}\{B_1(C) \geq 0\} + m_1(C)\mathrm{P}\{B_1(C) < 0\}\Big),\end{aligned} \tag{3.69}$$

where $B_0(C)$ and $B_1(C)$ are defined as in Theorem 1, m, m_0, and m_1 as in (3.11), and $\mathrm{P}(C)$ as in (2.24).

Proof: This is the immediate result of Theorem 1 and (3.23).

Theorem 7 *Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0+1, \ldots, n_0+n_1$ be a set of $n = n_0+n_1$ i.i.d. observations used to derive the classifier in (3.31). Then we have:*

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_0^2] = {} & \sum_C \lambda_2(C) \Big[m_0(C)\mathrm{P}\{B_0(C) \geq 0\} + m_0(C)\mathrm{P}\{B_0(C) < 0\} + m_1(C)\mathrm{P}\{B_1(C) \geq 0\} + \\
& + m_1(C)\mathrm{P}\{B_1(C) < 0\} + \big(m_0^2(C) - m_0(C)\big)\Big(\mathrm{P}\{T_{00}(C) \geq 0\} + \mathrm{P}\{T_{00}(C) < 0\}\Big\} + \\
& + \big(m_1^2(C) - m_1(C)\big)\Big(\mathrm{P}\{T_{11}(C) \geq 0\} + \mathrm{P}\{T_{11}(C) < 0\}\Big) + \\
& + 2m_0(C)m_1(C)\Big(\mathrm{P}\{T_{01}(C) \geq 0\} + \mathrm{P}\{T_{01}(C) < 0\}\Big)\Big] + \\
& + \sum_{C_1 \neq C_2} \lambda_3(C_1, C_2) \Bigg[\sum_{i,j=1}^{n_0} \mathrm{I}_{C_1(i)=0, C_2(j)=0} \mathrm{F}_{00}(i, j, C_1, C_2) + \\
& + \sum_{i,j=n_0+1}^{n_0+n_1} \mathrm{I}_{C_1(i)=0, C_2(j)=0} \mathrm{F}_{11}(i, j, C_1, C_2) + \sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C_1(i)=0, C_2(j)=0} \mathrm{F}_{01}(i, j, C_1, C_2) + \\
& + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j=1}^{n_0} \mathrm{I}_{C_1(i)=0, C_2(j)=0} \mathrm{F}_{01}(j, i, C_2, C_1) \Bigg],
\end{aligned} \tag{3.70}$$

where $\lambda_2(C) = \frac{\mathrm{P}(C)}{m^2(C)}$, $\lambda_3(C_1, C_2) = \frac{2\mathrm{P}(C_1)\mathrm{P}(C_2)}{m(C_1)m(C_2)}$, $B_0(C)$ and $B_1(C)$ are defined as in Theorem 1, $T_{ab}(C)$ as in Theorem 2, $\mathrm{F}_{ab}(i, j, C_1, C_2)$ as in Theorem 3, $a, b = 0, 1$; m, m_0, m_1 as in (3.11), and $\mathrm{P}(C)$ as in (2.24).

Proof: This is the immediate result of theorem 2, 3, and (3.24).

Theorem 8 *Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0+1, \ldots, n_0+n_1$ be a set of $n = n_0+n_1$ i.i.d. observations used to derive the classifier in (3.31). Then we* *have:*

$$\begin{aligned}\mathrm{E}\left[\hat{\varepsilon}_0\hat{\varepsilon}_r\right] = \sum_C \frac{\mathrm{P}(C)}{nm(C)}\Bigg[&\sum_{i,j=1}^{n_0} \mathrm{I}_{C(i)=0}\mathrm{G}_{00}(i,j,C) + \sum_{i,j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\mathrm{G}_{11}(i,j,C) + \\ &+ \sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\mathrm{G}_{01}(i,j,C) + \sum_{i=n_0+1}^{n_0+n_1}\sum_{j=1}^{n_0} \mathrm{I}_{C(i)=0}\mathrm{G}_{10}(i,j,C)\Bigg],\end{aligned} \tag{3.71}$$

where $\mathrm{G}_{ab}(i,j,C), a, b = 0, 1$ *are defined as in Theorem 4, m as in (3.11), and* $\mathrm{P}(C)$ *as in (2.24).*

Proof: This is the immediate result of theorem 3 and (3.25).

Theorem 9 *Let* $X_i \sim N(\mu_0, \sigma_0^2)$ *for* $i = 1, \dots, n_0$*, and* $X_i \sim N(\mu_1, \sigma_1^2)$ *for* $i = n_0+1, \dots, n_0+n_1$ *be a set of* $n = n_0 + n_1$ *i.i.d. observations used to derive the classifier in (3.31). Then we have:*

$$\begin{aligned}&\mathrm{E}[\hat{\varepsilon}_0\varepsilon] = \\ &= \sum_C \frac{\mathrm{P}(C)}{m(C)}\Big[(1-\gamma)\Big(m_0(C)\mathrm{K}_{00}(C) + m_1(C)\mathrm{K}_{10}(C)\Big) + \gamma\Big(m_1(C)\mathrm{K}_{11}(C) + m_0(C)\mathrm{K}_{01}(C)\Big)\Big],\end{aligned} \tag{3.72}$$

where $\mathrm{K}_{ab}(C), a, b = 0, 1$ *are defined as in theorem 5; m,* m_0*, and* m_1 *as in (3.11), and* $\mathrm{P}(C)$ *as in (2.24).*

Proof: This is the immediate result of theorem 5 and (3.26).

F. The Convex Bootstrap Error Estimation

We first rewrite in our notations the moments $\mathrm{E}[\varepsilon]$, $\mathrm{E}[\varepsilon^2]$, $\mathrm{E}[\hat{\varepsilon}_r]$, $\mathrm{E}[\hat{\varepsilon}_r^2]$, and $\mathrm{E}[\varepsilon\hat{\varepsilon}_r]$ which were were derived for the univariate model in [86]. This section then presents theorems to compute the first and second moments of the convex bootstrap estimator with arbitrary scalar w (3.10) and its correlation with the true error.

1. The Moments of the True Error

a. The First Moment

Under univariate Gaussian model, (3.20) becomes

$$\mathrm{E}[\varepsilon] = (1-\gamma)\Big(\mathrm{P}\{B_0(\overrightarrow{1}) > 0\} + \mathrm{P}\{B_0(\overrightarrow{1}) < 0\}\Big) + \gamma\Big(\mathrm{P}\{B_1(\overrightarrow{1}) > 0\} + \mathrm{P}\{B_1(\overrightarrow{1}) < 0\}\Big), \tag{3.73}$$

where B_0 and B_1 are defined in Theorem 1.

b. The Second Moment

Under univariate Gaussian model, (3.21) becomes

$$\mathrm{E}[\varepsilon^2] = (1-\gamma)^2\Big(\mathrm{P}\{R_{00} \geq 0\} + \mathrm{P}\{R_{00} < 0\}\Big) + 2\gamma(1-\gamma)\Big(\mathrm{P}\{R_{01} \geq 0\} + \mathrm{P}\{R_{01} < 0\}\Big) + \\ + \gamma^2\Big(\mathrm{P}\{R_{11} \geq 0\} + \mathrm{P}\{R_{11} < 0\}\Big), \tag{3.74}$$

where R_{00}, R_{11}, and R_{01} are trivariate Gaussian random variables with the means and covariance matrices as followings

$$\mathrm{E}[R_{00}] = \begin{bmatrix} \frac{-\mu}{2} \\ \mu \\ \frac{-\mu}{2} \end{bmatrix}, \Sigma_{R_{00}} = \begin{pmatrix} \left(1+\frac{1}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_1^2}{4n_1} + \frac{\sigma_0^2}{4n_0} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & \left(1+\frac{1}{4n_0}\right)\sigma_0^2 + \frac{\sigma_1^2}{4n_1} \end{pmatrix}, \tag{3.75}$$

$$\mathrm{E}[R_{11}] = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \end{bmatrix}, \Sigma_{R_{11}} = \begin{pmatrix} \left(1+\frac{1}{4n_1}\right)\sigma_1^2 + \frac{\sigma_0^2}{4n_0} & \frac{\sigma_1^2}{2n_1} - \frac{\sigma_0^2}{2n_0} & \frac{\sigma_1^2}{4n_1} + \frac{\sigma_0^2}{4n_0} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_1^2}{2n_1} - \frac{\sigma_0^2}{2n_0} \\ . & . & \left(1+\frac{1}{4n_1}\right)\sigma_1^2 + \frac{\sigma_0^2}{4n_0} \end{pmatrix}, \tag{3.76}$$

$$\mathrm{E}[R_{01}]=\begin{bmatrix}\frac{-\mu}{2}\\ \mu\\ \frac{-\mu}{2}\end{bmatrix},\Sigma_{R_{01}}=\begin{pmatrix}\left(1+\frac{1}{4n_0}\right)\sigma_0^2+\frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0}-\frac{\sigma_1^2}{2n_1} & -\frac{\sigma_1^2}{4n_1}-\frac{\sigma_0^2}{4n_0}\\ . & \frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1} & \frac{\sigma_1^2}{2n_1}-\frac{\sigma_0^2}{2n_0}\\ . & . & \left(1+\frac{1}{4n_1}\right)\sigma_1^2+\frac{\sigma_0^2}{4n_0}\end{pmatrix}, \quad (3.77)$$

2. The Moments of the Resubstitution Estimator

a. The First Moment

$$\begin{aligned}\mathrm{E}[\hat{\varepsilon}_r]&=\mathrm{E}\left[\frac{1}{n}\left(\sum_{i=1}^{n_0}I_{\psi(X_i)=1}+\sum_{i=n_0+1}^{n_0+n_1}I_{\psi(X_i)=0}\right)\right]\\ &=\frac{n_0}{n}\mathrm{P}\{\psi(X_1)=1|\,X\in\Pi_0\}+\frac{n_1}{n}\mathrm{P}\{\psi(X_{n_0+1})=0|\,X\in\Pi_1\}\\ &=\frac{n_0}{n}\Big(\mathrm{P}\{D_0\geq 0\}+\mathrm{P}\{D_0<0\}\Big)+\frac{n_1}{n}\Big(\mathrm{P}\{D_1\geq 0\}+\mathrm{P}\{D_1<0\}\Big),\end{aligned} \quad (3.78)$$

where D_0, D_1 are bivariate Gaussian vectors with the following means and covariance matrices:

$$\mathrm{E}[D_0]=\begin{bmatrix}\frac{\mu}{2}\\ -\mu\end{bmatrix},\ \Sigma_{D_0}=\begin{pmatrix}\left(1-\frac{3}{4n_0}\right)\sigma_0^2+\frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0}-\frac{\sigma_1^2}{2n_1}\\ . & \frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}\end{pmatrix}, \quad (3.79)$$

$$\mathrm{E}[D_1]=\begin{bmatrix}\frac{-\mu}{2}\\ \mu\end{bmatrix},\ \Sigma_{D_1}=\begin{pmatrix}\left(1-\frac{3}{4n_1}\right)\sigma_1^2+\frac{\sigma_0^2}{4n_0} & -\frac{\sigma_0^2}{2n_0}-\frac{\sigma_1^2}{2n_1}\\ . & \frac{\sigma_1^2}{n_1}+\frac{\sigma_0^2}{n_0}\end{pmatrix}, \quad (3.80)$$

where $\mu=\mu_0-\mu_1$.

b. The Second Moment

$$
\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_r^2] &= \mathrm{E}\left[\frac{1}{n^2}\left(\sum_{i=1}^{n_0} I_{\psi(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} I_{\psi(X_i)=0}\right)^2\right] \\
&= E\Bigg[\frac{1}{n^2}\Big(\sum_{i=1}^{n_0} I_{\psi(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} I_{\psi(X_i)=0} + \\
&\quad + \sum_{i=1}^{n_0}\sum_{j\neq i}^{n_0} I_{\psi(X_i)=1} I_{\psi(X_j)=1} + \sum_{i=n_0+1}^{n_0+n_1}\ \sum_{j\neq i}^{n_0+n_1} I_{\psi(X_i)=0} I_{\psi(X_j)=0} + \\
&\quad + \sum_{i=1}^{n_0}\ \sum_{j=n_0+1}^{n_0+n_1} I_{\psi(X_i)=1} I_{\psi(X_j)=0} + \sum_{i=n_0}^{n_0+n_1}\ \sum_{j=1}^{n_0} I_{\psi(X_i)=0} I_{\psi(X_j)=1}\Big)\Bigg] \\
&= \frac{n_0}{n^2}\mathrm{P}\{\psi(X_1)=1\} + \frac{n_1}{n^2}\mathrm{P}\{\psi(X_{n_0+1})=0\} + \\
&\quad + \frac{n_0(n_0-1)}{n^2}\mathrm{P}\{\psi(X_1)=1, \psi(X_2)=1\} + \\
&\quad + \frac{n_1(n_1-1)}{n^2}\mathrm{P}\{\psi(X_{n_0+1})=0, \psi(X_{n_0+2})=0\} + \\
&\quad + \frac{2n_0 n_1}{n^2}\mathrm{P}\{\psi(X_1)=1, \psi(X_{n_0+1})=0\},
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_r^2] = {} & \frac{n_0}{n^2}\Big(\mathrm{P}\{D_0 \geq 0\} + \mathrm{P}\{D_0 < 0\}\Big) + \frac{n_1}{n^2}\Big(\mathrm{P}\{D_1 \geq 0\} + \mathrm{P}\{D_1 < 0\}\Big) + \\
& + \frac{n_0(n_0-1)}{n^2}\Big(\mathrm{P}\{H_{00} \geq 0\} + \mathrm{P}\{H_{00} < 0\}\Big) + \\
& + \frac{n_1(n_1-1)}{n^2}\Big(\mathrm{P}\{H_{11} \geq 0\} + \mathrm{P}\{H_{11} < 0\}\Big) + \\
& + \frac{2n_0 n_1}{n^2}\Big(\mathrm{P}\{H_{01} \geq 0\} + \mathrm{P}\{H_{01} < 0\}\Big),
\end{aligned}
\tag{3.81}
$$

where H_{00}, H_{11}, H_{01} are trivariate Gaussian random variables defined as in (3.55), (3.56), and (3.57) respectively.

c. The Correlation with the True Error

$$\begin{aligned}
\mathrm{E}[\varepsilon\hat{\varepsilon}_r] &= \mathrm{E}\left[\left((1-\gamma)\varepsilon^0+\gamma\varepsilon^1\right)\times\frac{1}{n}\left(\sum_{i=1}^{n_0} I_{\psi(X_i)=1}+\sum_{i=n_0+1}^{n_0+n_1} I_{\psi(X_i)=0}\right)\right] \\
&= \frac{(1-\gamma)}{n}\sum_{i=1}^{n_0}\mathrm{E}[\varepsilon^0 I_{\psi(X_i)=1}]+\frac{(1-\gamma)}{n}\sum_{i=n_0+1}^{n_0+n_1}\mathrm{E}[\varepsilon^0 I_{\psi(X_i)=0}]+ \\
&\quad+\frac{\gamma}{n}\sum_{i=1}^{n_0}\mathrm{E}[\varepsilon^1 I_{\psi(X_i)=0}]+\frac{\gamma}{n}\sum_{i=n_0+1}^{n_0+n_1}\mathrm{E}[\varepsilon^1 I_{\psi(X_i)=0}] \\
&= \frac{1-\gamma}{n}\Big(n_0\mathrm{P}\{\psi(X^*)=1,\psi(X_1)=1\}+n_1\mathrm{P}\{\psi(X^*)=1,\psi(X_{n_0+1})=0\}\Big)+ \\
&\quad+\frac{\gamma}{n}\Big(n_0\mathrm{P}\{\psi(X^{**})=0,\psi(X_1)=1\}+n_1\mathrm{P}\{\psi(X^{**})=0,\psi(X_{n_0+1})=0\}\Big).
\end{aligned}$$

So,

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_r\varepsilon] = \frac{(1-\gamma)n_0}{n}&\Big(\mathrm{P}\{J_{00}\geq 0\}+\mathrm{P}\{J_{00}<0\}\Big)+\frac{\gamma n_1}{n}\Big(\mathrm{P}\{J_{11}\geq 0\}+\mathrm{P}\{J_{11}<0\}\Big)+ \\
&+\frac{\gamma n_0}{n}\Big(\mathrm{P}\{J_{01}\geq 0\}+\mathrm{P}\{J_{01}<0\}\Big)+\frac{(1-\gamma)n_1}{n}\Big(\mathrm{P}\{J_{10}\geq 0\}+\mathrm{P}\{J_{10}<0\}\Big),
\end{aligned} \tag{3.82}$$

where J_{00}, J_{11}, J_{01}, and J_{10} are trivariate Gaussian random variables defined as in (3.58), (3.59), (3.60) and (3.61) respectively.

3. The Moments of the Convex Estimator

This section presents the theorems to compute the first and second moments of the convex bootstrap estimate with arbitrary scalar w, as well as its correlation with the true error ε.

Theorem 10 *Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0 + 1, \ldots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (3.31). Then we have:*

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_w] = {} & \frac{n_0(1-w)}{n}\Big(\mathrm{P}\{D_0 \geq 0\} + \mathrm{P}\{D_0 < 0\}\Big) + \frac{n_1(1-w)}{n}\Big(\mathrm{P}\{D_1 \geq 0\} + \mathrm{P}\{D_1 < 0\}\Big) + \\
& + \sum_C \frac{w\mathrm{P}(C)}{m(C)}\Big(m_0(C)\mathrm{P}\{B_0(C) \geq 0\} + m_0(C)\mathrm{P}\{B_0(C) < 0\} + m_1(C)\mathrm{P}\{B_1(C) \leq 0\} + \\
& + m_1(C)\mathrm{P}\{B_1(C) > 0\}\Big),
\end{aligned} \tag{3.83}$$

where D_0 and D_1 are defined as in (3.79) and (3.80), $B_0(C)$ and $B_1(C)$ as in Theorem 1; m, m_0, and m_1 as in (3.11), and $\mathrm{P}(C)$ in (2.24).

Proof: This is the result of theorem 1, (3.27), and (3.78).

Theorem 11 *Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \dots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0+1, \dots, n_0+n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (3.31). Then we have:*

$$
\begin{aligned}
&\mathrm{E}[\hat{\varepsilon}_w^2] = \\
&(1-w)^2 \Big[\frac{n_0}{n^2} \Big(\mathrm{P}\{D_0 \geq 0\} + \mathrm{P}\{D_0 < 0\} \Big) + \frac{n_1}{n^2} \Big(\mathrm{P}\{D_1 \geq 0\} + \mathrm{P}\{D_1 < 0\} \Big) + \\
&+ \frac{n_0(n_0-1)}{n^2} \Big(\mathrm{P}\{H_{00} \geq 0\} + \mathrm{P}\{H_{00} < 0\} \Big) + \frac{n_1(n_1-1)}{n^2} \Big(\mathrm{P}\{H_{11} \geq 0\} + \mathrm{P}\{H_{11} < 0\} \Big) + \\
&+ \frac{2n_0 n_1}{n^2} \Big(\mathrm{P}\{H_{01} \geq 0\} + \mathrm{P}\{H_{01} < 0\} \Big) \Big] + \\
&+ \sum_C \frac{2w(1-w)\mathrm{P}(C)}{nm(C)} \Bigg[\sum_{i,j=1}^{n_0} \mathrm{I}_{C(i)=0} \mathrm{G}_{00}(i,j,C) + \sum_{i,j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0} \mathrm{G}_{11}(i,j,C) + \\
&+ \sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0} \mathrm{G}_{01}(i,j,C) + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j=1}^{n_0} \mathrm{I}_{C(i)=0} \mathrm{G}_{10}(i,j,C) \Bigg] + \\
&+ \sum_C \lambda_4(w,C) \Big[m_0(C)\mathrm{P}\{B_0(C) \geq 0\} + m_0(C)\mathrm{P}\{B_0(C) < 0\} + m_1(C)\mathrm{P}\{B_1(C) \geq 0\} + \\
&+ m_1(C)\mathrm{P}\{B_1(C) < 0\} + \big(m_0^2(C) - m_0(C) \big) \left(\mathrm{P}\{T_{00}(C) \geq 0\} + \mathrm{P}\{T_{00}(C) < 0\} \right) + \\
&+ \big(m_1^2(C) - m_1(C) \big) \left(\mathrm{P}\{T_{11}(C) \geq 0\} + \mathrm{P}\{T_{11}(C) < 0\} \right) + \\
&+ 2m_0(C)m_1(C) \left(\mathrm{P}\{T_{01}(C) \geq 0\} + \mathrm{P}\{T_{01}(C) < 0\} \right) \Big] + \\
&+ \sum_{C_1 \neq C_2} \lambda_5(w, C_1, C_2) \Big[\sum_{i,j}^{n_0} \mathrm{I}_{C_1(i)=0, C_2(j)=0} \mathrm{F}_{00}(i,j,C_1,C_2) + \\
&+ \sum_{i,j=n_0+1}^{n_0+n_1} \mathrm{I}_{C_1(i)=0, C_2(j)=0} \mathrm{F}_{11}(i,j,C_1,C_2) + \\
&+ \sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C_1(i)=0} \mathrm{I}_{C_2(j)=0} \mathrm{F}_{01}(i,j,C_1,C_2) + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j=1}^{n_0} \mathrm{I}_{C_2(j)=0} \mathrm{I}_{C_1(i)=0} \mathrm{F}_{01}(j,i,C_2,C_1) \Big],
\end{aligned}
$$

where D_0 and D_1 are defined as in (3.79) and (3.80), H_{ab}, $a,b = 0,1$ as in (3.55), (3.56), and (3.57), $B_0(C)$ and $B_1(C)$ as in Theorem 1, $T_{ab}(C)$, $a,b = 0,1$ as in theorem 2, $\mathrm{G}_{ab}(i,j,C)$ as in theorem 4, $\mathrm{F}_{ab}(i,j,C_1,C_2)$ as in theorem 3, and $\lambda_4(w,C) = \frac{w^2 \mathrm{P}(C)}{m^2(C)}$, $\lambda_5(w,C_1,C_2) = \frac{2w^2 \mathrm{P}(C_1)\mathrm{P}(C_2)}{m(C_1)m(C_2)}$, and $\mathrm{P}(C)$ as in (2.24).

Proof: This is the result of theorem 7, theorem 8, (3.28), and (3.81).

Theorem 12 *Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0+1, \ldots, n_0+n_1$ be a set of $n = n_0+n_1$ i.i.d. observations used to derive the classifier in (3.31). Then we have:*

$$
\begin{aligned}
&\mathrm{E}[\hat{\varepsilon}_w \varepsilon] = \\
&= (1-w)\Big[\frac{(1-\gamma)n_0}{n}\Big(\mathrm{P}\{J_{00} \geq 0\} + \mathrm{P}\{J_{00} < 0\}\Big) + \frac{\gamma n_1}{n}\Big(\mathrm{P}\{J_{11} \geq 0\} + \mathrm{P}\{J_{11} < 0\}\Big) + \\
&+ \frac{\gamma n_0}{n}\Big(\mathrm{P}\{J_{01} \geq 0\} + \mathrm{P}\{J_{01} < 0\}\Big) + \frac{(1-\gamma)n_1}{n}\Big(\mathrm{P}\{J_{10} \geq 0\} + \mathrm{P}\{J_{10} < 0\}\Big)\Big] + \\
&+ \sum_C \frac{w\mathrm{P}(C)}{m(C)}\Big[(1-\gamma)\Big(m_0(C)\mathrm{K}_{00}(C) + m_1(C)\mathrm{K}_{10}(C)\Big) + \gamma\Big(m_1(C)\mathrm{K}_{11}(C) + m_0(C)\mathrm{K}_{01}(C)\Big)\Big],
\end{aligned}
$$

where J_{ab} $a, b = 0, 1$ are defined as in (3.58), (3.59), (3.60), and (3.61); K_{ab}, $a, b = 0, 1$ as in theorem 5, m, m_0, and m_1 as in (3.11), and $\mathrm{P}(C)$ as in (2.24).

Proof: This is the result of theorem 9, (3.29), and (3.82)

G. The .632 Bootstrap Error Estimation

Setting $w = .632$ in the formulas of the convex estimator yields the moments of the classic .632 bootstrap estimate.

H. The Optimal Bootstrap Error Estimation

The above theorems allow one to compute the optimal weight w^*, which minimizes the root mean square of the deviation of the convex bootstrap estimate $\hat{\varepsilon}_w$ from the true error ε.

$$
w^* = \arg\min_w \mathrm{RMS}[\hat{\varepsilon}_w] = \arg\min_w \mathrm{RMS}^2[\hat{\varepsilon}_w]
$$

$$\begin{aligned}
\mathrm{RMS}^2[\hat{\varepsilon}_w] &= \mathrm{E}\left(\hat{\varepsilon}_w - \varepsilon\right)^2 \\
&= \mathrm{E}\left[\hat{\varepsilon}_w^2\right] - 2\mathrm{E}\left[\hat{\varepsilon}_w \varepsilon\right] + \mathrm{E}\left[\varepsilon^2\right] \\
&= (1-w)^2\mathrm{E}[\hat{\varepsilon}_r^2] + w^2\mathrm{E}[\hat{\varepsilon}_0^2] + 2w(1-w)\mathrm{E}[\hat{\varepsilon}_r\hat{\varepsilon}_0] + \\
&\quad - 2\left((1-w)\mathrm{E}[\varepsilon\hat{\varepsilon}_r] + w\mathrm{E}[\varepsilon\hat{\varepsilon}_0]\right) + \mathrm{E}[\varepsilon^2] \\
&= w^2\mathrm{E}\left(\hat{\varepsilon}_r - \hat{\varepsilon}_0\right)^2 + 2\mathrm{E}\left[-\varepsilon\hat{\varepsilon}_0 + \hat{\varepsilon}_r\hat{\varepsilon}_0 + \varepsilon\hat{\varepsilon}_r - \hat{\varepsilon}_r^2\right] w + \mathrm{E}\left(\hat{\varepsilon}_r - \varepsilon\right)^2
\end{aligned}$$

The root mean square of the convex estimator $\mathrm{RMS}^2[\hat{\varepsilon}_w]$ is a quadratic function of w. Thus, w^* can be found to be

$$\begin{aligned}
w^* &= -\frac{2\mathrm{E}\left[-\varepsilon\hat{\varepsilon}_0 + \hat{\varepsilon}_r\hat{\varepsilon}_0 + \varepsilon\hat{\varepsilon}_r - \hat{\varepsilon}_r^2\right]}{2\mathrm{E}\left(\hat{\varepsilon}_r - \hat{\varepsilon}_0\right)^2} \\
&= \frac{\mathrm{E}\left[\varepsilon\hat{\varepsilon}_0 - \hat{\varepsilon}_r\hat{\varepsilon}_0 - \varepsilon\hat{\varepsilon}_r + \hat{\varepsilon}_r^2\right]}{\mathrm{E}\left[\hat{\varepsilon}_r^2 - 2\hat{\varepsilon}_r\hat{\varepsilon}_0 + \hat{\varepsilon}_0^2\right]}
\end{aligned}$$

The optimal minimum RMS w^* can be computed using Theorem 7, 8, 9, and the results of (3.81), (3.82). In .632 bootstrap estimation, the combination scalar .632 was chosen heuristically, which represents the proportion of the original sample points in the bootstrap samples [99]. In .632+ bootstrap estimation, w was chosen heuristically adaptively in accordance with the *overfitting rate* [102]. While both of them have been shown to be among the best, they do not guarantee the minimum root mean square.

I. The Unbiased Bootstrap Error Estimation

While the minimized root mean square can be considered as the global criterion for estimation evaluation, unbiased estimation is also of interest to many. Based on Theorem 10, (3.78) and (3.20), we can find w_u that guarantees an unbiased bootstrap estimation.

$$\mathrm{E}[\hat{\varepsilon}_w - \varepsilon] = \mathrm{E}[(1-w_u)\hat{\varepsilon}_r + w_u\hat{\varepsilon}_0 - \varepsilon] = 0$$

$$w_u = \frac{\mathrm{E}[\varepsilon - \hat{\varepsilon}_r]}{\mathrm{E}[\hat{\varepsilon}_0 - \hat{\varepsilon}_r]} \tag{3.84}$$

The unbiased scalar w_u can be obtained based on Theorem 3.23, (3.78), (3.20). In order to compute $\mathrm{E}[\hat{\varepsilon}_0]$, we need to go through all Cs. However, note that

$$\mathrm{E}[\hat{\varepsilon}_0] = \sum_{C} P(C)\mathrm{E}[\varepsilon_C] = \sum_{(s_0,s_1)} \mathrm{P}(s_0,s_1)\mathrm{E}[\varepsilon_{C(s_0,s_1)}] \tag{3.85}$$

where $C(s_0,s_1)$ is any bootstrap vector C that satisfies (3.12). Since the number of all configurations of the vector (s_0,s_1) is much smaller than the number of all configurations of C, this provides the basis for an efficient way to calculate $E[\hat{\varepsilon}_0]$, *provided* that we have a method of directly calculating $P(s_0,s_1)$ without having to go through all C. Problem arises for large n, which can be greatly alleviated by using (3.85) and we create a method for computing $\mathrm{P}(s_0,s_1)$ efficiently that is described in the Appendix.

We present below examples of application of the formulas derived in the paper for the unbiased weight w_u. Figure I displays the exact w_u as a function of Bayes error for different sample sizes, and as a function of number of samples for different Bayes error, in the univariate case. We can see that for small Bayes error, the unbiased weight tends to be closer to the heuristic 0.632 weight than for large Bayes error. We also see that as sample size increases, the unbiased weight appears to be converging to a fixed value, which is not the heuristic 0.632 weight.

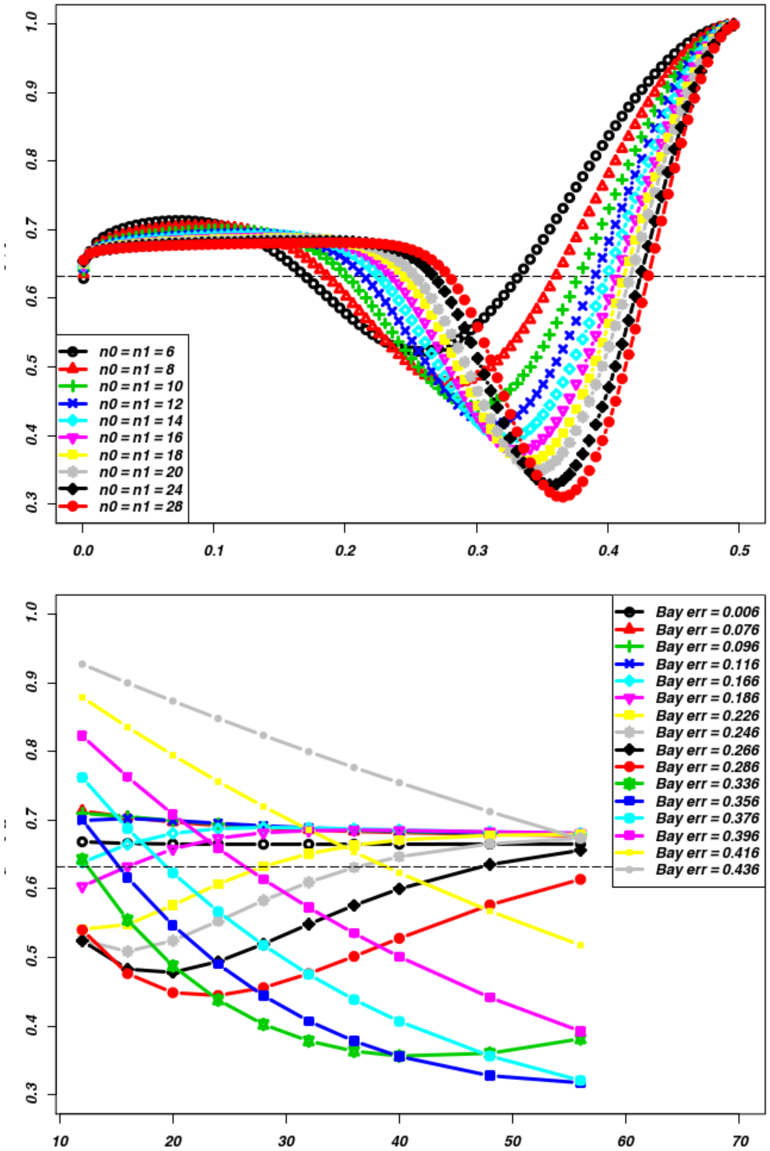


Fig. 1. Optimal weight w_u in the univariate case. The top figure displays w_u as a function of Bayes error for different sample sizes, whereas the bottom figure displays w_u as a function of the number of samples for different Bayes errors.

CHAPTER IV

BOOTSTRAP ERROR ESTIMATION - MULTIVARIATE MODEL *

This chapter presents the theoretical analysis of complete bootstrap error estimation for linear discriminant analysis under standard multivariate Gaussian model with the same covariance matrix. The covariance matrix of the label feature distribution is assumed to be known. The analysis is concerned with some bootstrap estimators including zero, .632, and convex bootstrap estimation. The results include the first moments, the second moments, the cross moments of these bootstrap estimators with the true error and the resubstitution estimator. As a result, we obtain the exact formulas for the bias, variance, and the root mean squared error of the estimation deviations from the true error, which are the usual metrics for evaluation of estimation methods. Also, we propose unbiased bootstrap estimation by zeroing the deviation bias and optimal bootstrap estimation by minimizing the root mean square of the deviation. Different from the univariate case, the formulas in the multivariate case are involved with doubly noncentral F random variables, including univariate and bivariate F. The efficient algorithm introduced in Chapter III is also applicable for the multivariate models. Finally, some figures of the optimal convex scalar for the unbiased bootstrap are provided for different number of samples under various multivariate Gaussian models.

* Part of this chapter is reprinted with permission from "Unbiased Bootstrap Error Estimation for Linear Discriminant Analysis." by T. T. Vu, U. M. Braga-Neto, and E. R. Dougherty, 2010. submitted, copyright 2010 of *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

A. Multivariate Model

1. Bootstrapped Linear Discriminant

Let $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0 + 1, \ldots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations. The covariance matrix Σ is assumed to be known. *Linear Discriminant Analysis* (LDA) employs Anderson's W discriminant, which is defined as follows:

$$W(X) = \left(X - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right)^T \Sigma^{-1} (\hat{\mu}_0 - \hat{\mu}_1) \tag{4.1}$$

where

$$\begin{aligned} \hat{\mu}_0 &= \frac{1}{n_0} \sum_{i=1}^{n_0} X_i, \\ \hat{\mu}_1 &= \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} X_i \end{aligned} \tag{4.2}$$

are the sample means of the sample sets S_0 and S_1, respectively. This defines the LDA classification rule, whereby the designed LDA classifier is defined by:

$$\psi(X) = \begin{cases} 1, & \text{if } W(X) < 0 \\ 0, & \text{if } W(X) \geq 0 \end{cases}, \tag{4.3}$$

The C-bootstrap LDA classifier is obtained by substituting $\hat{\mu}_i^C$, defined in (3.32), for $\hat{\mu}_i$, $i = 0, 1$, in (4.1):

$$W_C(X) = \left(X - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\right)^T \Sigma^{-1} \left(\hat{\mu}_0^C - \hat{\mu}_1^C\right) \tag{4.4}$$

$$\psi_C(X) = \begin{cases} 1, & \text{if } \left(X - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\right)^T \Sigma^{-1} (\hat{\mu}_0^C - \hat{\mu}_1^C) < 0 \\ 0, & \text{otherwise} \end{cases}. \tag{4.5}$$

2. Some Definitions

The multivariate model is different from the univariate one, in which we can break down all the moments of error estimators to Gaussian distribution. In multivariate case, they are involved with F distributions. Given the scarce literature of bivariate F distributions, we will need to define the following functions to help represent the results.

Suppose Z_j, $j \in \{1,2,3,4\}$ are jointly p-dimensional Gaussian random vectors with expectations $\mathrm{E}[Z_j]$, the covariance matrix Σ_j, and the cross-covariance matrices Σ_{ij}. Then $Y = [Z_1^T\, Z_2^T]^T$ and $Z = [Z_1^T\, Z_2^T\, Z_3^T\, Z_4^T]^T$ are Gaussian random vectors of dimension $2p$ and $4p$, respectively. We have:

$$\mathrm{E}[Y] = \left[\mathrm{E}[Z_1]^T\, \mathrm{E}[Z_2]^T\right]^T, \;\; \mathrm{E}[Z] = \left[\mathrm{E}[Z_1]^T\, \mathrm{E}[Z_2]^T\, \mathrm{E}[Z_3]^T\, \mathrm{E}[Z_4]^T\right]^T,$$

$$\Sigma_Y = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ . & \Sigma_2 \end{pmatrix}, \;\; \Sigma_Z = \begin{pmatrix} \Sigma_1 & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ . & \Sigma_2 & \Sigma_{23} & \Sigma_{24} \\ . & . & \Sigma_3 & \Sigma_{34} \\ . & . & . & \Sigma_4 \end{pmatrix}.$$

Because Z_js are jointly Gaussian distributed, $\mathrm{E}[Z_j]$s and Σ_Z fully specify Z_js and their relations.

Define the following probabilities:

$$\begin{aligned}
G_0(Y) &= P\{Z_1^T Z_1 - Z_2^T Z_2 < 0\}, \\
G_1(Y) &= P\{Z_1^T Z_1 - Z_2^T Z_2 > 0\}, \\
G_{00}(Z) &= P\{Z_1^T Z_1 - Z_2^T Z_2 < 0, Z_3^T Z_3 - Z_4^T Z_4 < 0\}, \\
G_{11}(Z) &= P\{Z_1^T Z_1 - Z_2^T Z_2 > 0, Z_3^T Z_3 - Z_4^T Z_4 > 0\}, \\
G_{01}(Z) &= P\{Z_1^T Z_1 - Z_2^T Z_2 < 0, Z_3^T Z_3 - Z_4^T Z_4 > 0\}, \\
G_{10}(Z) &= P\{Z_1^T Z_1 - Z_2^T Z_2 > 0, Z_3^T Z_3 - Z_4^T Z_4 < 0\}.
\end{aligned} \tag{4.6}$$

We can see that G_0, G_1 are probabilities of a $F = c\frac{Z_1^T Z_1}{Z_2^T Z_2}$ (c is a normalization constant) random variable if Z_1, Z_2 are independent; G_{00}, G_{11}, G_{01}, and G_{10} are probabilities of a bivariate $F = (F_1 F_2)$ random variable where $F_1 = c_1\frac{Z_1^T Z_1}{Z_2^T Z_2}, F_2 = c_2\frac{Z_3^T Z_3}{Z_4^T Z_4}$ if Z_1, Z_2 are independent, and so are Z_3, Z_4. Given the scarce literature of bivariate F distributions [142], we will use Gs, the previously defined functions of multivariate Gaussian distribution, as standard notations in the following results.

B. The C-bootstrap Linear Classifier

This section presents the theorems to compute the first, second moment of the hold-out error of the C-bootstrap linear classifier $\psi_C(X)$ and it cross moment with the resubstitution estimator and the true error of the original linear classifier $\psi(X)$.

Theorem 13 *Let $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0 + 1, \ldots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (4.5). Then given a bootstrap vector C, we have:*

$$\mathrm{E}[\hat{\varepsilon}_C] = \frac{1}{m(C)}\Big[m_0(C)G_0\left(Z_0(C)\right) + m_1(C)G_1\left(Z_1(C)\right)\Big], \tag{4.7}$$

where m, m_0, m_1 are defined as in (3.11), G_0 and G_1 in (4.6), and Z_0, Z_1 are 2p-dimensional Gaussian random vectors with $Z_i(C) = \left[(Z_i^1)^T (Z_i^2)^T\right]^T$, $i = 0, 1$ with

$$\mathrm{E}\left[Z_0^1\right] = \mathrm{E}\left[Z_1^2\right] = \left[s^{-\frac{1}{2}} + (s+4)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu,$$
$$\mathrm{E}\left[Z_0^2\right] = \mathrm{E}\left[Z_1^1\right] = \left[s^{-\frac{1}{2}} - (s+4)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu,$$
$$\Sigma_{Z_0} = \Sigma_{Z_1} = \begin{pmatrix} 2(1+\rho)I_p & \mathbf{0_{p\times p}} \\ . & 2(1-\rho)I_p \end{pmatrix},$$

where $\rho = \frac{s_0 - s_1}{\sqrt{s(s+4)}}$, s, s_0, s_1 are defined as in (3.12), and $\mu = \mu_0 - \mu_1$.

Proof: See appendix.

Theorem 14 *Let $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \dots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0 + 1, \dots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (4.5). Then we have:*

$$\begin{aligned} \mathrm{E}[\hat{\varepsilon}_C^2] = {} & \frac{m_0}{m^2} G_0(Z_0(C)) + \frac{m_1}{m^2} G_1(Z_1(C)) + \frac{m_0(m_0-1)}{m^2} G_{00}\left(T_{00}(C)\right) + \\ & + \frac{m_1(m_1-1)}{m^2} G_{11}\left(T_{11}(C)\right) + \frac{m_0 m_1}{m^2}\Big[G_{01}\left(T_{01}(C)\right) + G_{10}\left(T_{01}(C)\right)\Big], \end{aligned}$$

where the functions G are defined as in (4.6), m_0, m_1, and m in (3.11), Z_0 and Z_1 as in theorem 13, T_{ab}, $a, b = 0, 1$ are $4p$-dimensional Gaussian vectors with
$T_{ab} = \left[(T_{ab}^1)^T \, (T_{ab}^2)^T \, (T_{ab}^3)^T \, (T_{ab}^4)^T\right]^T$ and

$$\begin{aligned} \mathrm{E}\left[T_{00}^1\right] = \mathrm{E}\left[T_{00}^3\right] = \mathrm{E}\left[T_{11}^2\right] = \mathrm{E}\left[T_{11}^4\right] = \mathrm{E}\left[T_{01}^1\right] = \mathrm{E}\left[T_{01}^4\right] = \left[s^{-\frac{1}{2}} + (s+4)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \\ \mathrm{E}\left[T_{00}^2\right] = \mathrm{E}\left[T_{00}^4\right] = \mathrm{E}\left[T_{11}^1\right] = \mathrm{E}\left[T_{11}^3\right] = \mathrm{E}\left[T_{01}^2\right] = \mathrm{E}\left[T_{01}^3\right] = \left[s^{-\frac{1}{2}} - (s+4)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \end{aligned}$$

and

$$\Sigma_{T_{11}} = \Sigma_{T_{01}} = \Sigma_{T_{00}} = \begin{pmatrix} 2(1+\rho)I_p & \mathbf{0_{p\times p}} & \left(\frac{2s+4}{s+4} + \frac{2(s_1-s_0)}{\sqrt{s(s+4)}}\right)I_p & \frac{2s+4}{s+4}I_p \\ . & 2(1-\rho)I_p & \frac{2s+4}{s+4}I_p & \left(\frac{2s+4}{s+4} - \frac{2(s_1-s_0)}{\sqrt{s(s+4)}}\right)I_p \\ . & . & 2(1+\rho)I_p & \mathbf{0_{p\times p}} \\ . & . & . & 2(1-\rho)I_p \end{pmatrix},$$

where s, s_0, s_1 are defined as in (3.12), $\rho = \frac{s_0 - s_1}{\sqrt{s(s+4)}}$.

Proof: See appendix.

Theorem 15 *Let $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \dots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0 + 1, \dots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (4.5). Then for*

$C_1 \neq C_2$ we have:

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_{C_1}\hat{\varepsilon}_{C_2}] = \frac{1}{m(C_1)m(C_2)} \Big[& \sum_{i=1}^{n_0}\sum_{j=1}^{n_0} I_{C_1(i)=0,C_2(j)=0}\, G_{00}\big(F_{00}(C_1,C_2,i,j)\big) + \\
& + \sum_{i=n_0+1}^{n_0+n_1}\sum_{j=n_0+1}^{n_0+n_1} I_{C_1(i)=0,C_2(j)=0}\, G_{11}\big(F_{11}(C_1,C_2,i,j)\big) + \\
& + \sum_{i=1}^{n_0}\sum_{j=n_0}^{n_0+n_1} I_{C_1(i)=0,C_2(j)=0}\, G_{01}\big(F_{01}(C_1,C_2,i,j)\big) + \\
& + \sum_{i=n_0}^{n_0+n_1}\sum_{j=1}^{n_0} I_{C_1(i)=0,C_2(j)=0}\, G_{10}\big(F_{01}(C_2,C_1,j,i)\big) \Big],
\end{aligned}$$

where the functions G are defined as in (4.6), m as in (3.11), $F_{ab}(C_1,C_2,i,j)$, $a,b = 0,1$ are $4p$-dimensional Gaussian vectors with $F_{ab} = \left[(F^1_{ab})^T\,(F^2_{ab})^T\,(F^3_{ab})^T\,(F^4_{ab})^T\right]^T$, and

$$\begin{aligned}
\mathrm{E}\left[F^1_{00}\right] = \mathrm{E}\left[F^2_{11}\right] = \mathrm{E}\left[F^1_{01}\right] &= \left[s(C_1)^{-\frac{1}{2}} + (s(C_1)+4)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \\
\mathrm{E}\left[F^2_{00}\right] = \mathrm{E}\left[F^1_{11}\right] = \mathrm{E}\left[F^2_{01}\right] &= \left[s(C_1)^{-\frac{1}{2}} - (s(C_1)+4)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \\
\mathrm{E}\left[F^3_{00}\right] = \mathrm{E}\left[F^4_{11}\right] = \mathrm{E}\left[F^4_{01}\right] &= \left[s(C_2)^{-\frac{1}{2}} + (s(C_2)+4)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \\
\mathrm{E}\left[F^4_{00}\right] = \mathrm{E}\left[F^3_{11}\right] = \mathrm{E}\left[F^3_{01}\right] &= \left[s(C_2)^{-\frac{1}{2}} - (s(C_2)+4)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu,
\end{aligned}$$

and

$$\Sigma_{F_{ab}(C_1,C_2,i,j)} = \begin{pmatrix} 2(1+\rho(C_1))I_p & \mathbf{0_{p\times p}} & \kappa_{ab1}I_p & \kappa_{ab2}I_p \\ . & 2(1-\rho(C_1))I_p & \kappa_{ab3}I_p & \kappa_{ab4}I_p \\ . & . & 2(1+\rho(C_2))I_p & \mathbf{0_{p\times p}} \\ . & . & . & 2(1-\rho(C_2))I_p \end{pmatrix},$$

where

$$\kappa_{001}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}+\frac{\frac{2C_1(j)}{n_0}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}+\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}+\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{002}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}-\frac{\frac{2C_1(j)}{n_0}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}+\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}-\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{003}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}+\frac{\frac{2C_1(j)}{n_0}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}-\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}-\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{004}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}-\frac{\frac{2C_1(j)}{n_0}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}-\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}+\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{111}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}-\frac{\frac{2C_1(j)}{n_1}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}-\frac{\frac{2C_2(i)}{n_1}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}+\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_1}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{112}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}+\frac{\frac{2C_1(j)}{n_1}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}-\frac{\frac{2C_2(i)}{n_1}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}-\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_1}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{113}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}-\frac{\frac{2C_1(j)}{n_1}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}+\frac{\frac{2C_2(i)}{n_1}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}-\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_1}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{114}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}+\frac{\frac{2C_1(j)}{n_1}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}+\frac{\frac{2C_2(i)}{n_1}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}+\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_1}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{011}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}-\frac{\frac{2C_1(j)}{n_1}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}+\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}+\frac{r_0+r_1-\frac{2C_1(j)}{n_1}-\frac{2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{012}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}+\frac{\frac{2C_1(j)}{n_1}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}+\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}-\frac{r_0+r_1-\frac{2C_1(j)}{n_1}-\frac{2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{013}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}-\frac{\frac{2C_1(j)}{n_1}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}-\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}-\frac{r_0+r_1-\frac{2C_1(j)}{n_1}-\frac{2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

$$\kappa_{014}=\Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}+\frac{\frac{2C_1(j)}{n_1}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}-\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}+\frac{r_0+r_1-\frac{2C_1(j)}{n_1}-\frac{2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big),$$

where s, s_0, s_1, r_0, r_1 are defined as in (3.12), (3.13), $\rho(C)=\frac{s_0-s_1}{\sqrt{s(s+4)}}$.

Proof: See appendix.

Theorem 16 *Let $X_i \sim N(\mu_0,\Sigma)$ for $i=1,\ldots,n_0$, and $X_i \sim N(\mu_1,\Sigma)$ for $i=n_0+1,\ldots,n_0+$*

n_1 be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (4.5). Then we have:

$$\mathrm{E}[\hat{\varepsilon}_C \hat{\varepsilon}_r] = \frac{1}{nm(C)} \Bigg[\sum_{i=1}^{n_0} \sum_{j=1}^{n_0} I_{C(i)=0}\, G_{00}(M_{00}(C,i,j)) + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j=n_0+1}^{n_0+n_1} I_{C(i)=0}\, G_{11}(M_{11}(C,i,j)) + \\ + \sum_{i=1}^{n_0} \sum_{j=n_0}^{n_0+n_1} I_{C(i)=0}\, G_{01}(M_{01}(C,i,j)) + \sum_{i=n_0}^{n_0+n_1} \sum_{j=1}^{n_0} I_{C(j)=0}\, G_{10}(M_{10}(C,i,j)) \Bigg],$$

where the functions G are defined as in (4.6), m as in (3.11), $M_{ab}(C,i,j), a,b = 0,1$ are $4p$-dimensional Gaussian vectors with

$$\begin{aligned} M_{ab} &= [(M_{ab}^1)^T\,(M_{ab}^2)^T\,(M_{ab}^3)^T\,(M_{ab}^4)^T]^T, \\ \mathrm{E}\left[M_{00}^1\right] = \mathrm{E}\left[M_{11}^2\right] = \mathrm{E}\left[M_{01}^1\right] = \mathrm{E}\left[M_{10}^2\right] &= \left[s^{-\frac{1}{2}} + (s+4)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu, \\ \mathrm{E}\left[M_{00}^2\right] = \mathrm{E}\left[M_{11}^1\right] = \mathrm{E}\left[M_{01}^2\right] = \mathrm{E}\left[M_{10}^1\right] &= \left[s^{-\frac{1}{2}} - (s+4)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu, \\ \mathrm{E}\left[M_{00}^3\right] = \mathrm{E}\left[M_{10}^4\right] &= \left[\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(1 - \frac{3}{4n_0} + \frac{1}{4n_1}\right)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu, \\ \mathrm{E}\left[M_{00}^4\right] = \mathrm{E}\left[M_{10}^3\right] &= \left[\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(1 - \frac{3}{4n_0} + \frac{1}{4n_1}\right)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu, \\ \mathrm{E}\left[M_{11}^4\right] = \mathrm{E}\left[M_{01}^4\right] &= \left[\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(1 - \frac{3}{4n_1} + \frac{1}{4n_0}\right)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu, \\ \mathrm{E}\left[M_{11}^3\right] = \mathrm{E}\left[M_{01}^3\right] &= \left[\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(1 - \frac{3}{4n_1} + \frac{1}{4n_0}\right)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu, \end{aligned}$$

and s, s_0 and s_1 are defined as in (3.12), and

$$\Sigma_{M_{ab}(C,i,j)} = \begin{pmatrix} 2(1+\rho)I_p & \mathbf{0_{p\times p}} & \eta_{ab1} I_p & \eta_{ab2} I_p \\ . & 2(1-\rho)I_p & \eta_{ab3} I_p & \eta_{ab4} I_p \\ . & . & 2(1+\rho_b)I_p & \mathbf{0_{p\times p}} \\ . & . & . & 2(1-\rho_b)I_p \end{pmatrix}.$$

where $\rho_0 = \sqrt{\frac{n}{4n_0n_1-3n_1+n_0}}$, $\rho_1 = \sqrt{\frac{n}{4n_0n_1-3n_0+n_1}}$, $\rho = \frac{s_0-s_1}{\sqrt{s(s+4)}}$ *with*

$$\eta_{001} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} + \frac{2n_1C(j)-n_1+n_0}{\sqrt{s(4n_0n_1-3n_1+n_0)}} + \sqrt{\frac{n}{s+4}} + \frac{4n_0n_1I_{i=j}-2n_1C(j)-n_1+n_0}{\sqrt{(s+4)(4n_0n_1-3n_1+n_0)}}\right),$$

$$\eta_{002} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} - \frac{2n_1C(j)-n_1+n_0}{\sqrt{s(4n_0n_1-3n_1+n_0)}} + \sqrt{\frac{n}{s+4}} - \frac{4n_0n_1I_{i=j}-2n_1C(j)-n_1+n_0}{\sqrt{(s+4)(4n_0n_1-3n_1+n_0)}}\right),$$

$$\eta_{003} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} + \frac{2n_1C(j)-n_1+n_0}{\sqrt{s(4n_0n_1-3n_1+n_0)}} - \sqrt{\frac{n}{s+4}} - \frac{4n_0n_1I_{i=j}-2n_1C(j)-n_1+n_0}{\sqrt{(s+4)(4n_0n_1-3n_1+n_0)}}\right),$$

$$\eta_{004} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} - \frac{2n_1C(j)-n_1+n_0}{\sqrt{s(4n_0n_1-3n_1+n_0)}} - \sqrt{\frac{n}{s+4}} + \frac{4n_0n_1I_{i=j}-2n_1C(j)-n_1+n_0}{\sqrt{(s+4)(4n_0n_1-3n_1+n_0)}}\right),$$

$$\eta_{111} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} - \frac{2n_0C(j)-n_0+n_1}{\sqrt{s(4n_0n_1-3n_0+n_1)}} - \sqrt{\frac{n}{s+4}} + \frac{4n_0n_1I_{i=j}-2n_0C(j)-n_0+n_1}{\sqrt{(s+4)(4n_1n_0-3n_0+n_1)}}\right),$$

$$\eta_{112} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} + \frac{2n_0C(j)-n_0+n_1}{\sqrt{s(4n_1n_0-3n_0+n_1)}} - \sqrt{\frac{n}{s+4}} - \frac{4n_1n_0I_{i=j}-2n_0C(j)-n_0+n_1}{\sqrt{(s+4)(4n_1n_0-3n_0+n_1)}}\right),$$

$$\eta_{113} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} - \frac{2n_0C(j)-n_0+n_1}{\sqrt{s(4n_1n_0-3n_0+n_1)}} + \sqrt{\frac{n}{s+4}} - \frac{4n_1n_0I_{i=j}-2n_0C(j)-n_0+n_1}{\sqrt{(s+4)(4n_1n_0-3n_0+n_1)}}\right),$$

$$\eta_{114} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} + \frac{2n_0C(j)-n_0+n_1}{\sqrt{s(4n_1n_0-3n_0+n_1)}} + \sqrt{\frac{n}{s+4}} + \frac{4n_1n_0I_{i=j}-2n_0C(j)-n_0+n_1}{\sqrt{(s+4)(4n_1n_0-3n_0+n_1)}}\right),$$

$$\eta_{011} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} - \frac{2n_0C(j)-n_0+n_1}{\sqrt{s(4n_0n_1-3n_0+n_1)}} + \sqrt{\frac{n}{s+4}} + \frac{n_0-2n_0C(j)-n_1}{\sqrt{(4n_0n_1-3n_0+n_1)(s+4)}}\right),$$

$$\eta_{012} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} + \frac{2n_0C(j)-n_0+n_1}{\sqrt{s(4n_0n_1-3n_0+n_1)}} + \sqrt{\frac{n}{s+4}} - \frac{n_0-2n_0C(j)-n_1}{\sqrt{(4n_0n_1-3n_0+n_1)(s+4)}}\right),$$

$$\eta_{013} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} - \frac{2n_0C(j)-n_0+n_1}{\sqrt{s(4n_0n_1-3n_0+n_1)}} - \sqrt{\frac{n}{s+4}} - \frac{n_0-2n_0C(j)-n_1}{\sqrt{(4n_0n_1-3n_0+n_1)(s+4)}}\right),$$

$$\eta_{014} = \frac{1}{\sqrt{n_0n_1}}\left(\sqrt{\frac{n}{s}} + \frac{2n_0C(j)-n_0+n_1}{\sqrt{s(4n_0n_1-3n_0+n_1)}} - \sqrt{\frac{n}{s+4}} + \frac{n_0-2n_0C(j)-n_1}{\sqrt{(4n_0n_1-3n_0+n_1)(s+4)}}\right),$$

$$\eta_{101} = \frac{1}{\sqrt{n_0 n_1}}\left(\sqrt{\frac{n}{s}} + \frac{2n_1 C(j) - n_1 + n_0}{\sqrt{s(4n_0 n_1 - 3n_1 + n_0)}} - \sqrt{\frac{n}{s+4}} + \frac{n_1 - 2n_1 C(j) - n_0}{\sqrt{(4n_0 n_1 - 3n_1 + n_0)(s+4)}}\right),$$

$$\eta_{102} = \frac{1}{\sqrt{n_0 n_1}}\left(\sqrt{\frac{n}{s}} - \frac{2n_1 C(j) - n_1 + n_0}{\sqrt{s(4n_0 n_1 - 3n_1 + n_0)}} - \sqrt{\frac{n}{s+4}} - \frac{n_1 - 2n_1 C(j) - n_0}{\sqrt{(4n_0 n_1 - 3n_1 + n_0)(s+4)}}\right),$$

$$\eta_{103} = \frac{1}{\sqrt{n_0 n_1}}\left(\sqrt{\frac{n}{s}} + \frac{2n_1 C(j) - n_1 + n_0}{\sqrt{s(4n_0 n_1 - 3n_1 + n_0)}} + \sqrt{\frac{n}{s+4}} - \frac{n_1 - 2n_1 C(j) - n_0}{\sqrt{(4n_0 n_1 - 3n_1 + n_0)(s+4)}}\right),$$

$$\eta_{104} = \frac{1}{\sqrt{n_0 n_1}}\left(\sqrt{\frac{n}{s}} - \frac{2n_1 C(j) - n_1 + n_0}{\sqrt{s(4n_0 n_1 - 3n_1 + n_0)}} + \sqrt{\frac{n}{s+4}} + \frac{n_1 - 2n_1 C(j) - n_0}{\sqrt{(4n_0 n_1 - 3n_1 + n_0)(s+4)}}\right).$$

Proof: See appendix.

Theorem 17 *Let $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0 + 1, \ldots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (4.5). Then we have:*

$$\begin{aligned} \mathrm{E}[\hat{\varepsilon}_C \varepsilon] = {} & \frac{m_0(1-\gamma)}{m} G_{00}(K_{00}(C)) + \frac{m_1 \gamma}{m} G_{11}(K_{11}(C)) + \\ & + \frac{m_0 \gamma}{m} G_{01}(K_{01}(C)) + \frac{m_1(1-\gamma)}{m} G_{10}(K_{10}(C)), \end{aligned}$$

where the functions G are defined as in (4.6), m as in (3.11), $K_{ab}(C,i,j), a,b = 0,1$ are $4p$-dimensional Gaussian vectors with $K_{ab} = [(K^1_{ab})^T \, (K^2_{ab})^T \, (K^3_{ab})^T \, (K^4_{ab})^T]^T$, and

$$\mathrm{E}\left[K^1_{00}\right] = \mathrm{E}\left[K^2_{11}\right] = \mathrm{E}\left[K^1_{01}\right] = \mathrm{E}\left[K^2_{10}\right] = \left[\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(4 + \frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu,$$

$$\mathrm{E}\left[K^2_{00}\right] = \mathrm{E}\left[K^1_{11}\right] = \mathrm{E}\left[K^2_{01}\right] = \mathrm{E}\left[K^1_{10}\right] = \left[\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(4 + \frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu,$$

$$\mathrm{E}\left[K^3_{00}\right] = \mathrm{E}\left[K^4_{11}\right] = \mathrm{E}\left[K^4_{01}\right] = \mathrm{E}\left[K^3_{10}\right] = \left[s^{-\frac{1}{2}} + (s+4)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu,$$

$$\mathrm{E}\left[K^4_{00}\right] = \mathrm{E}\left[K^3_{11}\right] = \mathrm{E}\left[K^3_{01}\right] = \mathrm{E}\left[K^4_{10}\right] = \left[s^{-\frac{1}{2}} - (s+4)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu,$$

and

$$\Sigma_{K_{ab}(C,i,j)} = \begin{pmatrix} 2(1+\rho_*)I_p & \mathbf{0_{p\times p}} & \zeta_{ab1}I_p & \zeta_{ab2}I_p \\ . & 2(1-\rho_*)I_p & \zeta_{ab3}I_p & \zeta_{ab4}I_p \\ . & . & 2(1+\rho)I_p & \mathbf{0_{p\times p}} \\ . & . & . & 2(1-\rho)I_p \end{pmatrix}.$$

where $\rho_* = \frac{n_1-n_0}{\sqrt{n(n+4n_0n_1)}}$, $\rho = \frac{s_0-s_1}{\sqrt{s(s+4)}}$ *with*

$$\zeta_{001} = \zeta_{011} = \zeta_{112} = \zeta_{102} = \frac{\sqrt{s+4}+\sqrt{s}}{\sqrt{n_0n_1s(s+4)}}\left(\sqrt{n}+\frac{n_0-n_1}{\sqrt{4n_0n_1+n}}\right),$$

$$\zeta_{002} = \zeta_{012} = \zeta_{111} = \zeta_{101} = \frac{\sqrt{s+4}-\sqrt{s}}{\sqrt{n_0n_1s(s+4)}}\left(\sqrt{n}+\frac{n_0-n_1}{\sqrt{4n_0n_1+n}}\right),$$

$$\zeta_{003} = \zeta_{013} = \zeta_{114} = \zeta_{104} = \frac{\sqrt{s+4}+\sqrt{s}}{\sqrt{n_0n_1s(s+4)}}\left(\sqrt{n}-\frac{n_0-n_1}{\sqrt{4n_0n_1+n}}\right),$$

$$\zeta_{004} = \zeta_{014} = \zeta_{113} = \zeta_{103} = \frac{\sqrt{s+4}-\sqrt{s}}{\sqrt{n_0n_1s(s+4)}}\left(\sqrt{n}-\frac{n_0-n_1}{\sqrt{4n_0n_1+n}}\right),$$

where s, s_0*, and* s_1 *are defined as in (3.12),* $\mu = \mu_0 - \mu_1$.

Proof: See appendix.

C. The Zero Bootstrap Error Estimation

Theorem 18 *Let* $X_i \sim N(\mu_0,\Sigma)$ *for* $i = 1,\ldots,n_0$*, and* $X_i \sim N(\mu_1,\Sigma)$ *for* $i = n_0+1,\ldots,n_0+n_1$ *be a set of* $n = n_0+n_1$ *i.i.d. observations used to derive the classifier in (4.5). Then given a bootstrap vector C, we have:*

$$\mathrm{E}\left[\hat{\varepsilon}_0\right] = \sum_C \frac{P(C)}{m(C)}\Big(m_0(C)G_0\left(Z_0(C)\right)+m_1(C)G_1\left(Z_1(C)\right)\Big), \tag{4.8}$$

where the functions G are defined as in (4.6), m_0*,* m_1*, and m as in (3.11),* $Z_0(C)$ *and* $Z_1(C)$ *as in theorem 13.*

Proof: This is the immediate result of theorem 13 and (3.23).

Theorem 19 *Let $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0 + 1, \ldots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (4.5). Then given a bootstrap vector C, we have:*

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_0^2] = & \sum_C \frac{\mathrm{P}(C)}{m^2(C)} \Big(m_0(C) G_0\left(Z_0(C)\right) + m_1(C) G_1\left(Z_1(C)\right) + \\
& + \sum_{i=1}^{n_0} \sum_{j \neq i}^{n_0} \mathrm{I}_{C(i)=0, C(j)=0} G_{00}\left(T_{00}(C,i,j)\right) + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j \neq i}^{n_0+n_1} \mathrm{I}_{C(i)=0, C(j)=0} G_{11}\left(T_{11}(C,i,j)\right) + \\
& + \sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0, C(j)=0} G_{01}\left(T_{01}(C,i,j)\right) + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j=1}^{n_0} \mathrm{I}_{C(i)=0, C(j)=0} G_{10}\left(T_{01}(C,j,i)\right) \Big) + \\
& + \sum_{C_1 \neq C_2} \frac{2\mathrm{P}(C_1)\mathrm{P}(C_2)}{m(C_1)m(C_2)} \left[\sum_{i,j=1}^{n_0} \mathrm{I}_{C_1(i)=0, C_2(j)=0}\, G_{00}\left(F_{00}(C_1, C_2, i, j)\right) + \right. \\
& + \sum_{i,j=n_0+1}^{n_0+n_1} \mathrm{I}_{C_1(i)=0, C_2(j)=0}\, G_{11}\left(F_{11}(C_1, C_2, i, j)\right) + \\
& + \sum_{i=1}^{n_0} \sum_{j=n_0}^{n_0+n_1} \mathrm{I}_{C_1(i)=0, C_2(j)=0}\, G_{01}\left(F_{01}(C_1, C_2, i, j)\right) + \\
& \left. + \sum_{i=n_0}^{n_0+n_1} \sum_{j=1}^{n_0} \mathrm{I}_{C_1(i)=0, C_2(j)=0}\, G_{10}\left(F_{01}(C_2, C_1, j, i)\right) \right]
\end{aligned}$$

where the functions G are defined as in (4.6), Z_0 and Z_1 as in theorem 13, T_{ab} as in theorem 14, and F_{ab}, $a, b = 0, 1$ as in theorem 15, m_0, m_1, and m in (3.11), and $\mathrm{P}(C)$ in (2.24).

Proof: This is the immediate result of theorem 14, 15 and (3.24).

Theorem 20 *Let $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0 + 1, \ldots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (4.5). Then* *given a bootstrap vector C, we have:*

$$\mathrm{E}\left[\hat{\varepsilon}_0\hat{\varepsilon}_r\right] = \sum_C \frac{P(C)}{nm(C)}\Big[\sum_{i,j=1}^{n_0} \mathrm{I}_{C(i)=0}\, G_{00}(M_{00}(C,i,j)) + \sum_{i,j=n_0+1}^{n_0+n_1} \mathrm{I}_{C(i)=0}\, G_{11}(M_{11}(C,i,j)) +$$

$$+\sum_{i=1}^{n_0}\sum_{j=n_0}^{n_0+n_1} \mathrm{I}_{C(i)=0}\, G_{01}(M_{01}(C,i,j)) + \sum_{i=n_0}^{n_0+n_1}\sum_{j=1}^{n_0} \mathrm{I}_{C(i)=0}\, G_{10}(M_{10}(C,i,j))\Big],$$

where m is defined as in (3.11), the functions G as in (4.6), the random vectors M_{ab}s as in theorem 16.

Proof: This is the immediate result of theorem 16 and (3.25).

Theorem 21 *Let $X_i \sim N(\mu_0,\Sigma)$ for $i = 1,\ldots,n_0$, and $X_i \sim N(\mu_1,\Sigma)$ for $i = n_0+1,\ldots,n_0+n_1$ be a set of $n = n_0+n_1$ i.i.d. observations used to derive the classifier in (4.5). Then given a bootstrap vector C, we have:*

$$\mathrm{E}\left[\hat{\varepsilon}_0\varepsilon\right] = \sum_C \frac{P(C)}{m(C)}\Big[m_0(C)(1-\gamma)G_{00}(K_{00}(C)) + m_1(C)\gamma G_{11}(K_{11}(C)) +$$

$$+m_0(C)\gamma G_{01}(K_{01}(C)) + m_1(C)(1-\gamma)G_{10}(K_{10}(C))\Big],$$

where the functions G are defined as in (4.6), the random vectors K_{ab}, $a,b = 0,1$, as in theorem 17, m, m_0, and m_1 in (3.11), $\mathrm{P}(C)$ *in (2.24).*

Proof: This is the immediate result of theorem 17 and (3.26).

D. The Convex Bootstrap Error Estimation

This section first presents the exact formulas for the moments of the true error and the resubstitution estimators. Given that the proofs for these results are similar to that of theorem 15 provided in appendix B, they are omitted here. The theorems to compute the moments of the convex bootstrap estimator and its correlation with the true error are then presented.

1. The Moments of the True Error

a. The First Moment

Under multivariate Gaussian model, (3.20) becomes

$$\mathrm{E}[\varepsilon] = (1-\gamma)\mathrm{G}_0(Z_0(\overrightarrow{1})) + \gamma \mathrm{G}_1(\mathrm{Z}_1(\overrightarrow{1})) \tag{4.9}$$

where Z_0 and Z_1 are defined in theorem 13.

b. The Second Moment

Under multivariate Gaussian model, (3.21) becomes

$$\mathrm{E}[\varepsilon^2] = (1-\gamma)^2\mathrm{G}_{00}(R_{00}) + 2\gamma(1-\gamma)\mathrm{G}_{01}(R_{01}) + \gamma^2\mathrm{G}_{11}(R_{11}) \tag{4.10}$$

where the functions G_{ab}, $a,b = 0,1$ are defined as in (4.6), R_{00}, R_{11}, and R_{01} are 4-dimensional Gaussian random variables with the means and covariance matrices as followings:

$$\begin{aligned} &E[R_{00}^1] = E[R_{00}^3] = E[R_{11}^2] = E[R_{11}^4] = E[R_{01}^1] = E[R_{01}^4] = \\ &\quad = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(1+\frac{1}{4n_0}+\frac{1}{4n_1}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \end{aligned} \tag{4.11}$$

$$\begin{aligned} &E[R_{00}^2] = E[R_{00}^4] = E[R_{11}^1] = E[R_{11}^3] = E[R_{01}^2] = E[R_{01}^3] = \\ &\quad = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(1+\frac{1}{4n_0}+\frac{1}{4n_1}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \end{aligned} \tag{4.12}$$

and

$$\Sigma_{R_{00}} = \Sigma_{R_{11}} = \Sigma_{R_{01}} = \begin{pmatrix} 2(1+\rho_*)I_p & \mathbf{0_{p\times p}} & \nu_1 I_p & \nu_2 I_p \\ . & 2(1-\rho_*)I_p & \nu_3 I_p & \nu_4 I_p \\ . & . & 2(1+\rho_*)I_p & \mathbf{0_{p\times p}} \\ . & . & . & 2(1-\rho_*)I_p \end{pmatrix}, \quad (4.13)$$

where $\rho_* = \frac{n_1-n_0}{\sqrt{n(4n_0n_1+n)}}$, and

$$\nu_1 = \frac{2(n_0-n_1)}{\sqrt{n(4n_0n_1+n)}} + \frac{4n_0n_1+2n}{4n_0n_1+n} \quad (4.14)$$

$$\nu_2 = \nu_3 = \frac{4n_0n_1+2n}{4n_0n_1+n} \quad (4.15)$$

$$\nu_4 = \frac{2(n_1-n_0)}{\sqrt{n(4n_0n_1+n)}} + \frac{4n_0n_1+2n}{4n_0n_1+n} \quad (4.16)$$

2. The Moments of the Resubstitution Estimator

a. The First Moment

$$\begin{aligned} \mathrm{E}\left[\hat{\varepsilon}_r\right] &= \frac{n_0}{n} P\{\psi(X^*) = 1\} + \frac{n_1}{n} P\{\psi(X^{**}) = 0\} \\ &= \frac{n_0}{n} G_0\left(D_0\right) + \frac{n_1}{n} G_1\left(D_1\right), \end{aligned} \quad (4.17)$$

where D_0, D_1 are $2p$-dimensional Gaussian vectors, $D_i(C) = \left[(D_i^1)^T (D_i^2)^T\right]^T$, $i = 0, 1$ with

$$\mathrm{E}\left[D_0^1\right] = \left[\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(1 - \frac{3}{4n_0} + \frac{1}{4n_1}\right)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu, \quad (4.18)$$

$$\mathrm{E}\left[D_0^2\right] = \left[\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(1 - \frac{3}{4n_0} + \frac{1}{4n_1}\right)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu, \quad (4.19)$$

$$\mathrm{E}\left[D_1^2\right] = \left[\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(1 - \frac{3}{4n_1} + \frac{1}{4n_0}\right)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu, \quad (4.20)$$

$$\mathrm{E}\left[D_1^1\right] = \left[\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(1 - \frac{3}{4n_1} + \frac{1}{4n_0}\right)^{-\frac{1}{2}}\right] \Sigma^{-\frac{1}{2}} \mu, \quad (4.21)$$

and

$$\Sigma_{D_i} = \begin{pmatrix} 2(1+\rho_i)I_p & \mathbf{0_{p\times p}} \\ . & 2(1-\rho_i)I_p \end{pmatrix}, \tag{4.22}$$

where $\rho_0 = \frac{n_1-n_0}{\sqrt{n(4n_0n_1-3n_1+n_0)}}$, $\rho_1 = \frac{n_1-n_0}{\sqrt{n(4n_0n_1-3n_0+n_1)}}$.

b. The Second Moment

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_r^2] &= \mathrm{E}\left[\frac{1}{n^2}\left(\sum_{i=1}^{n_0} I_{\psi(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} I_{\psi(X_i)=0}\right)^2\right] \\
&= E\Bigg[\frac{1}{n^2}\Big(\sum_{i=1}^{n_0} I_{\psi(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} I_{\psi(X_i)=0} + \\
&\quad + \sum_{i=1}^{n_0}\sum_{j\neq i}^{n_0} I_{\psi(X_i)=1} I_{\psi(X_j)=1} + \sum_{i=n_0+1}^{n_0+n_1}\sum_{j\neq i}^{n_0+n_1} I_{\psi(X_i)=0} I_{\psi(X_j)=0} + \\
&\quad + \sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n_0+n_1} I_{\psi(X_i)=1} I_{\psi(X_j)=0} + \sum_{i=n_0}^{n_0+n_1}\sum_{j=1}^{n_0} I_{\psi(X_i)=0} I_{\psi(X_j)=1}\Big)\Bigg] \\
&= \frac{n_0}{n^2}\mathrm{P}\{\psi(X_1)=1\} + \frac{n_1}{n^2}\mathrm{P}\{\psi(X_{n_0+1})=0\} + \\
&\quad + \frac{n_0(n_0-1)}{n^2}\mathrm{P}\{\psi(X_1)=1, \psi(X_2)=1\} + \\
&\quad + \frac{n_1(n_1-1)}{n^2}\mathrm{P}\{\psi(X_{n_0+1})=0, \psi(X_{n_0+2})=0\} + \\
&\quad + \frac{2n_0n_1}{n^2}\mathrm{P}\{\psi(X_1)=1, \psi(X_{n_0+1})=0\},
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_r^2] &= \frac{n_0}{n^2}G_0(D_0) + \frac{n_1}{n^2}G_1(D_1) + \frac{n_0(n_0-1)}{n^2}G_{00}(H_{00}) + \\
&\quad + \frac{n_1(n_1-1)}{n^2}G_{11}(H_{11}) + \frac{2n_0n_1}{n^2}G_{01}(H_{01}),
\end{aligned} \tag{4.23}$$

where $H_{ab} = \left[(H_{ab}^1)^T\,(H_{ab}^2)^T\,(H_{ab}^3)^T\,(H_{ab}^4)^T\right]^T$, $a,b = 0,1$ are 4p-dimensional Gaussian vectors, with the means and covariance matrices as followings:

$$\mathrm{E}\left[H_{00}^1\right] = \mathrm{E}\left[H_{00}^3\right] = \mathrm{E}\left[H_{10}^1\right] = \mathrm{E}\left[H_{10}^3\right] = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(1-\frac{3}{4n_0}+\frac{1}{4n_1}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu,$$

$$\mathrm{E}\left[H_{00}^2\right] = \mathrm{E}\left[H_{00}^4\right] = \mathrm{E}\left[H_{10}^2\right] = \mathrm{E}\left[H_{10}^4\right] = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(1-\frac{3}{4n_0}+\frac{1}{4n_1}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu,$$

$$\mathrm{E}\left[H_{11}^1\right] = \mathrm{E}\left[H_{11}^3\right] = \mathrm{E}\left[H_{01}^1\right] = \mathrm{E}\left[H_{01}^3\right] = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(1-\frac{3}{4n_1}+\frac{1}{4n_0}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu,$$

$$\mathrm{E}\left[H_{11}^2\right] = \mathrm{E}\left[H_{11}^4\right] = \mathrm{E}\left[H_{01}^2\right] = \mathrm{E}\left[H_{01}^4\right] = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(1-\frac{3}{4n_1}+\frac{1}{4n_0}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu,$$

$$\Sigma_{H_{ab}} = \begin{pmatrix} 2(1+\rho_a)I_p & \mathbf{0_{p\times p}} & \alpha_{ab1}I_p & \alpha_{ab2}I_p \\ . & 2(1-\rho_a)I_p & \alpha_{ab3}I_p & \alpha_{ab4}I_p \\ . & . & 2(1+\rho_b)I_p & \mathbf{0_{p\times p}} \\ . & . & . & 2(1-\rho_b)I_p \end{pmatrix}, \tag{4.24}$$

where $\rho_0 = \frac{n_1-n_0}{\sqrt{n(4n_0n_1-3n_1+n_0)}}$, $\rho_1 = \frac{n_1-n_0}{\sqrt{n(4n_0n_1-3n_0+n_1)}}$

$$\alpha_{001} = \left(1+2\sqrt{\frac{n_0+n_1}{4n_0n_1-3n_1+n_0}}+\frac{n_0-3n_1}{4n_0n_1-3n_1+n_0}\right), \tag{4.25}$$

$$\alpha_{002} = \alpha_{003} = \frac{4n_0n_1}{4n_0n_1-3n_1+n_0}, \tag{4.26}$$

$$\alpha_{004} = \left(1-2\sqrt{\frac{n_0+n_1}{4n_0n_1-3n_1+n_0}}+\frac{n_0-3n_1}{4n_0n_1-3n_1+n_0}\right), \tag{4.27}$$

$$\alpha_{111} = \left(1-2\sqrt{\frac{n_0+n_1}{4n_0n_1-3n_1+n_0}}+\frac{n_1-3n_0}{4n_0n_1-3n_1+n_0}\right), \tag{4.28}$$

$$\alpha_{112} = \alpha_{113} = \frac{4n_0n_1}{4n_0n_1-3n_0+n_1}, \tag{4.29}$$

$$\alpha_{114} = \left(1+2\sqrt{\frac{n_0+n_1}{4n_0n_1-3n_1+n_0}}+\frac{n_1-3n_0}{4n_0n_1-3n_1+n_0}\right), \tag{4.30}$$

$$\alpha_{011} = \alpha_{014} = \frac{4n_0n_1}{4n_0n_1 - 3n_1 + n_0}, \tag{4.31}$$

$$\alpha_{012} = \left(1 - 2\sqrt{\frac{n_0 + n_1}{4n_0n_1 - 3n_1 + n_0}} + \frac{n_1 + n_0}{4n_0n_1 - 3n_1 + n_0}\right), \tag{4.32}$$

$$\alpha_{013} = \left(1 + 2\sqrt{\frac{n_0 + n_1}{4n_0n_1 - 3n_1 + n_0}} + \frac{n_1 + n_0}{4n_0n_1 - 3n_1 + n_0}\right). \tag{4.33}$$

c. The Cross Moment

$$\begin{aligned}
\mathrm{E}\left[\varepsilon\hat{\varepsilon}_r\right] &= \mathrm{E}\left[\left((1-\gamma)\varepsilon^0 + \gamma\varepsilon^1\right) \times \frac{1}{n}\left(\sum_{i=1}^{n_0} I_{\psi(X_i)=1} + \sum_{i=n_0+1}^{n_0+n_1} I_{\psi(X_i)=0}\right)\right] \\
&= \frac{(1-\gamma)}{n}\sum_{i=1}^{n_0}\mathrm{E}\left[\varepsilon^0 I_{\psi(X_i)=1}\right] + \frac{(1-\gamma)}{n}\sum_{i=n_0+1}^{n_0+n_1}\mathrm{E}\left[\varepsilon^0 I_{\psi(X_i)=1}\right] + \\
&\quad + \frac{\gamma}{n}\sum_{i=1}^{n_0}\mathrm{E}\left[\varepsilon^1 I_{\psi(X_i)=0}\right] + \frac{\gamma}{n}\sum_{i=n_0+1}^{n_0+n_1}\mathrm{E}\left[\varepsilon^1 I_{\psi(X_i)=0}\right] \\
&= \frac{1-\gamma}{n}\Big(n_0 P\{\psi(X^*) = 1, \psi(X_1) = 1\} + n_1 P\{\psi(X^*) = 1, \psi(X_{n_0+1}) = 0\}\Big) + \\
&\quad + \frac{\gamma}{n}\Big(n_0 P\{\psi(X^{**}) = 0, \psi(X_1) = 1\} + n_1 P\{\psi(X^{**}) = 0, \psi(X_{n_0+1}) = 0\}\Big),
\end{aligned}$$

So,

$$\mathrm{E}\left[\hat{\varepsilon}_r\varepsilon\right] = \frac{(1-\gamma)n_0}{n}G_{00}(J_{00}) + \frac{\gamma n_1}{n}G_{11}(J_{11}) + \frac{\gamma n_0}{n}G_{01}(J_{01}) + \frac{(1-\gamma)n_1}{n}G_{10}(J_{10}). \tag{4.34}$$

where $J_{ab} = \left[(J^1_{ab})^T\,(J^2_{ab})^T\,(J^3_{ab})^T\,(J^4_{ab})^T\right]^T$, $a,b = 0,1$ are 4p-dimensional Gaussian vectors with the means and covariance matrices as followings:

$$E[J_{00}^1] = E[J_{01}^1] = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(1-\frac{3}{4n_0}+\frac{1}{4n_1}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \tag{4.35}$$

$$E[J_{00}^2] = E[J_{01}^2] = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(1-\frac{3}{4n_0}+\frac{1}{4n_1}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \tag{4.36}$$

$$E[J_{11}^1] = E[J_{10}^1] = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(1-\frac{3}{4n_1}+\frac{1}{4n_0}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \tag{4.37}$$

$$E[J_{11}^2] = E[J_{10}^2] = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(1-\frac{3}{4n_1}+\frac{1}{4n_0}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \tag{4.38}$$

$$E[J_{00}^3] = E[J_{10}^3] = E[J_{11}^4] = E[J_{01}^4] = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} + \left(1+\frac{1}{4n_0}+\frac{1}{4n_1}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \tag{4.39}$$

$$E[J_{00}^4] = E[J_{10}^4] = E[J_{11}^3] = E[J_{01}^3] = \left[\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} - \left(1+\frac{1}{4n_0}+\frac{1}{4n_1}\right)^{-\frac{1}{2}}\right]\Sigma^{-\frac{1}{2}}\mu, \tag{4.40}$$

and

$$\Sigma_{J_{ab}} = \begin{pmatrix} 2(1+\rho_a)I_p & \mathbf{0_{p\times p}} & \beta_{ab1}I_p & \beta_{ab2}I_p \\ . & 2(1-\rho_a)I_p & \beta_{ab3}I_p & \beta_{ab4}I_p \\ . & . & 2(1+\rho_*)I_p & \mathbf{0_{p\times p}} \\ . & . & . & 2(1-\rho_*)I_p \end{pmatrix}. \tag{4.41}$$

where $\rho_* = \frac{n_1-n_0}{\sqrt{n(n+4n_0n_1)}}$, $\rho_0 = \sqrt{\frac{n}{4n_0n_1-3n_1+n_0}}$, $\rho_1 = \sqrt{\frac{n}{4n_0n_1-3n_0+n_1}}$, and

$$\beta_{001} = \beta_{011} = \left(\sqrt{n}+\frac{n_0-n_1}{\sqrt{4n_0n_1+n}}\right)\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{4n_0n_1-3n_1+n_0}}\right), \tag{4.42}$$

$$\beta_{002} = \beta_{012} = \left(\sqrt{n}-\frac{n_0-n_1}{\sqrt{4n_0n_1+n}}\right)\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{4n_0n_1-3n_1+n_0}}\right), \tag{4.43}$$

$$\beta_{003} = \beta_{013} = \left(\sqrt{n}+\frac{n_0-n_1}{\sqrt{4n_0n_1+n}}\right)\left(\frac{1}{\sqrt{n}}-\frac{1}{\sqrt{4n_0n_1-3n_1+n_0}}\right), \tag{4.44}$$

$$\beta_{004} = \beta_{014} = \left(\sqrt{n}-\frac{n_0-n_1}{\sqrt{4n_0n_1+n}}\right)\left(\frac{1}{\sqrt{n}}-\frac{1}{\sqrt{4n_0n_1-3n_1+n_0}}\right), \tag{4.45}$$

$$\beta_{111} = \beta_{101} = \left(\sqrt{n} + \frac{n_0 - n_1}{\sqrt{4n_0n_1 + n}}\right)\left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{4n_0n_1 - 3n_0 + n_1}}\right), \tag{4.46}$$

$$\beta_{112} = \beta_{102} = \left(\sqrt{n} - \frac{n_0 - n_1}{\sqrt{4n_0n_1 + n}}\right)\left(\frac{1}{\sqrt{n}} - \frac{1}{\sqrt{4n_0n_1 - 3n_0 + n_1}}\right), \tag{4.47}$$

$$\beta_{113} = \beta_{103} = \left(\sqrt{n} + \frac{n_0 - n_1}{\sqrt{4n_0n_1 + n}}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{4n_0n_1 - 3n_0 + n_1}}\right), \tag{4.48}$$

$$\beta_{114} = \beta_{104} = \left(\sqrt{n} - \frac{n_0 - n_1}{\sqrt{4n_0n_1 + n}}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{4n_0n_1 - 3n_0 + n_1}}\right). \tag{4.49}$$

3. The Moments of the Convex Estimator

Theorem 22 *Let $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0 + 1, \ldots, n_0 + n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (4.5). Then given a bootstrap vector C, we have:*

$$\begin{aligned} \mathrm{E}\left[\hat{\varepsilon}_w\right] = & \frac{n_0(1-w)}{n} G_0(D_0) + \frac{n_1(1-w)}{n} G_1(D_1) + \\ & + \sum_C \frac{w\mathrm{P}(C)}{m(C)} \Big(m_0(C) G_0\left(Z_0(C)\right) + m_1(C) G_1\left(Z_1(C)\right) \Big), \end{aligned}$$

where the functions G_0 and G_1 are defined as in (4.6), the random variables $Z_0(C)$ and $Z_1(C)$ in theorem 13, D_0 and D_1 in (4.17), m_0 and m_1 in (3.11), $\mathrm{P}(C)$ *in (2.24).*

Proof: This is the immediate result of theorem 13 and (3.27).

Theorem 23 *Let $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0+1, \ldots, n_0+n_1$ be a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (4.5). Then given a bootstrap vector C, we have:*

$$
\begin{aligned}
&\mathrm{E}\left[\hat{\varepsilon}_w^2\right] = \\
&= (1-w)^2 \left[\frac{n_0}{n^2} G_0\left(D_0\right) + \frac{n_1}{n^2} G_1\left(D_1\right) + \right. \\
&\left. + \frac{n_0(n_0-1)}{n^2} G_{00}(H_{00}) + \frac{n_1(n_1-1)}{n^2} G_{11}(H_{11}) + \frac{2n_0 n_1}{n^2} G_{01}(H_{01}) \right] + \\
&+ \sum_C \frac{2w(1-w)\mathrm{P}(C)}{nm(C)} \left[\sum_{i,j=1}^{n_0} I_{C(i)=0}\, G_{00}(M_{00}(C,i,j)) + \sum_{i,j=n_0+1}^{n_0+n_1} I_{C(i)=0}\, G_{11}(M_{11}(C,i,j)) + \right. \\
&\left. + \sum_{i=1}^{n_0} \sum_{j=n_0}^{n_0+n_1} I_{C(i)=0}\, G_{01}(M_{01}(C,i,j)) + \sum_{i=n_0}^{n_0+n_1} \sum_{j=1}^{n_0} I_{C(i)=0}\, G_{10}(M_{10}(C,i,j)) \right] + \\
&+ \sum_C \frac{w^2 \mathrm{P}(C)}{m^2(C)} \left[m_0(C) G_0\left(Z_0(C)\right) + m_1(C) G_1\left(Z_1(C)\right) + \right. \\
&+ \sum_{i=1}^{n_0} \sum_{j \neq i}^{n_0} I_{C(i)=0} I_{C(j)=0} G_{00}\left(T_{00}(C,i,j)\right) + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j \neq i}^{n_0+n_1} I_{C(i)=0} I_{C(j)=0} G_{11}\left(T_{11}(C,i,j)\right) + \\
&\left. + \sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n_0+n_1} I_{C(i)=0} I_{C(j)=0} G_{01}\left(T_{01}(C,i,j)\right) + \sum_{i=n_0+1}^{n_0+n_1} \sum_{j=1}^{n_0} I_{C(i)=0} I_{C(j)=0} G_{10}\left(T_{01}(C,j,i)\right) \right] + \\
&+ \sum_{C_1 \neq C_2} \frac{w^2}{m(C_1)m(C_2)} \left[\sum_{i,j=1}^{n_0} I_{C_1(i)=0} I_{C_2(j)=0}\, G_{00}\left(F_{00}(C_1,C_2,i,j)\right) + \right. \\
&+ \sum_{i,j=n_0+1}^{n_0+n_1} I_{C_1(i)=0} I_{C_2(j)=0}\, G_{11}\left(F_{11}(C_1,C_2,i,j)\right) + \\
&+ \sum_{i=1}^{n_0} \sum_{j=n_0}^{n_0+n_1} I_{C_1(i)=0} I_{C_2(j)=0}\, G_{01}\left(F_{01}(C_1,C_2,i,j)\right) + \\
&\left. + \sum_{i=n_0}^{n_0+n_1} \sum_{j=1}^{n_0} I_{C_1(i)=0} I_{C_2(j)=0}\, G_{10}\left(F_{01}(C_2,C_1,j,i)\right) \right],
\end{aligned}
$$

where the functions G are defined as in (4.6), the random variables Z_0 and Z_1 in theorem 13, D_0 and D_1 in (4.17), T_{ab}, F_{ab}, $a,b = 0,1$ in theorem 14 and 15, respectively, m_0 and m_1 in (3.11), $\mathrm{P}(C)$ *in (2.24).*

Proof: This is the immediate result of theorem 19, theorem 20, and (3.28).

Theorem 24 *Let $X_i \sim N(\mu_0,\Sigma)$ for $i = 1,\dots,n_0$, and $X_i \sim N(\mu_1,\Sigma)$ for $i = n_0+1,\dots,n_0+n_1$ be a set of $n = n_0+n_1$ i.i.d. observations used to derive the classifier in (4.5). Then given a bootstrap vector C, we have:*

$$\begin{aligned}\mathrm{E}[\hat{\varepsilon}_w\varepsilon] = {} & (1-w)\Big[\frac{(1-\gamma)n_0}{n}G_{00}(J_{00})+\frac{\gamma n_1}{n}G_{11}(J_{11})+\frac{\gamma n_0}{n}G_{01}(J_{01})+\frac{(1-\gamma)n_1}{n}G_{10}(J_{10})\Big]+\\ & +\sum_C\frac{wP(C)}{nm(C)}\Big[m_0(C)(1-\gamma)G_{00}(K_{00}(C))+m_1(C)\gamma G_{11}(K_{11}(C))+\\ & +m_0(C)\gamma G_{01}(K_{01}(C))+m_1(C)(1-\gamma)G_{10}(K_{10}(C))\Big],\end{aligned}$$

where the functions G *are defined in (4.6), the random variables J_{ab} and K_{ab} as in (4.34) and theorem 17, respectively, m as in (3.11).*

Proof: This is the immediate result of theorem 21, and (3.29).

E. The .632 Bootstrap Error Estimation

Similarly to the univariate case, setting $w = .632$ in the formulas of the convex estimator yields the moments of the classic .632 bootstrap estimate.

F. The Optimal Bootstrap Error Estimation

The above theorems allow one to compute the optimal weight w^*, which minimizes the root mean square of the deviation of the convex bootstrap estimate $\hat{\varepsilon}_w$ from the true error ε.

$$w^* = \arg\min_w \mathrm{RMS}[\hat{\varepsilon}_w] = \arg\min_w \mathrm{RMS}^2[\hat{\varepsilon}_w]$$

$$
\begin{aligned}
\mathrm{RMS}^2[\hat{\varepsilon}_w] &= \mathrm{E}\left(\hat{\varepsilon}_w - \varepsilon\right)^2 \\
&= \mathrm{E}\left[\hat{\varepsilon}_w^2\right] - 2\mathrm{E}\left[\hat{\varepsilon}_w \varepsilon\right] + \mathrm{E}\left[\varepsilon^2\right] \\
&= (1-w)^2\mathrm{E}[\hat{\varepsilon}_r^2] + w^2\mathrm{E}[\hat{\varepsilon}_0^2] + 2w(1-w)\mathrm{E}[\hat{\varepsilon}_r\hat{\varepsilon}_0] + \\
&\quad - 2\left((1-w)\mathrm{E}[\varepsilon\hat{\varepsilon}_r] + w\mathrm{E}[\varepsilon\hat{\varepsilon}_0]\right) + \mathrm{E}[\varepsilon^2] \\
&= w^2\mathrm{E}\left(\hat{\varepsilon}_r - \hat{\varepsilon}_0\right)^2 + 2\mathrm{E}\left[-\varepsilon\hat{\varepsilon}_0 + \hat{\varepsilon}_r\hat{\varepsilon}_0 + \varepsilon\hat{\varepsilon}_r - \hat{\varepsilon}_r^2\right] w + \mathrm{E}\left(\hat{\varepsilon}_r - \varepsilon\right)^2
\end{aligned}
$$

The root mean square of the convex estimator $\mathrm{RMS}^2[\hat{\varepsilon}_w]$ is a quadratic function of w. Thus, w^* can be found to be

$$
\begin{aligned}
w^* &= -\frac{2\mathrm{E}\left[-\varepsilon\hat{\varepsilon}_0 + \hat{\varepsilon}_r\hat{\varepsilon}_0 + \varepsilon\hat{\varepsilon}_r - \hat{\varepsilon}_r^2\right]}{2\mathrm{E}\left(\hat{\varepsilon}_r - \hat{\varepsilon}_0\right)^2} \\
&= \frac{\mathrm{E}\left[\varepsilon\hat{\varepsilon}_0 - \hat{\varepsilon}_r\hat{\varepsilon}_0 - \varepsilon\hat{\varepsilon}_r + \hat{\varepsilon}_r^2\right]}{\mathrm{E}\left[\hat{\varepsilon}_r^2 - 2\hat{\varepsilon}_r\hat{\varepsilon}_0 + \hat{\varepsilon}_0^2\right]}
\end{aligned}
$$

In principle for the multivariate case, the optimal minimum RMS w^* can be computed using Theorem 19, 20, 21, and the results of (4.23), (4.34). In .632 bootstrap estimation, the combination scalar .632 was chosen heuristically, which represents the proportion of the original sample points in the bootstrap samples. In .632+ bootstrap estimation, w was chosen heuristically adaptively in accordance with the *overfitting rate*. While both of them have been shown to be among the best, they do not guarantee the minimum root mean square.

G. The Unbiased Bootstrap Error Estimation

By Theorem 10, (4.17) and and (4.9), the unbiased bootstrap scalar w_u for the multivariate case can be found similarly using (3.84). An issue that arises in the multivariate case is the computation of the probabilities in (4.17), (4.9), and (4.8). This computation is very difficult since it involves the ratio of noncentral chi-square random variables, which has a doubly noncentral F distribution. Computation of this distribution is a hard problem.

Moran proposes in [143] a complex procedure, based on work by Price [144], to compute this probability, which only applies to even dimensionality p. To compute (4.8), we employ an accurate approximation, based on the use of the Imhof-Pearson three-moment method [145]. This consists of approximating a non-central $\chi_p^2(\lambda)$ random variable with a central χ_h^2 random variable, by equating the first three moments of their distributions. This approach, which was originally employed in [146], is not restricted to even dimensionality p. For example,

$$P\left(\frac{W_1}{W_2} > \frac{1-\rho_e}{1+\rho_e}\right) \simeq P(\chi_h^2 < y)\,, \tag{4.50}$$

where W_1 and W_2 are two independent noncentral χ_p^2 with non-centrality parameters λ_1 and λ_2, repectively, and χ_h^2 is a central chi-square random variable with h degrees of freedom, with

$$\begin{aligned} h &= \frac{c_2^3}{c_3^2}\,, \\ y &= h + c_1\sqrt{\frac{h}{c_2}}\,, \end{aligned} \tag{4.51}$$

and

$$c_i = \left(\frac{1+\rho_e}{2}\right)^i (p+i\lambda_1) + \left(\frac{1-\rho_e}{2}\right)^i (p+i\lambda_2)\,, \quad i = 1,2,3\,. \tag{4.52}$$

The approximation is valid only for $c_3 > 0$ [145]. However, since $\lambda_1, \lambda_2 \geq 0, -1 \leq \rho_e \leq 1$, so it is always the case that $c_3 > 0$ and the approximation applies. The same approximation applies to (4.17), (4.9), and (4.8) by substituting the appropriate values.

Figure 2 is as Figure 1, but displays the multivariate case, with $p = 2$. The plots are exact save for the accurate Imhof-Pearson approximation described in the previous section. Some of the same behavior observed in the univariate case is seen here. Unlike the univariate case, here the unbiased weight can be quite far from the heuristic 0.632 weight, even for small Bayes error.

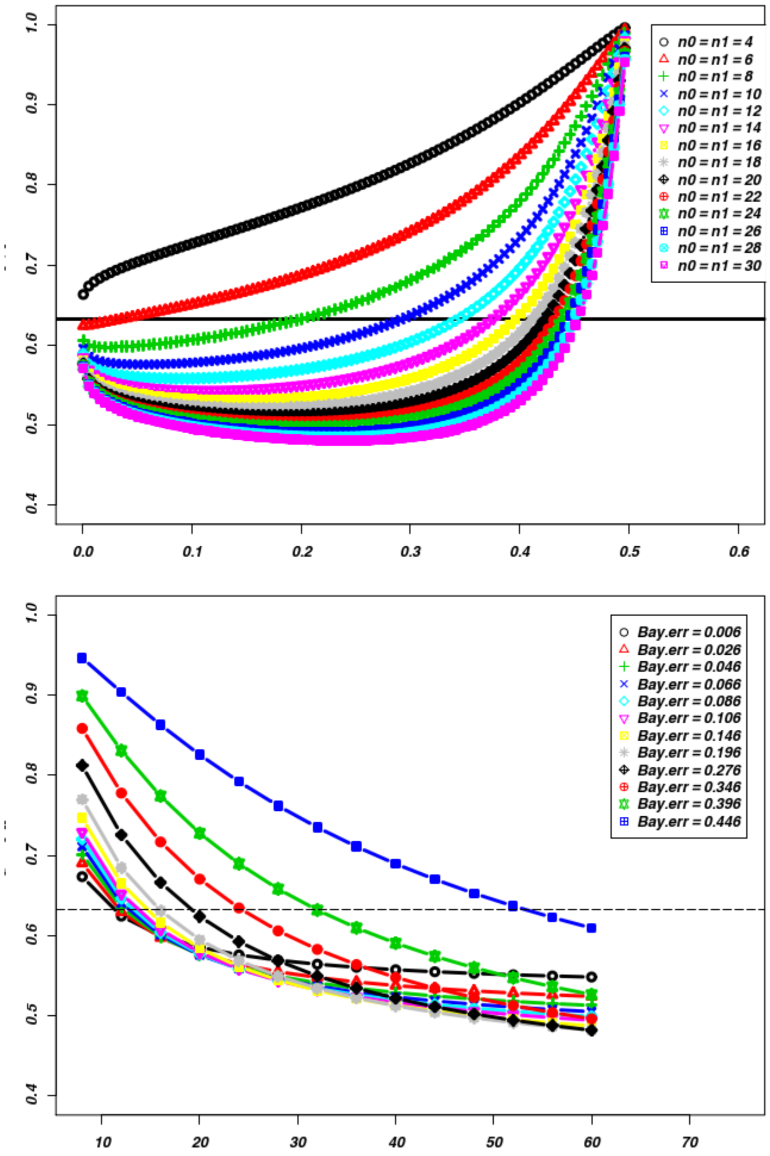


Fig. 2. Optimal weight w_u in the multivariate case, $p = 2$. The top figure displays w_u as a function of Bayes error for different sample sizes, whereas the bottom figure displays w_u as a function of the number of samples for different Bayes errors.

CHAPTER V

SMALL-SAMPLE PERFORMANCE OF BAGGING CLASSIFICATION RULES *

There has been considerable interest recently in the application of bagging in the classification of both gene-expression data and protein-abundance mass spectrometry data. The approach is often justified by the improvement it produces on the performance of unstable, overfitting classification rules under small-sample situations. However, the question of real practical interest is whether the ensemble scheme will improve performance of those classifiers sufficiently to beat the performance of single stable, non-overfitting classifiers, in the case of small-sample genomic and proteomic data sets. To investigate that question, we conducted a detailed empirical study, using publicly-available data sets from published genomic and proteomic studies. We observed that, under t-test and RELIEF filter-based feature selection, bagging generally does a good job of improving the performance of unstable, overfitting classifiers, such as CART decision trees and neural networks, but that improvement was not sufficient to beat the performance of single stable, non-overfitting classifiers, such as diagonal and plain linear discriminant analysis, or 3-nearest neighbors. Furthermore, as expected, the ensemble method did not improve the performance of these classifiers significantly. Representative experimental results are presented and discussed here, whereas the full results of the empirical study are available on a companion website http://www.ece.tamu.edu/∼ulisses/bagging/index.html.

* Reprinted with permission from "Is Bagging Effective in the Classification of Small-sample Genomic and Proteomic Data?" by T. T. Vu and U. M. Braga-neto, 2009. volume 2009, p.1–10, Copyright 2009 of *EURASIP Journal on Bioinformatics and Systems Biology.*

A. Introduction

Randomized ensemble methods for classifier design combine the decision of an ensemble of classifiers designed on randomly perturbed versions of the available data [147, 148, 149, 150, 151]. The combination is often done by means of majority-voting among the individual classifier decisions [150, 152, 151], whereas the data perturbation usually employs the bootstrap resampling approach, which corresponds to sampling uniformly with replacement from the original data [96, 153]. The combination of bootstrap resampling and majority-voting is known as bootstrap aggregate or *bagging* [150, 151].

There has been considerable interest recently in the application of bagging in the classification of both gene-expression data [154, 155, 156, 157] and protein-abundance mass spectrometry data [158, 159, 160, 161, 162, 163]. However, there is scant theoretical justification for the use of this heuristic, other than the expectation that combining the decision of several classifiers will regularize and improve the performance of unstable, overfitting classification rules, such asunpruned decision trees, provided one uses a large enough number of classifiers in the ensemble [150, 151]. It is also claimed that ensemble rules "do not overfit", meaning that classification error converges as the number of component classifiers tends to infinity [151].

However, the main performance issue is not whether the ensemble scheme improves the classification error of a single unstable, overfitting classifier, or whether its classification error converges to a fixed limit; these are important questions, which have been studied in the literature (in particular when the component classifiers are decision trees) [164, 165, 151, 166, 167, 168], but the question of main practical interest is whether the ensemble scheme will improve the performance of unstable, overfitting classifiers *sufficiently* to beat the performance of single stable, non-overfitting classifiers, particularly in small-sample settings. Therefore, there is a pressing need to examine rigorously the suitability and validity of the ensemble approach in the classification of small-sample genomic and proteomic data. In this chapter, we present results from a comprehensive empirical study concerning the effect of bagging on the performance of several classification rules, including diagonal and plain linear discriminant analysis, 3-nearest neighbors, CART decision trees, and neural networks, using real data from published microarray and mass spectrometry studies. Here we are concerned exclusively with the performance in terms of the true classification error, and therefore we employ filter-based feature selection and holdout estimation based on large samples in order to allow accurate classification error estimation. Similar studies recently published [156, 157] rely on small-sample wrapper feature selection and small-sample error estimation methods, which will obscure the issue of how bagging really affects the true classification error. In particular, there is evidence that filter-based feature selection outperforms wrapper feature selection in small sample settings [169]. In our experiments, we employ the one-tailed paired t-test to assess whether the expected true classification error is significantly smaller for the bagged classifier as opposed to the original base classifier, under different number of samples, dimensionality, and number of classifiers in the ensemble. Clearly, the heuristic is beneficial for the particular classification rule if and only there is a significant decrease in expected classification error, otherwise the procedure is to be avoided; however the magnitude of improvement is also a factor — a small improvement in performance may not be worth it the extra computation required (which is roughly m times larger for the bagging classifier, where m is the number of classifiers in the ensemble).

B. Randomized Ensemble Classification Rules

Randomization approaches based on resampling can be seen as drawing i.i.d. samples $S_k^* = \{(X_1^*, Y_1^*), (X_2^*, Y_2^*), \ldots, (X_k^*, Y_k^*)\}$ from a surrogate joint-feature label distribution F^*, which is a function of the original training data S_n. In the bootstrap resampling approach, one has $k = n$, and the randomized sample S_n^* corresponds to sampling uniformly n training points from S_n *with* replacement. This corresponds to using the *empirical distribution* of the data S_n as the surrogate joint-feature label distribution F^*; the empirical distribution assigns discrete probability mass $\frac{1}{n}$ at each observed data point in S_n. Some of the original training points may appear multiple times, whereas others may not appear at all in the *bootstrap sample* S_n^*. Note that, given S_n, the bootstrap sample S_n^* is conditionally independent from the original feature-label distribution F.

In aggregation by majority voting, a classifier is obtained based on majority voting among individual classifiers designed on the randomized samples S_k^* using the original classification rule Ψ_n. This leads to an *ensemble classification rule* Ψ_n^R, such that

$$\psi_n^R(x) = \Psi_n^R(S_n)(x) = \begin{cases} 1\,, & E[\Psi_n(S_k^*)(x) \mid S_n] > \frac{1}{2} \\ 0\,, & \text{otherwise} \end{cases} \tag{5.1}$$

for $x \in V$, where expectation is with respect to the random mechanism F^*, fixed at the observed value of S_n. For bootstrap majority voting, or bagging, the expectation in (5.1) usually has to be approximated by Monte-Carlo sampling, which leads to the "bagged" classifier:

$$\psi_{n,m}^B(x) = \begin{cases} 1\,, & \frac{1}{m}\sum_{j=1}^m \psi_n^{*(j)}(x) > \frac{1}{2} \\ 0\,, & \text{otherwise} \end{cases} \tag{5.2}$$

where the classifiers $\psi_n^{*(j)}$ are designed by the original classification rule Ψ_n on bootstrap samples $S_n^{*(j)}$, for $j = 1, \ldots, m$, for large enough m (notice the parallel with the development in [99], particularly eqs. (2.8)–(2.10) and accompanying discussion).

The issue of how large m has to be so that (5.2) is a good Monte-Carlo approximation is a critical issue in the application of bagging. Note that m represents the number of classifiers that must be designed to be part of the ensemble, so that a computational problem may emerge if *m* is made too large. In addition, even if a suitable *m* is found, the performance of the ensemble must be compared to that of the base classification rule, to see if there is significant improvement. Even more importantly, the performance of the ensemble has to compared to that of other classification rules; that the ensemble improves the performance of an unstable, overfitting classifier is of small value if it can be bested by a single stable, non-overfitting classifier. In the next section, we present a comprehensive empirical study that addresses these questions.

C. Experimental Study

In this section, we report the results obtained from a large simulation study based on publicly-available patient data from genomic and proteomic studies, which measured the performance of the bagging heuristic through the expected classification error, for varying number of component classifiers, sample size, and dimensionality.

1. Methods

We considered in our experiment several classification rules, listed here in order of complexity: diagonal linear discriminant analysis (DLDA), linear discriminant analysis (LDA), 3-nearest-neighbors (3NN), decision trees (CART), and neural networks (NNET) [61, 170]. DLDA is an extension of LDA where only the diagonal elements (the variances) of the covariance matrix are estimated, while the off-diagonal elements (the covariances) are assumed to be zero. Bagging is applied to each of these base classification rules and its performance recorded for varying number of individual classifiers. The neural network consists of a one-hidden layer with 4 nodes and standard sigmoids as nonlinearities. The network is trained by Levenberg-Marquardt optimization with a maximum of 30 iterations.

CART is applied with a stopping criterion: splitting is stopped when there are fewer than 3 points in a given node. This is distinct from the approach advocated in [151] for random forests, where unpruned, fully-grown trees are used instead; the reason for this is that we did not attempt to implement the approach in [151] (which involves concepts as random node splitting and is thus specific to decision trees), but rather to study the behavior of bagging, which is the centerpiece of such ensemble methods, across different classification rules. Resampling is done by means of *balanced* bootstrapping, where all samples are made to appear exactly the same number of times in the computation [171].

We selected data sets with large number N of samples (see below) in order to be able to estimate the true error accurately using held out testing data. In each case, 1000 training data sets of size $n = 20, 40,$ and 60 were drawn uniformly and independently from the total pool of N samples. The training data are drawn in a stratified fashion, following the approximate proportion of each class in the original data. Based on the training data, a filter-based gene selection step is employed to select the top p discriminating genes; we considered in this study $p = 2, 3, 5, 8$. The univariate feature selection methods used in the filter step are the Welch two-sample t-test [172] and the RELIEF method [173] — in the latter case, we employ the 1-nearest-neighbor method when searching for hits and misses. After classifier design, the true classification error for each data set of size n is approximated by a holdout estimator, whereby the $N - n$ sample points not drawn are used as the test set (a good approximation to the classification error, given that $N >> n$). The expected classification error is then estimated as the sample mean of classification error over the 1000 training data sets. The sample size n is kept small, as we are interested in the small-sample properties of bagging. Note also that we also must have $N >> n$ in order to provide for large enough testing sets, as well as to make sure that consecutive training sets do not significantly overlap, so that the expected classification error can be accurately approximated. As can be easily verified, the expected ratio of overlapping sample points between two samples of size n from a population of size N is given simply by n/N. In all cases considered here the expected overlap is around 20% less, which we consider to be acceptable, except in the case of the lung cancer data set with $n = 60$. This latter case is therefore not included in our results. An unpaired one-tailed t-test is employed to assess whether the ensemble classifier has an expected error that is significantly smaller than that of the corresponding individual classifier.

2. Data Sets

We utilized the following publicly-available data sets from published studies in order to study the performance of bagging in the context of genomics and proteomics applications.

- **Breast Cancer Gene Expression Data.** These data come from the breast cancer classification study in [174], which analyzed $N = 295$ gene-expression microarrays containing a total of 25760 transcripts each. Filter-based feature selection was performed on a 70-gene prognosis profile, previously published by the same authors in [175]. Classification is between the good-prognosis class (115 samples), and the poor-prognosis class (180 samples), where prognosis is determined retrospectively in terms of survivability [174].

- **Lung Cancer Gene Expression Data.** We employed here the data set "A" from the study in [176] on non-small-cell lung carcinomas (NSCLC), which analyzed $N = 186$ gene-expression microarrays containing a total of 12600 transcripts each. NSCLC is subclassified as adenocarcinomas, squamous cell carcinomas and large-cell carcinomas , of which adenocarcinomas are the most common subtypes and of interest to classify from other subtypes of NSCLC. Classification is thus between adenocarcinomas (139 samples) and non-adenocarcinomas (47 samples).

- **Prostrate Cancer Protein Abundance Data.** Given the recent keen interest on deriving serum-based proteomic biomarkers for the diagnosis of cancer [177], we also included in this study data from a proteomic study of prostate cancer reported in [178]. It consists of SELDI-TOF mass spectrometry of $N = 326$ samples, which yield mass spectra for 45,000 n/z (mass over charge) values. Filter-based feature selection is employed to find the top discriminatory n/z values to be used in the experiment. Classification is between prostate cancer patients (167 samples) and non-prostate patients, including benign prostatic hyperplasia and healthy patients (159 samples). We use the raw spectra values, without baseline subtraction, as we found that this leads to better classification rates.

3. Results and Discussion

We present results for sample sizes $n = 20$ and $n = 40$ and dimensionality $p = 2$ and $p = 5$, which are representative of the full set of results, available on the companion website http://www.ece.tamu.edu/∼ulisses/bagging/index.html. The case $p = 2$ is displayed in Tables 1–3, each of which corresponds to a different data set. Each table displays the expected classification error as a function of the number m of classifiers used in the ensemble, for different base classification rules, feature selection methods, and sample sizes. We used in all cases an odd number m of classifiers in the ensembles, to avoid tie-breaking issues. Errors that are smaller for the ensemble classifier as compared to a single classifier at a 99% significance level, according to a one-tailed paired t-test, are indicated by bold-face type. This allows one to immediately observe that bagging is able to improve the performance of the unstable, overfitting CART and NNET classifiers; in most cases, a small ensemble is required, and the improvement in performance is substantial. In contrast, bagging does not improve the performance of the stable, non-overfitting DLDA, LDA, and 3NN classifiers, except via a large ensemble; and even so the improvement in magnitude is quite small, and

Table I. Expected classification error of selected experiments with breast cancer gene expression data (full results available on the companion website). Bold-face type indicates the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test.

Rule	p	n	Single	$m=5$	$m=11$	$m=15$	$m=21$	$m=25$	$m=31$	$m=35$	$m=41$	$m=45$	$m=51$
LDA	2	20	0.212	0.237	0.224	0.220	0.217	0.217	0.216	0.216	0.215	0.215	0.214
LDA	2	40	0.204	0.217	0.209	0.208	0.207	0.206	0.206	0.206	0.205	0.205	0.205
LDA	2	60	0.203	0.212	0.207	0.205	0.205	0.204	0.204	0.204	0.204	0.204	0.204
LDA	5	20	0.240	0.285	0.261	0.255	0.251	0.249	0.247	0.248	0.246	0.246	0.245
LDA	5	40	0.207	0.233	0.219	0.216	0.213	0.212	0.212	0.211	0.211	0.210	0.210
LDA	5	60	0.196	0.216	0.205	0.203	0.201	0.201	0.200	0.199	0.199	0.199	0.199
3NN	2	20	0.230	0.281	0.246	0.241	0.235	0.234	0.231	0.231	0.230	0.229	**0.229**
3NN	2	40	0.228	0.274	0.241	0.235	0.231	0.229	0.228	0.227	**0.226**	**0.226**	**0.225**
3NN	2	60	0.225	0.269	0.238	0.232	0.228	0.227	0.225	**0.224**	**0.224**	**0.223**	**0.222**
3NN	5	20	0.220	0.270	0.235	0.229	0.224	0.223	0.221	0.220	0.219	**0.219**	**0.219**
3NN	5	40	0.217	0.262	0.229	0.224	0.220	0.219	0.217	0.216	**0.216**	**0.215**	**0.215**
3NN	5	60	0.219	0.261	0.230	0.225	0.221	0.220	0.219	0.218	**0.217**	**0.217**	**0.216**
CART	2	20	0.259	0.297	0.263	0.256	**0.250**	**0.247**	**0.246**	**0.244**	**0.243**	**0.242**	**0.242**
CART	2	40	0.257	0.294	0.258	**0.252**	**0.245**	**0.244**	**0.242**	**0.240**	**0.239**	**0.239**	**0.237**
CART	2	60	0.255	0.287	0.256	**0.249**	**0.243**	**0.241**	**0.237**	**0.236**	**0.235**	**0.234**	**0.234**
CART	5	20	0.261	0.291	**0.257**	**0.248**	**0.240**	**0.238**	**0.235**	**0.235**	**0.233**	**0.232**	**0.231**
CART	5	40	0.260	0.287	**0.249**	**0.240**	**0.233**	**0.231**	**0.228**	**0.226**	**0.225**	**0.224**	**0.223**
CART	5	60	0.262	0.290	**0.248**	**0.240**	**0.232**	**0.229**	**0.226**	**0.225**	**0.223**	**0.222**	**0.221**
NNET	2	20	0.252	0.293	**0.246**	**0.240**	**0.230**	**0.230**	**0.225**	**0.224**	**0.223**	**0.222**	**0.221**
NNET	2	40	0.226	0.256	0.225	**0.219**	**0.215**	**0.213**	**0.212**	**0.210**	**0.210**	**0.209**	**0.209**
NNET	2	60	0.216	0.241	0.216	**0.211**	**0.208**	**0.206**	**0.204**	**0.204**	**0.203**	**0.203**	**0.203**
NNET	5	20	0.282	0.321	**0.265**	**0.250**	**0.242**	**0.239**	**0.235**	**0.233**	**0.231**	**0.230**	**0.229**
NNET	5	40	0.253	0.286	**0.238**	**0.228**	**0.221**	**0.218**	**0.215**	**0.213**	**0.212**	**0.210**	**0.209**
NNET	5	60	0.236	0.268	**0.226**	**0.218**	**0.212**	**0.210**	**0.208**	**0.206**	**0.205**	**0.204**	**0.204**

certainly does not justify the extra computational cost (note that in the case of the simplest classification rule, DLDA, there is no improvement at all). This is in agreement with what is known about the ensemble approach (e.g., see [151]).

However, of larger interest here is the performance of the ensemble against a single instance of the stable, non-overfitting classifiers. This can be better visualized in the plots of Figures 3–5, which display the expected classification errors as a function of number of component classifiers in the ensemble, for the case $p = 5$. The error of a single classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for

Table II. Expected classification error of selected experiments with the lung cancer gene expression data (full results available on the companion website). Bold-face type indicates the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test.

Rule	p	n	Single	$m=5$	$m=11$	$m=15$	$m=21$	$m=25$	$m=31$	$m=35$	$m=41$	$m=45$	$m=51$
LDA	2	20	0.201	0.206	0.203	0.203	0.202	0.202	0.203	0.202	0.202	0.202	0.203
LDA	2	40	0.192	0.194	0.193	0.193	0.193	0.193	0.192	0.192	0.193	0.192	0.192
LDA	2	60	0.190	0.191	0.190	0.190	0.190	0.190	0.190	0.190	0.190	0.190	0.190
LDA	5	20	0.227	0.241	0.232	0.231	0.230	0.228	0.228	0.227	0.228	0.227	0.227
LDA	5	40	0.200	0.205	0.202	0.201	0.200	0.200	0.200	0.200	0.200	0.200	0.200
LDA	5	60	0.194	0.197	0.196	0.195	0.194	0.194	0.194	0.194	0.194	0.194	0.194
3NN	2	20	0.122	0.151	0.130	0.126	0.124	0.123	0.122	0.121	0.121	0.121	**0.120**
3NN	2	40	0.123	0.147	0.129	0.127	0.125	0.124	0.123	0.123	**0.122**	**0.122**	**0.121**
3NN	2	60	0.128	0.148	0.132	0.130	0.128	0.127	**0.126**	**0.126**	**0.125**	**0.126**	**0.125**
3NN	5	20	0.126	0.160	0.136	0.132	0.129	0.128	0.127	0.127	0.126	0.126	0.126
3NN	5	40	0.123	0.147	0.130	0.127	0.125	0.125	0.123	0.123	0.122	0.122	0.122
3NN	5	60	0.125	0.147	0.130	0.128	0.126	0.125	0.124	**0.124**	**0.123**	**0.123**	**0.123**
CART	2	20	0.160	0.182	0.161	**0.155**	**0.152**	**0.151**	**0.150**	**0.149**	**0.148**	**0.148**	**0.147**
CART	2	40	0.156	0.177	0.155	**0.150**	**0.146**	**0.145**	**0.144**	**0.143**	**0.142**	**0.142**	**0.142**
CART	2	60	0.158	0.177	**0.154**	**0.149**	**0.146**	**0.144**	**0.143**	**0.142**	**0.141**	**0.141**	**0.140**
CART	5	20	0.161	0.181	0.159	**0.154**	**0.151**	**0.149**	**0.148**	**0.148**	**0.147**	**0.146**	**0.146**
CART	5	40	0.158	0.181	0.156	**0.151**	**0.148**	**0.146**	**0.144**	**0.143**	**0.143**	**0.142**	**0.141**
CART	5	60	0.159	0.178	**0.154**	**0.148**	**0.143**	**0.143**	**0.140**	**0.140**	**0.139**	**0.138**	**0.138**
NNET	2	20	0.216	0.244	0.235	0.232	0.231	0.229	0.228	0.228	0.227	0.227	0.226
NNET	2	40	0.195	0.232	0.215	0.212	0.208	0.207	0.205	0.204	0.203	0.202	0.202
NNET	2	60	0.187	0.222	0.200	0.194	0.189	0.188	0.185	0.184	**0.182**	**0.182**	0.183
NNET	5	20	0.244	0.255	0.252	0.251	0.251	0.250	0.251	0.249	0.250	0.250	0.250
NNET	5	40	0.238	0.254	0.251	0.250	0.250	0.250	0.249	0.249	0.249	0.249	0.249
NNET	5	60	0.228	0.254	0.250	0.248	0.248	0.248	0.247	0.246	0.247	0.247	0.246

Table III. Expected classification error of selected experiments with prostate cancer protein abundance data (full results available on the companion website). Bold-face type indicates the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test.

Rule	p	n	Single	$m=5$	$m=11$	$m=15$	$m=21$	$m=25$	$m=31$	$m=35$	$m=41$	$m=45$	$m=51$
LDA	2	20	0.212	0.241	0.225	0.222	0.219	0.218	0.216	0.216	0.215	0.215	0.215
LDA	2	40	0.198	0.224	0.210	0.208	0.205	0.204	0.203	0.202	0.202	0.202	0.201
LDA	2	60	0.194	0.216	0.204	0.202	0.199	0.199	0.198	0.197	0.197	0.196	0.196
LDA	5	20	0.214	0.254	0.229	0.223	0.219	0.217	0.216	0.215	0.213	0.212	**0.212**
LDA	5	40	0.183	0.212	0.193	0.189	0.187	0.185	0.184	0.183	0.183	**0.182**	**0.181**
LDA	5	60	0.166	0.192	0.175	0.171	0.169	0.168	0.167	0.167	0.166	0.166	**0.165**
3NN	2	20	0.187	0.251	0.203	0.195	0.192	0.189	0.187	0.187	0.186	**0.185**	**0.185**
3NN	2	40	0.153	0.208	0.168	0.162	0.158	0.156	0.154	0.153	**0.152**	**0.152**	**0.151**
3NN	2	60	0.148	0.199	0.160	0.154	0.150	0.149	0.148	0.147	**0.146**	**0.146**	**0.145**
3NN	5	20	0.184	0.249	0.205	0.197	0.193	0.191	0.189	0.189	0.187	0.187	0.186
3NN	5	40	0.143	0.187	0.157	0.152	0.149	0.147	0.146	0.145	0.144	0.143	0.143
3NN	5	60	0.128	0.164	0.139	0.135	0.131	0.130	0.129	0.128	0.128	0.127	0.127
CART	2	20	0.232	0.247	**0.223**	**0.218**	**0.213**	**0.210**	**0.209**	**0.209**	**0.208**	**0.209**	**0.208**
CART	2	40	0.213	0.219	**0.198**	**0.194**	**0.189**	**0.189**	**0.187**	**0.185**	**0.185**	**0.185**	**0.184**
CART	2	60	0.204	0.205	**0.185**	**0.180**	**0.176**	**0.175**	**0.172**	**0.172**	**0.172**	**0.171**	**0.171**
CART	5	20	0.220	0.244	**0.216**	**0.210**	**0.206**	**0.204**	**0.201**	**0.200**	**0.199**	**0.198**	**0.199**
CART	5	40	0.187	0.215	0.188	**0.182**	**0.179**	**0.176**	**0.174**	**0.173**	**0.172**	**0.172**	**0.171**
CART	5	60	0.169	0.192	**0.166**	**0.160**	**0.156**	**0.154**	**0.152**	**0.151**	**0.150**	**0.150**	**0.149**
NNET	2	20	0.297	0.300	**0.271**	**0.266**	**0.260**	**0.259**	**0.256**	**0.256**	**0.254**	**0.254**	**0.253**
NNET	2	40	0.277	0.274	**0.254**	**0.248**	**0.244**	**0.244**	**0.240**	**0.241**	**0.239**	**0.239**	**0.239**
NNET	2	60	0.276	**0.268**	**0.246**	**0.243**	**0.239**	**0.238**	**0.236**	**0.235**	**0.234**	**0.234**	**0.234**
NNET	5	20	0.305	0.307	**0.270**	**0.261**	**0.255**	**0.250**	**0.248**	**0.247**	**0.246**	**0.243**	**0.244**
NNET	5	40	0.288	**0.274**	**0.249**	**0.242**	**0.238**	**0.235**	**0.233**	**0.233**	**0.231**	**0.230**	**0.230**
NNET	5	60	0.281	**0.267**	**0.244**	**0.238**	**0.234**	**0.232**	**0.229**	**0.228**	**0.227**	**0.227**	**0.226**

the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test. One observes that as ensemble size increases, classification error decreases and tends to converge to a fixed value (in agreement with [151]). But we can also see that the error is usually larger at very small ensemble sizes, as compared to the error of the individual classifier. We can again observe that, in most cases, bagging is able to improve the performance of CART and NNET, but that is not significantly so, or at all, for DLDA, LDA and 3NN. More importantly, we can see that the improvement on the performance of CART and NNET is not sufficient to beat the performance of single DLDA, LDA, or 3NN classifiers (with the exception of the prostate cancer data with RELIEF feature selection, which we comment on below).

As we can see in Figures 3–5, the breast cancer gene-expression data produces linear features that favor single DLDA and LDA classifiers (the latter do not perform so well at $n = 20$, due to the difficulty of estimating the entire covariance matrix at this sample size, which affects DLDA less), while the lung cancer gene-expression data produces nonlinear features, in which case, according to the results, the best option overall is to use a single 3NN classifier, followed closely by a bagged NNET in t-test feature selection and a bagged CART in RELIEF feature selection. The case of the prostate cancer proteomic data is peculiar in that it presents the only case where the best option was not a DLDA, LDA, or 3NN classifier, but in fact a single CART classifier, namely, the case $n = 20$ (with either $p = 2$ or $p = 5$) for RELIEF feature selection (the results for t-test feature selection, on the other hand, are very similar to the ones obtained for the lung cancer data set). Note that, in this case, the best performance is achieved by a single CART classifier, rather than the ensemble CART scheme. We also point out that the classification errors obtained with t-test feature selection are smaller than the ones obtained with RELIEF feature selection, indicating that RELIEF is not a good option in this case due to the very small sample size (in fact, there is evidence that t-test filter-based feature selection may be the method of choice in small sample cases [169]) In the case $n = 40$, the difference between 3NN and CART essentially disappears. It is also interesting that in the case $n = 20$ and $p = 5$, for RELIEF feature selection, bagging is able to improve the performance of LDA by a good margin in the case of the prostate cancer data. This is due to the fact that the combination LDA and RELIEF feature selection produces a unstable, overfitting classification rule at this acute small-sample scenario.

The results obtained with t-test feature selection are consistent across all data sets. When using RELIEF feature selection, there is a degree of contrast between the results for the prostate cancer protein-abundance data set and the ones for the gene-expression data sets, which may be attributed to the differences in technology as well as the fact that we do not employ baseline subtraction for the proteomics data in order to achieve better classification rates.

We remark that results are not expected to change much if ensemble sizes are increased further (beyond $m = 51$), as can be seen from convergence of the expected classification error curves in Figures 3–5.

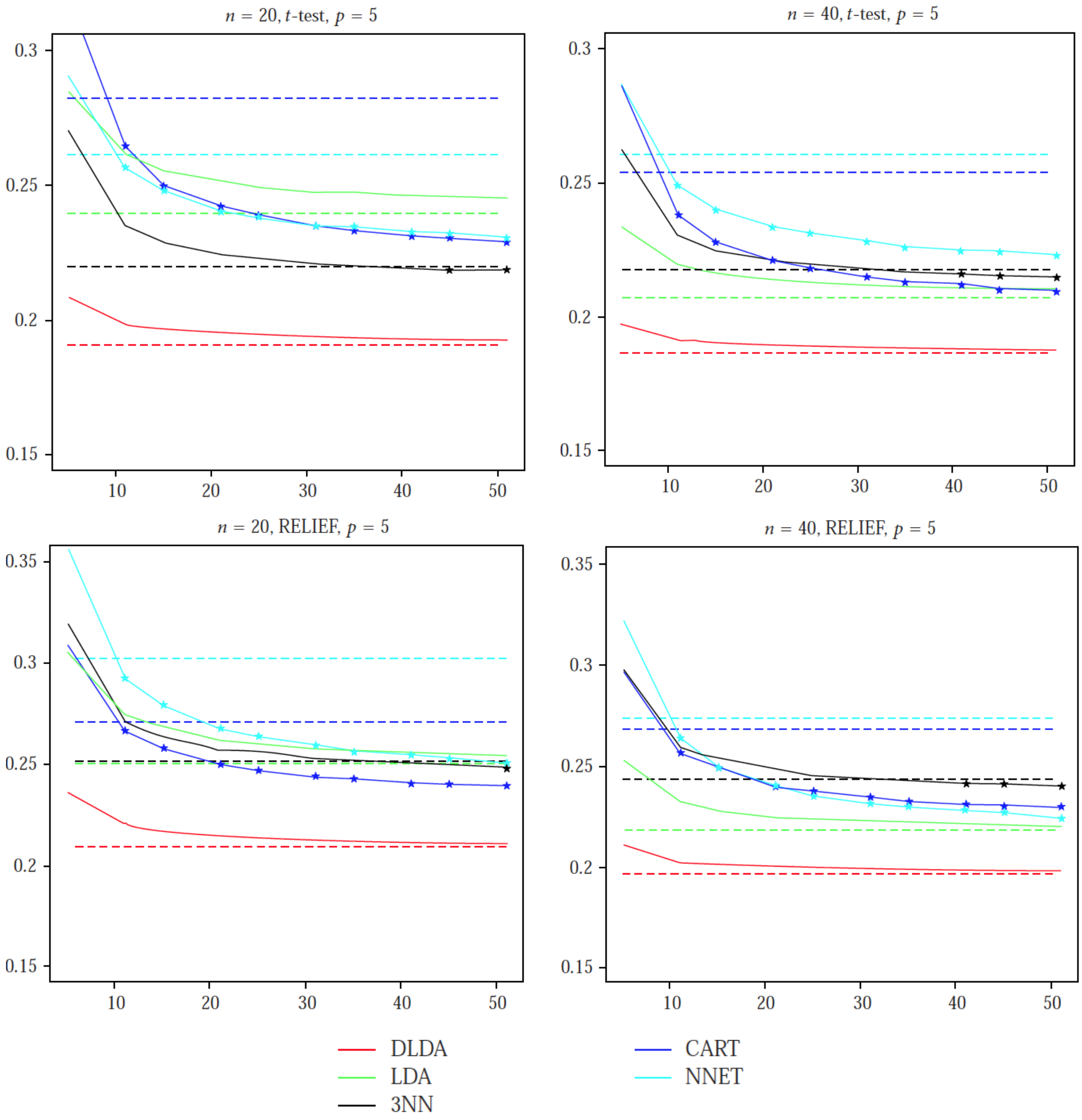


Fig. 3. Expected classification error as a function of number of component classifiers in the ensemble for selected experiments with the breast cancer gene expression data (full results available on the companion website). Error of single component classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test.

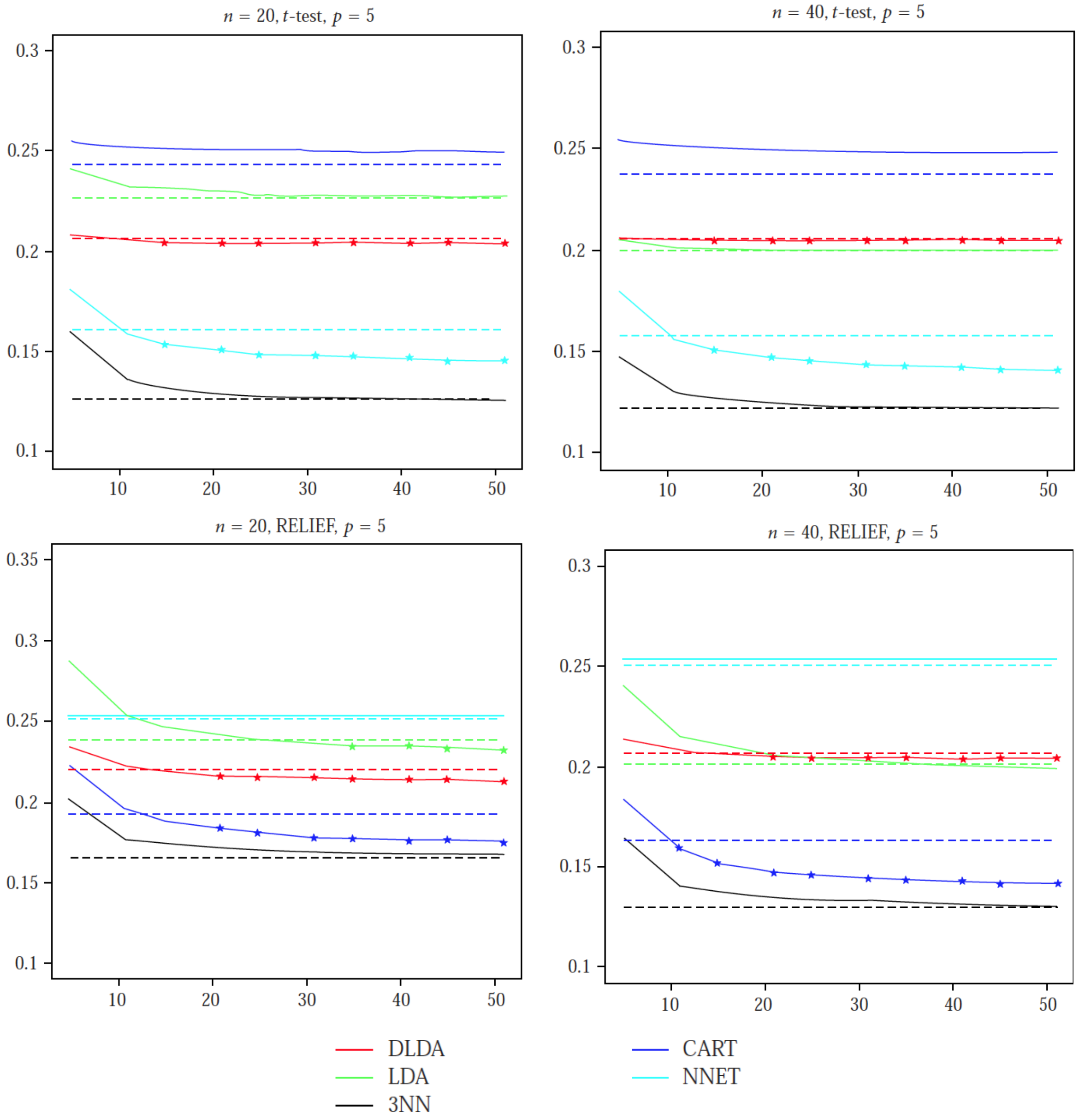


Fig. 4. Expected classification error as a function of number of component classifiers in the ensemble for selected experiments with the lung cancer gene expression data (full results available on the companion website). Error of single component classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test.

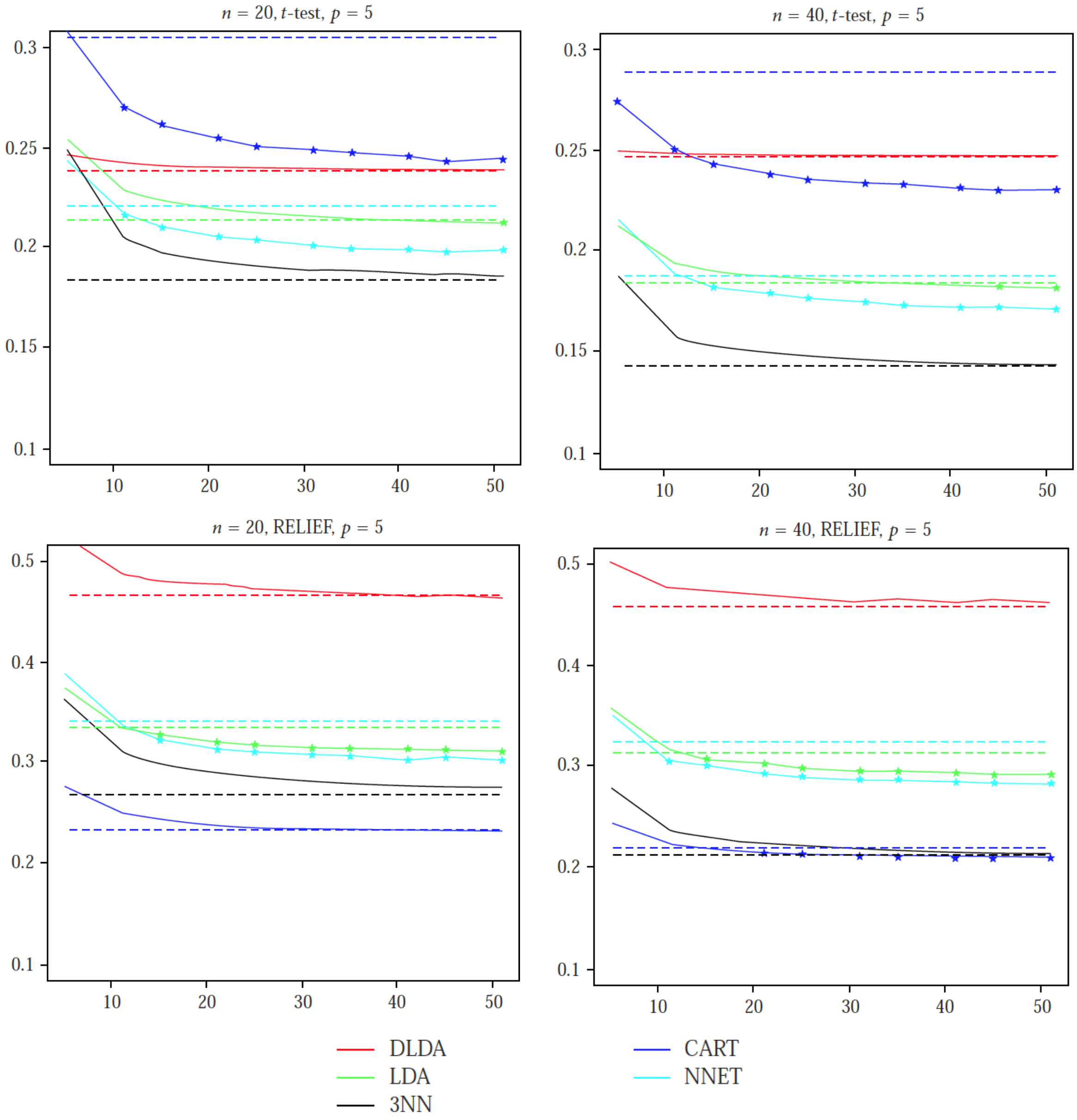


Fig. 5. Expected classification error as a function of number of component classifiers in the ensemble for selected experiments with the prostate cancer protein abundance data (full results available on the companion website). Error of single component classifier is indicated by a horizontal dashed line. Marks indicate the values that are smaller for the ensemble classifier as compared to a single component classifier at a 99% significance level, according to a one-tailed paired t-test.

D. Conclusion

In this chapter we conducted a detailed empirical study of the ensemble approach to classification of small-sample genomic and proteomic data. The main performance issue is not whether the ensemble scheme improves the classification error of an unstable, overfitting classifier (e.g., CART, NNET), or whether its classification error converges to a fixed limit; but rather whether the ensemble scheme will improve performance of the unstable, overfitting classifier *sufficiently* to beat the performance of single stable, non-overfitting classifiers (e.g., DLDA, LDA, 3NN). We observed that this never was the case for any of the data sets and experimental conditions considered here, except in the case of the proteomics data set with RELIEF feature selection in acute small-sample cases, when nevertheless the performance of a single unstable, overfitting classifier (in this case, CART) was better or comparable to the corresponding ensemble classifier. We observed that in most cases bagging does a good (sometimes, admirable) job of improving the performance of unstable, overfitting classifiers, but that improvement was not enough to beat the performance of single stable, non-overfitting classifiers.

The main message to be gleaned from this study by practitioners is that the use of bagging in classification of small-sample genomics and proteomics data increases computational cost, but is not likely to improve overall classification accuracy over other, more simple, approaches. The solution we recommend is to use simple classification rules and avoid bagging in these scenarios. It is important to stress that we do not give a definitive recommendation on the use of the random forest method for small-sample genomics and proteomics data; however, we do think that this study does provide a step in that direction, since the random forest method depends partly, if not significantly, for its success on the effectiveness of bagging. Further research is needed to investigate this question.

CHAPTER VI

SMALL-SAMPLE ERROR ESTIMATION FOR BAGGING CLASSIFICATION RULES [*]

Application of ensemble classification rules in gene-expression microarray classification problems has become increasingly common. Among ensemble classification rules, bootstrap aggregating ("bagging") is the most popular, and has generated a considerable amount of literature. However, the problem of error estimation for these classification rules, particularly under the small-sample settings prevalent in genomics, is not well understood. Breiman proposed a general method, which he called "out-of-bag", for estimating statistics of bagged classifiers, which was subsequently applied by other authors to estimate the classification error. In this chapter, we give an explicit definition of the out-of-bag estimator that is intended to remove estimator bias, by formulating carefully how the error count is normalized. We conducted an extensive simulation study of bagging of common classification rules, including LDA, 3NN, and CART, applied on both synthetic and real patient data, corresponding to the use of common error estimators such as resubstitution, leave-one-out, cross-validation, basic bootstrap, bootstrap 632, bootstrap 632 plus, bolstering, semi-bolstering, in addition to the out-of-bag estimator. The results from the numerical experiments indicated that the performance of the out-of-bag estimator is very similar to that of leave-one-out; in particular, the out-of-bag estimator is slightly pessimistically biased. The performance of the other estimators are consistent with their performance with the corresponding single classifiers, as reported in other studies. Bolstered error estima-

[*] Reprinted with permission from "Small-sample Error Estimation for Bagged Classification Rules," by T. T. Vu and U. M. Braga-Neto, 2009. volume 2010, 12 pages, Copyright 2010 of *EURASIP Journal on Bioinformatics and Systems Biology.*

tors showed consistent superior performance to the others, in terms of accuracy (RMS) and computational cost.

A. Introduction

Ensemble classification methods combine the decision of multiple classifiers designed on randomly perturbed versions of the available data [147, 148, 149, 150, 151]. The most popular version of this scheme is known as bootstrap aggregating, or "bagging" [150, 151] where the ensemble classifier corresponds to a majority-vote among classifiers designed on bootstrap samples [96] from the available training data.

There has been considerable interest recently in the application of bagging in the classification of both gene-expression data [154, 155, 156, 157] and protein-abundance mass spectrometry data [158, 159, 160, 161, 162, 163]. The popularity of bagging is based on the expectation that combining the decision of several classifiers will regularize and improve the performance of unstable, overfitting classification rules (the so-called "weak learners"). In Chapter V, we have investigated this claim, in the context of small-sample genomics and proteomics data. On the other hand, a different issue is the performance of error estimators for bagged classifiers. Accurate error estimation is a critical issue in Genomics, as it decisively impacts the scientific validity of hypotheses derived from application of pattern recognition methods to biomedical data [43, 179, 180]. On the topic of error estimation, Breiman proposed a general method, which he called "out-of-bag", for estimating statistics of bagged classifiers [181], and, subsequently, other authors applied it to the estimation of the classification error [182, 183]. In this chapter, we give an explicit definition of the out-of-bag estimator that is intended to remove estimator bias, which is done by formulating carefully how the error count is normalized. The performance of out-of-bag estimators with general bagged classification rules is not in fact well understood, especially in connection with bagging ensemble classifiers derived from classification rules other than decision trees (which was Breiman's primary interest). In addition, to our knowledge, no studies have attempted to assess the performance of error estimators for bagged classifiers in the context of Genomics data, particularly in the prevalent small-sample setting usually found in these applications.

To investigate these issues, we conducted an extensive simulation study of bagging of common classification rules, including LDA, 3NN, and CART, applied on both synthetic and real patient data, corresponding to the use of common error estimators such as resubstitution, leave-one-out, cross-validation, basic bootstrap, bootstrap 632, bootstrap 632 plus, bolstering, semi-bolstering, in addition to the out-of-bag estimator itself. We present here selected representative results; the full set of results can be found on the companion website, at http://gsp.tamu.edu/Publications/supplementary/oob. The results from the numerical experiments indicated that the performance of the out-of-bag error estimator is very similar to that of leave-one-out; in particular, the out-of-bag estimator is slightly pessimistically biased. The performance of the other estimators are for the most part consistent with their performance with the corresponding single classification rules assessed in other studies, with the best performance being provided by the bolstered error estimators, in terms of root mean square error.

B. Error Estimation for Bagging Classification Rule

1. Classical Methods

Classical error estimation methods including resubstitution, cross-validation, and bootstrap are reviewed in Chapter II. Readers are encouraged to refer back to chapter II for more details. All these estimation methods are to be applied to the bagging classification rules.

2. Bolstered Error Estimation

Bolstered estimation was proposed in [120]. It has shown promising performance for small sample sizes in terms of root mean square error. While it is comparable to bootstrap methods in many cases, bolstered estimators are typically much more computationally efficient than the bootstrap. The main idea of bolstering is to put a kernel at each of the sample point, called "bolstering kernel" to smooth the variance of counting-based estimation methods (in this chapter, we adopt Gaussian bolstering kernels). When the classifiers are overfitted, and hence, resubstitution estimates are optimistically biased, then bolstering at a misclassified point will increase this bias. Semi-bolstering is suggested for correcting this, by conducting no bolstering at misclassified points. We refer the reader to [120] for the full details (in this chapter, we employ the bolstered and semi-bolstered resubstitution estimators of [120]).

3. Out-of-bag Error Estimation

Breiman [181] originally proposed the out-of-bag method to estimate the generalization error of bagged predictors of CART and the node priority probabilities. Bylander [182] later did a simulation study comparing out-of-bag and cross-validation for tree classification C4.5 and concluded that both are biased. Banfield et al [183] used out-of-bag in a large simulation of investigating performances of a variety of ensemble methods. Martinez [184], in an attempt to find the optimal number of components of ensembles, employed out-of-bag as the optimization criterion. Despite that, the properties of the out-of-bag estimator remain largely unclear, in particular, the issue of bias. We propose in the sequel a modification to the standard out-of-bag estimator that removes nearly all of its bias (as evidenced by the numerical experiments in Section C).

In bagging, component classifiers are designed based on bootstrap sets, each of which contain on average 63% of the original sample set. Hence, there are approximately 37%

of the data which are not used to build the classifier and are therefore uncorrelated with it. Out-of-bag estimates are obtained by testing the majority-voting classifier via those individual classifiers in the ensemble that are uncorrelated with the testing point, i.e., those classifiers whose training sets do not contain the testing points. Suppose we resample the original sample set k times, leading to k bootstrap sample sets S^{*j}. Let $P_i^j = 1$ if sample i appears in the bootstrap sample S^{*j}, and $P_i^j = 0$, otherwise, for $i = 1,\ldots,n$. Denote

$$\begin{aligned}
A_0(i) &= \sum_{j=1}^{k} I_{\{P_i^j=0\}} I_{\{Y_i=0\}} \\
B_0(i) &= \sum_{j=1}^{k} I_{\{P_i^j=0\}} I_{\{\Psi_n(S^{*j})(X_i)=1\}} I_{\{Y_i=0\}} \\
A_1(i) &= \sum_{j=1}^{k} I_{\{P_i^j=0\}} I_{\{Y_i=1\}} \\
B_1(i) &= \sum_{j=1}^{k} I_{\{P_i^j=0\}} I_{\{\Psi_n(S^{*j})(X_i)=0\}} I_{\{Y_i=1\}}
\end{aligned} \tag{6.1}$$

for $i = 1,\ldots,n$. Notice that $A_m(i)$ is equal to the number of times that sample i in class m appears across all bootstrap sample sets, while $B_m(i)$ is equal to the number of times that sample i in class m appears and is *misclassified* across all bootstrap sample sets. Then the out-of-bag error estimator, as proposed by Breiman in [150], can be written as

$$\hat{\varepsilon}_{\text{oob}} = \frac{1}{n} \sum_{i=1}^{n} \left[I_{\{B_0(i) \geq \frac{A_0(i)}{2}\}} I_{\{A_0(i)>0\}} + I_{\{B_1(i) \geq \frac{A_1(i)}{2}\}} I_{\{A_1(i)>0\}} \right]. \tag{6.2}$$

The estimator, as formulated above, will be optimistically biased, in general, according to the following rationale. Clearly, when $Y_i = j$ and $A_j(i) = 0$, then the i-th sample point belongs to all of the bootstrap samples, so there are no individual classifiers to test on the i-th point. In other words, the "out-of-bag ensemble" of classifiers for that point is empty in this case. That means that, with training sample size of n, we often have fewer than n samples to perform the out-of-bag estimation. In computing the proportion of incorrect classification by the ensemble, one should therefore divide not by n as in (6.2), but rather by n minus the number of times when the out-of-bag ensembles are empty, which leads to the following modified out-of-bag estimator:

$$\hat{\varepsilon}_{\mathrm{oob}}^{\mathrm{m}} = \frac{1}{n - \sum_{i=1}^{n} \left[I_{\{A_0(i)=0\}} + I_{\{A_1(i)=0\}} \right]} \sum_{i=1}^{n} \left[I_{\{B_0(i) \geq \frac{A_0(i)}{2}\}} I_{\{A_0(i)>0\}} + I_{\{B_1(i) \geq \frac{A_1(i)}{2}\}} I_{\{A_1(i)>0\}} \right]. \tag{6.3}$$

As shown by the numerical results in Section C, this estimator has approximately the bias of leave-one-out, i.e., it is only slightly pessimistically biased. As far as we know, this formulation of the out-of-bag estimator has not been explicitly given in the literature.

C. Simulation Study

This section reports the results of an extensive simulation study, which were conducted on both synthetic and publicly available microarray data and protein abundance mass spectrometry data. We present here selected representative results; the full set of results can be found on the companion website, at http://gsp.tamu.edu/Publications/supplementary/oob. We simulated bagged ensembles of linear discriminant analysis (LDA), 3-nearest-neighbors (3NN), and decision trees (CART) [60], and computed actual and estimated errors, according to the different estimation methods. These estimators were evaluated based on the distribution of their deviation from the true error, and in terms of bias, variance, and root-mean-square (RMS) errors.

1. Methods

We compared the performances of estimators for varying number of training samples with different dimensions of the feature space. The dimensionality and number of samples are selected to be compatible with a small-sample scenario (in this chapter, the dimensionality is kept fixed at $p = 2$). For patient data, a small number of features (once again, $p = 2$ in this chapter) are first selected by the t-test. We afterwards randomly draw a number of samples to be used as the training set and employed the rest as a testing set. The number of training points are chosen to be small to keep the small sample setting, and to have a large enough testing set. This was repeated 1000 times to get the empirical deviation distribution [43], that is, the distribution of estimated minus actual errors, for the different error estimators. The results are presented in forms of beta-fit curves, box-plots, and bias, variance, and RMS curves in order to provide as detailed as possible a picture of the empirical deviation distributions of the error estimators.

2. Simulation Based on Synthetic Data

We employ here the spherical gaussian model, where the covariance matrix is identity and the two mean vector are symmetric over the origin. With that assumption, we varied the Bayes error of the model by changing the distance between the two means. Models with different Bayes errors and dimension are compared over varying number of samples. The feature-label distribution is known and this allows us to exactly compute the true error of the designed classifier, which is then used to derive the empirical deviation distribution for the different estimators.

3. Simulation Based on Patient Data

We utilize the same three patients data sets described in Chapter V in order to study the performance of bagging in the context of genomics and proteomics applications.

4. Results and Discussion

a. Synthetic Data

The various error estimators can be grouped into four groups according to performance. The first group corresponds to resubstitution, which showed to be optimistically biased for the bagged LDA, 3NN, and CART classifiers, with a root mean square error that increases substantially with increasing Bayes error; resubstitution had been previously known to behave as such for single LDA, 3NN, and CART classifiers. The second group contains leave-one-out, five-fold cross-validation and out-of-bag. As we can see from Figure 6, the out-of-bag estimator, with the formulation given in (6.3), is almost identical to leave-one-out. This second group shows very small bias but considerably high variance. The resemblance of out-of-bag to cross-validation, which had been pointed out already in [182], is explained by the similar way of partitioning the sample set. This group shows much smaller bias than resubstitution, and this is consistent as the Bayes error increases. However, this group displayed larger variability than resubstitution and the bootstrap group, as we already knew from [179] on single classification rules. The third group includes the basic bootstrap, bootstrap 632 and bootstrap 632 plus; this group displays very competitive performance in terms of root mean square error. Even though they often perform better than the two previous groups, the estimators in this group took the longest time to compute across all experiments. The last group consists of the bolstered and semi-bolstered error estimators, which exhibit superior performance to the other groups, in terms of RMS error, despite being far less computationally expensive than cross-validation and bootstrap estimators.

Generally, for a fixed model, almost all the estimates work better when the sample size increase and this holds for all three bagged classifiers. In Figure 7, we see that there is a consistent trend: as the Bayes error increases or, equivalently, the classification problem becomes harder, error estimation performance decreases steadily, in term of root mean square error; this is true for all error estimation methods. Bolstered error estimators showed consistent superior performance to the others, in terms of accuracy (RMS) and computational cost. These conclusions are also supported by Figures 8 and 9.

We observed that the performance of error estimators other than out-of-bag (which

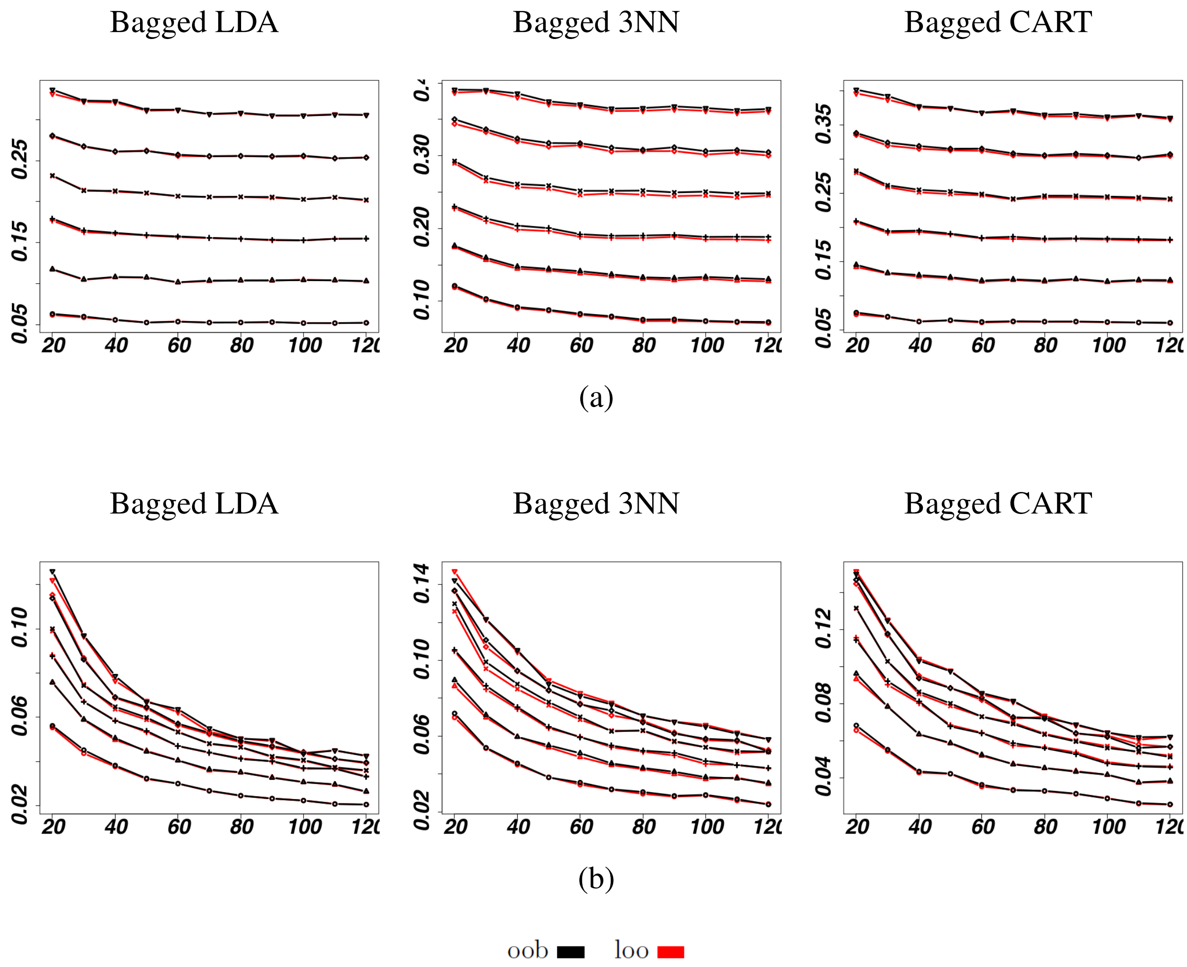
Table IV. Bias, variance (standard deviation), and RMS for different error estimators, with different base classification rules, for breast cancer gene expression data, dimensionality $p = 2$.

Rule	n	stat	resb	boot	bresb	loo	b632	oob	sbresb	b632plus	cv5
lda	20	bias	-0.0388	0.0287	-0.0104	0.0063	0.0039	0.0076	0.0244	0.0092	0.0143
		sd	0.0908	0.0944	0.0789	0.1004	0.0912	0.1003	0.0933	0.0938	0.1140
		rms	0.0988	0.0986	0.0795	0.1006	0.0913	0.1006	0.0964	0.0942	0.1149
lda	40	bias	-0.0198	0.0082	-0.0084	-0.0012	-0.0021	0.0002	0.0168	-0.0011	-0.0044
		sd	0.0657	0.0642	0.0614	0.0671	0.0638	0.0673	0.0676	0.0641	0.0714
		rms	0.0686	0.0647	0.0620	0.0671	0.0639	0.0673	0.0696	0.0641	0.0716
lda	60	bias	-0.0157	-0.0000	-0.0097	-0.0045	-0.0058	-0.0036	0.0104	-0.0054	-0.0011
		sd	0.0577	0.0559	0.0544	0.0580	0.0560	0.0581	0.0586	0.0560	0.0586
		rms	0.0598	0.0559	0.0553	0.0582	0.0563	0.0582	0.0595	0.0563	0.0587
cart	20	bias	-0.1554	0.0456	-0.0330	0.0226	-0.0284	0.0267	-0.0225	0.0096	0.0094
		sd	0.0653	0.1047	0.0671	0.1210	0.0798	0.1229	0.0700	0.1059	0.1187
		rms	0.1686	0.1142	0.0747	0.1231	0.0847	0.1258	0.0735	0.1063	0.1190
cart	40	bias	-0.1583	0.0323	-0.0358	0.0095	-0.0378	0.0143	-0.0284	-0.0094	0.0058
		sd	0.0484	0.0697	0.0502	0.0774	0.0533	0.0799	0.0516	0.0671	0.0810
		rms	0.1655	0.0769	0.0616	0.0780	0.0653	0.0812	0.0589	0.0677	0.0812
cart	60	bias	-0.1722	0.0211	-0.0377	0.0001	-0.0501	0.0043	-0.0317	-0.0232	-0.0050
		sd	0.0400	0.0624	0.0473	0.0705	0.0473	0.0701	0.0472	0.0590	0.0695
		rms	0.1768	0.0658	0.0605	0.0705	0.0689	0.0703	0.0569	0.0634	0.0697
3nn	20	bias	-0.0964	0.0575	-0.0478	0.0270	0.0009	0.0269	-0.0176	0.0273	0.0076
		sd	0.0716	0.0996	0.0649	0.1174	0.0835	0.1167	0.0778	0.1005	0.1156
		rms	0.1201	0.1150	0.0806	0.1204	0.0835	0.1197	0.0798	0.1041	0.1159
3nn	40	bias	-0.0952	0.0406	-0.0481	0.0109	-0.0094	0.0139	-0.0214	0.0075	0.0036
		sd	0.0529	0.0687	0.0493	0.0787	0.0590	0.0785	0.0577	0.0669	0.0801
		rms	0.1089	0.0798	0.0689	0.0794	0.0598	0.0797	0.0615	0.0673	0.0802
3nn	60	bias	-0.0962	0.0316	-0.0504	0.0034	-0.0154	0.0054	-0.0261	-0.0012	-0.0008
		sd	0.0432	0.0625	0.0452	0.0693	0.0526	0.0693	0.0514	0.0595	0.0680
		rms	0.1054	0.0701	0.0677	0.0694	0.0548	0.0695	0.0576	0.0595	0.0680

can only be applied to ensemble rules) were consistent with their performance with the corresponding single classifier, as reported in other studies [43, 120].

b. Patient Data

The results for the real patient data sets were entirely consistent with those for the synthetic data, as can be seen in Figures 10–12 and Tables 4–6. We again observed the division of the error estimators in the same four groups according to performance. We also observed that the bolstered error estimator group displayed the best performance, as measured by RMS.



(a)

(b)

Fig. 6. Comparison of out-of-bag and leave-one-out for different Gaussian models over the number of samples $p = 2$ (a) Sample mean, (b) Sample standard deviation.

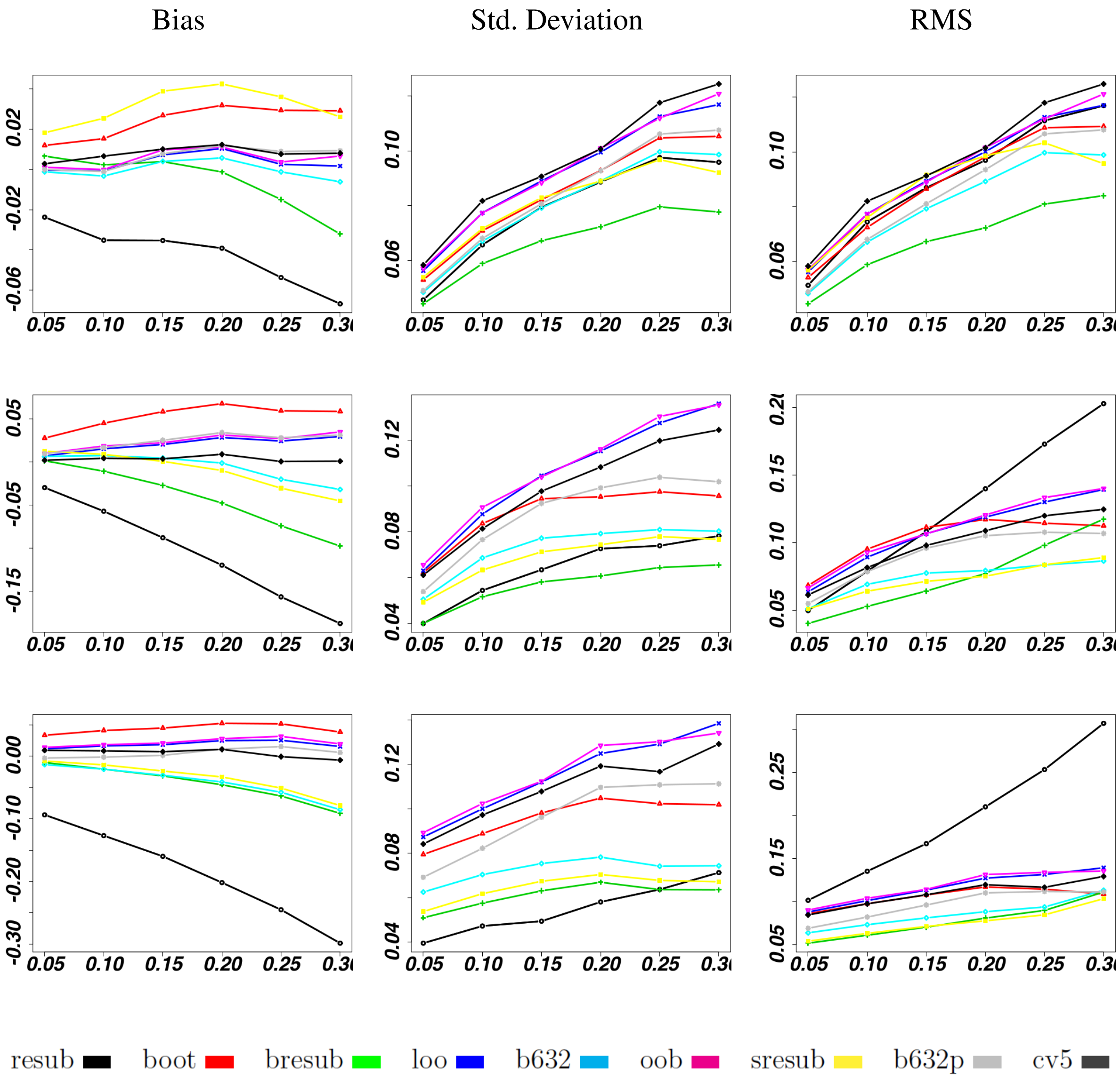


Fig. 7. Bias, variance (standard deviation), and RMS of as a function of the bayes error, for the synthetic data, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules: LDA, 3NN, and CART on the first, second, and third row, respectively.

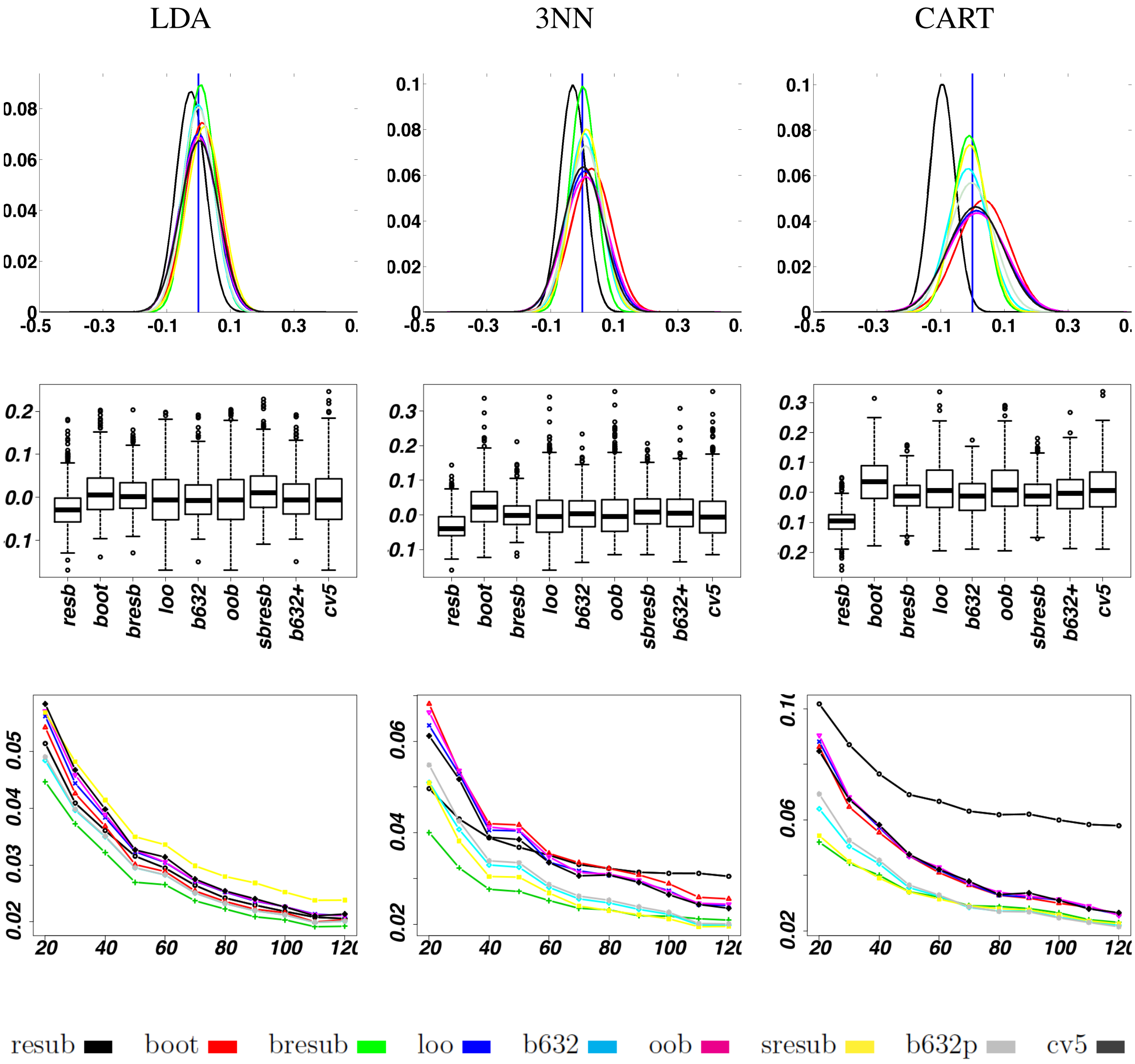


Fig. 8. Empirical deviation distribution (top row), box plots (middle row), and RMS as a function of sample size (bottom row), for synthetic Gaussian model with Bayes error = 0.05, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules.

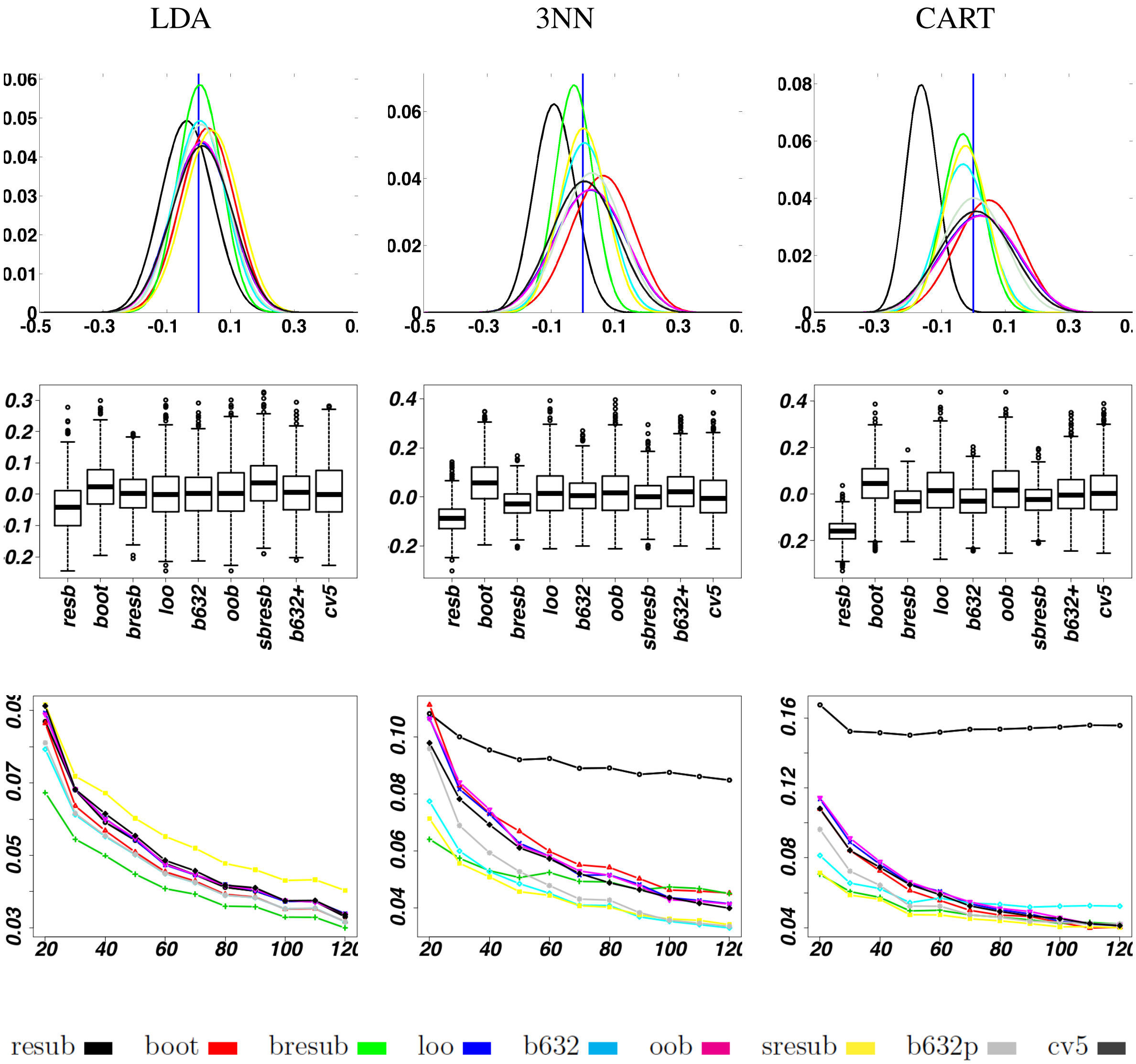


Fig. 9. Empirical deviation distribution (top row), box plots (middle row), and RMS as a function of sample size (bottom row), for synthetic Gaussian model with Bayes error = 0.15, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules.

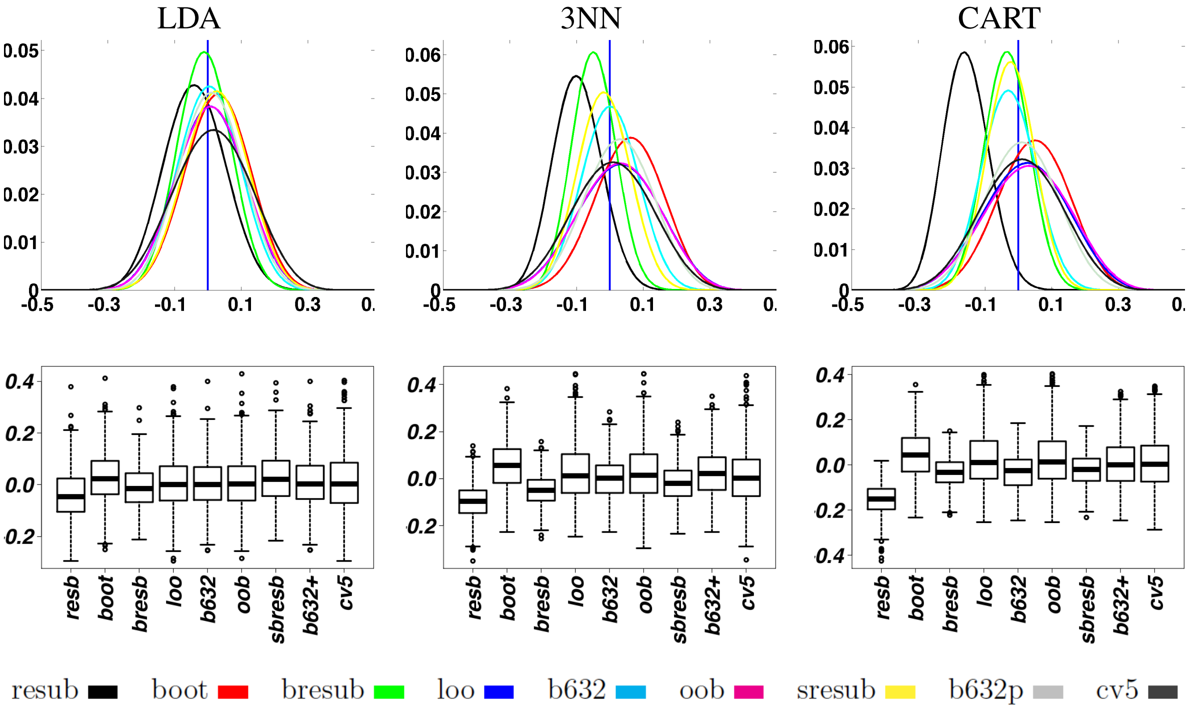


Fig. 10. Empirical deviation distribution (top row) and box plots (bottom row), for breast cancer gene-expression data, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules.

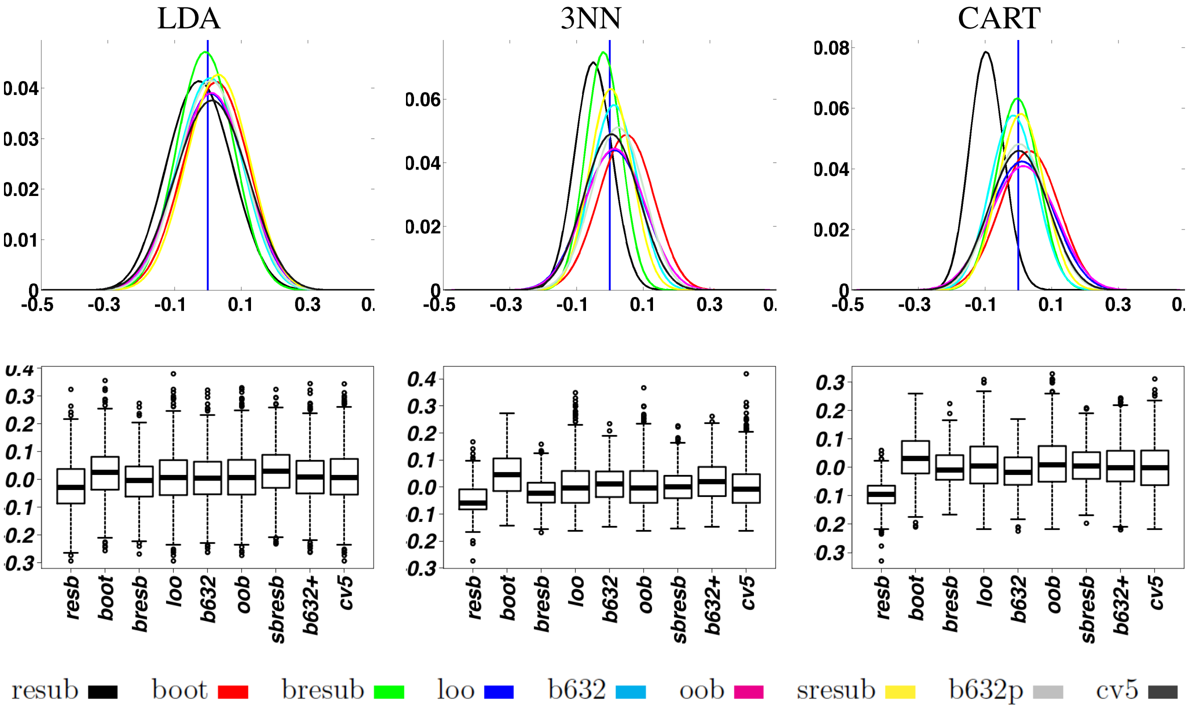


Fig. 11. Empirical deviation distribution (top row) and box plots (bottom row), for lung cancer gene-expression data, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules.

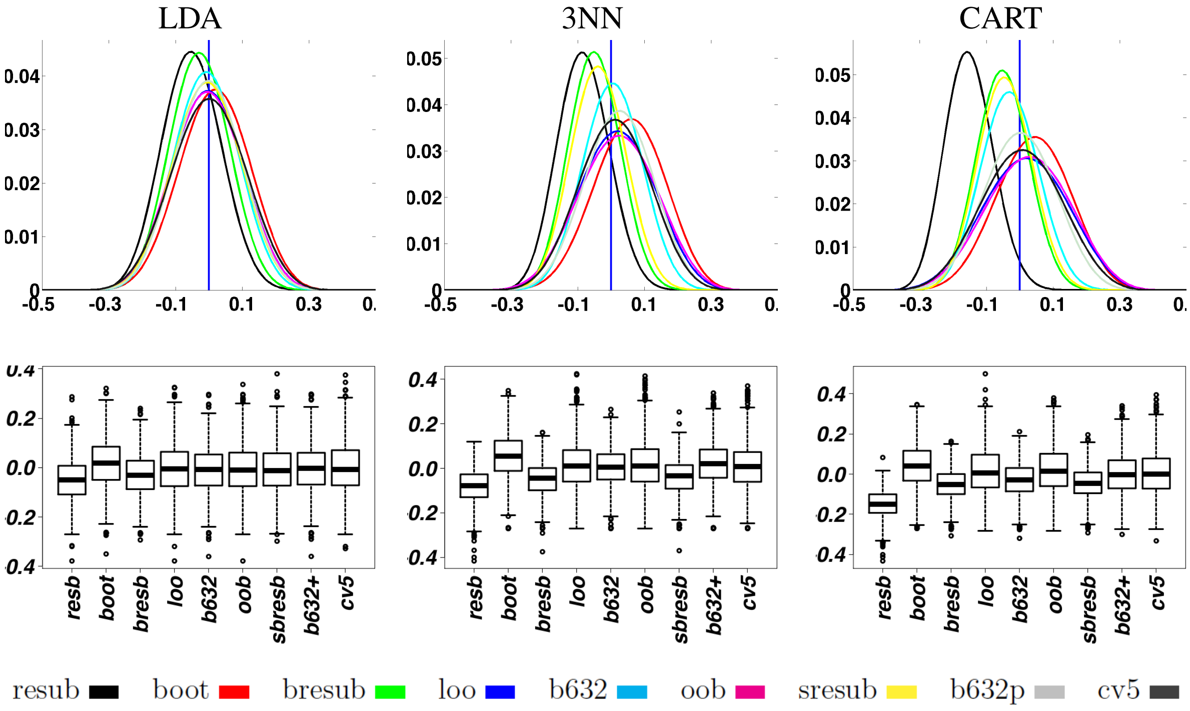


Fig. 12. Empirical deviation distribution (top row) and box plots (bottom row), for prostate cancer mass-spectrometry data, sample size $n = 20$, dimensionality $p = 2$, with different base classification rules.

Table V. Bias, variance (standard deviation), and RMS for different error estimators, with different base classification rules, for lung cancer gene expression data, dimensionality $p = 2$.

Rule	n	stat	resb	boot	bresb	loo	b632	oob	sbresb	b632plus	cv5
lda	20	bias	-0.0243	0.0238	-0.0070	0.0075	0.0061	0.0103	0.0294	0.0094	0.0106
		sd	0.0938	0.0938	0.0827	0.0989	0.0923	0.0988	0.0910	0.0932	0.1025
		rms	0.0969	0.0967	0.0830	0.0992	0.0925	0.0993	0.0956	0.0937	0.1031
lda	40	bias	-0.0118	0.0109	0.0012	0.0017	0.0025	0.0044	0.0273	0.0033	0.0045
		sd	0.0675	0.0655	0.0628	0.0684	0.0656	0.0685	0.0652	0.0656	0.0694
		rms	0.0685	0.0664	0.0628	0.0684	0.0657	0.0686	0.0707	0.0657	0.0695
lda	60	bias	-0.0092	0.0067	0.0023	-0.0004	0.0009	0.0015	0.0235	0.0012	0.0020
		sd	0.0606	0.0587	0.0570	0.0608	0.0590	0.0608	0.0586	0.0590	0.0610
		rms	0.0613	0.0591	0.0570	0.0608	0.0591	0.0609	0.0632	0.0590	0.0610
cart	20	bias	-0.0945	0.0321	-0.0025	0.0100	-0.0145	0.0139	0.0076	0.0031	0.0017
		sd	0.0502	0.0852	0.0623	0.0916	0.0683	0.0945	0.0676	0.0811	0.0849
		rms	0.1069	0.0911	0.0623	0.0921	0.0699	0.0955	0.0681	0.0812	0.0849
cart	40	bias	-0.0926	0.0226	-0.0230	0.0071	-0.0198	0.0088	-0.0141	-0.0071	0.0022
		sd	0.0384	0.0630	0.0439	0.0694	0.0504	0.0705	0.0472	0.0577	0.0654
		rms	0.1003	0.0670	0.0496	0.0698	0.0542	0.0710	0.0493	0.0581	0.0655
cart	60	bias	-0.0938	0.0202	-0.0277	0.0043	-0.0218	0.0068	-0.0210	-0.0103	0.0012
		sd	0.0335	0.0544	0.0397	0.0590	0.0438	0.0597	0.0414	0.0496	0.0571
		rms	0.0996	0.0580	0.0484	0.0592	0.0490	0.0601	0.0464	0.0507	0.0571
3nn	20	bias	-0.0483	0.0474	-0.0185	0.0114	0.0122	0.0132	0.0027	0.0238	0.0040
		sd	0.0552	0.0803	0.0529	0.0876	0.0677	0.0870	0.0623	0.0765	0.0787
		rms	0.0734	0.0932	0.0561	0.0884	0.0688	0.0880	0.0624	0.0802	0.0788
3nn	40	bias	-0.0489	0.0236	-0.0270	0.0043	-0.0031	0.0055	-0.0094	0.0027	-0.0004
		sd	0.0435	0.0602	0.0411	0.0626	0.0519	0.0624	0.0484	0.0555	0.0593
		rms	0.0655	0.0646	0.0492	0.0627	0.0520	0.0626	0.0493	0.0555	0.0593
3nn	60	bias	-0.0500	0.0198	-0.0317	0.0031	-0.0059	0.0036	-0.0147	-0.0009	-0.0028
		sd	0.0381	0.0526	0.0383	0.0555	0.0459	0.0553	0.0439	0.0486	0.0514
		rms	0.0629	0.0562	0.0497	0.0556	0.0462	0.0555	0.0463	0.0486	0.0514

Table VI. Bias, variance (standard deviation), and RMS for different error estimators, with different base classification rules, for prostate cancer mass-spectrometry data, dimensionality $p = 2$.

Rule	n	stat	resb	boot	bresb	loo	b632	oob	sbresb	b632plus	cv5
lda	20	bias	-0.0506	0.0181	-0.0277	-0.0033	-0.0072	-0.0044	-0.0050	-0.0019	0.0006
		sd	0.0871	0.1025	0.0879	0.1031	0.0949	0.1037	0.0993	0.0985	0.1071
		rms	0.1007	0.1041	0.0921	0.1031	0.0951	0.1038	0.0994	0.0985	0.1071
lda	40	bias	-0.0283	0.0079	-0.0189	-0.0051	-0.0054	-0.0042	-0.0029	-0.0039	-0.0031
		sd	0.0609	0.0688	0.0626	0.0673	0.0647	0.0683	0.0674	0.0655	0.0693
		rms	0.0672	0.0693	0.0654	0.0675	0.0649	0.0684	0.0675	0.0656	0.0694
lda	60	bias	-0.0192	0.0045	-0.0141	-0.0042	-0.0042	-0.0044	-0.0008	-0.0035	-0.0017
		sd	0.0514	0.0572	0.0524	0.0542	0.0542	0.0549	0.0559	0.0546	0.0577
		rms	0.0549	0.0573	0.0542	0.0544	0.0544	0.0550	0.0560	0.0547	0.0577
cart	20	bias	-0.1504	0.0409	-0.0500	0.0164	-0.0295	0.0248	-0.0441	0.0014	0.0059
		sd	0.0693	0.1082	0.0765	0.1198	0.0847	0.1223	0.0791	0.1053	0.1169
		rms	0.1655	0.1157	0.0914	0.1209	0.0897	0.1247	0.0905	0.1054	0.1170
cart	40	bias	-0.1412	0.0320	-0.0436	0.0047	-0.0317	0.0096	-0.0418	-0.0108	0.0044
		sd	0.0461	0.0701	0.0497	0.0753	0.0539	0.0773	0.0503	0.0646	0.0787
		rms	0.1485	0.0771	0.0661	0.0755	0.0625	0.0779	0.0654	0.0655	0.0788
cart	60	bias	-0.1397	0.0284	-0.0404	0.0021	-0.0334	0.0088	-0.0393	-0.0155	0.0049
		sd	0.0347	0.0580	0.0418	0.0626	0.0441	0.0648	0.0424	0.0521	0.0636
		rms	0.1439	0.0646	0.0581	0.0627	0.0554	0.0654	0.0578	0.0544	0.0637
3nn	20	bias	-0.0820	0.0554	-0.0488	0.0165	0.0048	0.0200	-0.0371	0.0233	0.0104
		sd	0.0748	0.1041	0.0757	0.1100	0.0871	0.1129	0.0805	0.0993	0.1037
		rms	0.1110	0.1179	0.0901	0.1112	0.0872	0.1147	0.0886	0.1020	0.1043
3nn	40	bias	-0.0673	0.0405	-0.0377	0.0029	0.0008	0.0067	-0.0271	0.0099	0.0040
		sd	0.0458	0.0643	0.0460	0.0679	0.0536	0.0695	0.0504	0.0585	0.0644
		rms	0.0814	0.0760	0.0595	0.0680	0.0536	0.0698	0.0572	0.0593	0.0645
3nn	60	bias	-0.0660	0.0304	-0.0375	0.0015	-0.0051	0.0040	-0.0269	0.0016	0.0006
		sd	0.0389	0.0534	0.0393	0.0560	0.0451	0.0563	0.0435	0.0482	0.0557
		rms	0.0766	0.0614	0.0543	0.0560	0.0454	0.0564	0.0511	0.0482	0.0557

D. Conclusion

We presented an extensive study of several error estimation methods for bagged ensembles of typical classifiers. We provided here an explicit formulation for the out-of-bag error estimator, which is intended to remove estimator bias. We observed that this out-of-bag error estimator was almost identical to leave-one-out, under spherical Gaussian models, and conjectured a very close relationship between the two. The results of our simulation study were consistent between synthetic and real patient data, and the performance of error estimators that can be applied to single classifiers (i.e., all of them save for the out-of-bag estimator) with the bagged classifiers was comparable to their performance with the corresponding single classifier, as reported elsewhere. The bolstered error estimators exhibited the best performance, in terms of RMS error, in our simulation study, despite being far less computationally expensive than cross-validation and bootstrap estimators. We hope this work will provide useful guidance to practitioners working with bagged ensemble classifiers designed on small-sample data.

CHAPTER VII

CONCLUSION

In this dissertation, we have presented a study of bootstrap technique in error estimation and ensemble classification methods. This study is aimed at applications in Genomics and Proteomics where the small-sample challenge is prevalent. Reuse of data is expected to increase the accuracy and reliability of error estimation and classification.

In the first part, we have provided the exact formulas for the moments of the variants of bootstrap error estimators, which have been empirically known among the best methods. Based on these results, we obtained the closed form of RMS, which allows us to evaluate the methods globally and hence, to find the optimal bootstrap estimator with the minimum RMS. We believe that this is the first time, as far as we are aware of, that such analysis of bootstrap error estimation is provided.

The second part give us more insights into the bagging classification rules, with respect to the resampling efficiency for different classification rules used to build members of the ensemble. It also provides new observations of the problem of error estimation for bagging classifiers.

Some issues remain to be addressed. In the first part, we assumed the covariance matrix is known. In the case of unknown covariance matrix, the bootstrap estimators have more complexed distributional properties, which require different techniques to solve. Also, our analysis provided here is based on the complete bootstrap, while the bootstrap methods in practice is often its Monte Carlo approximation. Moreover, in the multivariate case, the results are in the forms of noncentral bivariate F distributions, the computations of which are needed to establish. These problems are to be under consideration.

REFERENCES

[1] J. C. Wooley and H. Lin, *Catalyzing inquiry at the interface of computing and biology*, Washington DC: National Academy Press, 2005.

[2] J. van der Greef, P. Stroobant, and R. van der Heijden, “The role of analytical sciences in medical systems biology,” *Curr. Opin. Chem. Biol*, vol. 8, no. 5, pp. 559–565, 2004.

[3] U. M. Braga-Neto, *Lecture notes of ELEN 649*, Department of Electrical Engineering, Texas A&M University, College Station, TX, January 2009.

[4] Affymetrix Inc., “Genechip®human genome u133 arrays,” http://media.affymetrix.com/support/technical/datasheets, November 2010.

[5] Agilent Inc., “Human genome cgh microarray kit 244a,” http://www.genomics.agilent.com/, November 2010.

[6] Illumina Inc., “Revb beadchip kits infinium humanmethylation27,” http://www.illumina.com/products/, November 2010.

[7] E. A. Zerhouni, “US biomedical research: basic, translational, and clinical sciences,” *JAMA*, vol. 294, no. 11, pp. 1352–1358, 2005.

[8] C. Sotiriou and L. Pusztai, “Gene-expression signatures in breast cancer,” *New England Journal of Medicine*, vol. 360, no. 8, pp. 790–800, 2009.

[9] C. Desmedt, E. Ruiz-Garcia, and F. Andre, “Gene expression predictors in breast cancer: current status, limitations and perspectives,” *European Journal of Cancer*, vol. 44, no. 18, pp. 2714–2720, 2008.

[10] S. Koscielny, "Critical review of microarray-based prognostic tests and trials in breast cancer," *Current Opinion in Obstetrics and Gynecology*, vol. 20, no. 1, pp. 47–50, 2008.

[11] L. Harris, H. Fritsche, R. Mennel, L. Norton, P. Ravdin, S. Taube, M. R. Somerfield, D. F. Hayes, and R. C. Bast, "American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer," *Journal of Clinical Oncology*, vol. 25, no. 33, pp. 5287–5312, 2007.

[12] M. Dowsett and A. K. Dunbier, "Emerging biomarkers and new understanding of traditional markers in personalized therapy for breast cancer," *Clinical Cancer Research*, vol. 14, no. 24, pp. 8019–8026, 2008.

[13] J. Subramanian and R. Simon, "Gene expression-based prognostic signatures in lung cancer: ready for clinical use?," *Journal of the National Cancer Institute*, vol. 102, no. 7, pp. 464–474, 2010.

[14] B. J. Wouters, B. Lowenberg, and R. Delwel, "A decade of genome-wide gene expression profiling in acute myeloid leukemia: flashback and prospects," *Blood*, vol. 113, no. 2, pp. 291–298, 2009.

[15] K. S. Hoek, "DNA microarray analyses of melanoma gene expression: a decade in the mines," *Pigment Cell Research*, vol. 20, no. 6, pp. 466–484, 2007.

[16] P. A. Konstantinopoulos, D. Spentzos, and S. A. Cannistra, "Gene-expression profiling in epithelial ovarian cancer," *Nature Clinical Practice Oncology*, vol. 5, no. 10, pp. 577–587, 2008.

[17] A. Walther, E. Johnstone, C. Swanton, R. Midgley, I. Tomlinson, and D. Kerr, "Genetic prognostic and predictive markers in colorectal cancer," *Nature Reviews Can-*

cer, vol. 9, no. 7, pp. 489–499, 2009.

[18] A. Butte, “The use and analysis of microarray data,” *Nature Reviews Drug Discovery*, vol. 1, no. 12, pp. 951–960, 2002.

[19] E. A. Perez, L. Pusztai, and M. van De Vijver, “Improving patient care through molecular diagnostics,” in *Seminars in Oncology*. Elsevier, 2004, vol. 31, pp. 14–20.

[20] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al., “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[21] R. G. W. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, et al., “Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1,” *Cancer Cell*, vol. 17, no. 1, pp. 98–110, 2010.

[22] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, et al., “Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13790–13795, 2001.

[23] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, et al., “Molecular classification of cutaneous malignant melanoma by gene expression profiling,” *Nature*, vol. 406, no. 6795, pp. 536–540, 2000.

[24] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, et al., "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10869–10874, 2001.

[25] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[26] S. Classen, A. Staratschek-Jox, and J. L. Schultze, "Use of genome-wide high-throughput technologies in biomarker development," *Biomarkers Med.*, vol. 2, no. 5, pp. 509–524, 2008.

[27] R. Simon, "Roadmap for developing and validating therapeutically relevant genomic classifiers," *Journal of Clinical Oncology*, vol. 23, no. 29, pp. 7332–7341, 2005.

[28] R. Simon, "Advances in clinical trial designs for predictive biomarker discovery and validation," *Current Breast Cancer Reports*, vol. 1, no. 4, pp. 216–221, 2009.

[29] W. N. Van Wieringen, D. Kun, R. Hampel, and A. L. Boulesteix, "Survival prediction using gene expression data: a review and comparison," *Computational Statistics & Data Analysis*, vol. 53, no. 5, pp. 1590–1603, 2009.

[30] D. G. Beer, S.L.R. Kardia, C. C. Huang, T. J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine*, vol. 8, no. 8, pp. 816–824, 2002.

[31] N. D. Price, L. B. Edelman, I. Lee, H. Yoo, D. Hwang, G. Carlson, D. J. Galas, J. R. Heath, and L. Hood, "Systems biology and systems medicine," in *Essentials of Genomic and Personalized Medicine*, S. G. Geoffrey and F. W. Huntington, Eds., pp. 131 – 141. San Diego: Academic Press, 2010.

[32] C. Sotiriou and M. J. Piccart, "Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?," *Nature Reviews Cancer*, vol. 7, no. 7, pp. 545–553, 2007.

[33] A. Dupuy and R. M. Simon, "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting," *Journal of the National Cancer Institute*, vol. 99, no. 2, pp. 147–157, 2007.

[34] D. F. Ransohoff, "How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design," *Journal of Clinical Epidemiology*, vol. 60, no. 12, pp. 1205–1219, 2007.

[35] R. Simon, "Microarray-based expression profiling and informatics," *Current opinion in biotechnology*, vol. 19, no. 1, pp. 26–29, 2008.

[36] V. M. Coyle and P. G. Johnston, "Genomic markers for decision making: what is preventing us from using markers?," *Nature Reviews Clinical Oncology*, vol. 7, pp. 90–97, 2010.

[37] R. Simon, "Validation of pharmacogenomic biomarker classifiers for treatment selection," *Cancer Biomarkers*, vol. 2, no. 3, pp. 89–96, 2006.

[38] R. Simon, "Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling," *JNCI Cancer Spectrum*, vol. 98, no. 17, pp. 1169, 2006.

[39] S. Raudys and A. K. Jain, “Small sample size effects in statistical pattern recognition: recommendations for practitioners,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, 2002.

[40] A. Jain and D. Zongker, “Feature selection: evaluation, application, and small sample performance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 2002.

[41] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane, “Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification,” *Journal of the National Cancer Institute*, vol. 95, no. 1, pp. 14–18, 2003.

[42] E. R. Dougherty, “Small sample issues for microarray-based classification,” *Comparative and Functional Genomics*, vol. 2, no. 1, pp. 28–34, 2001.

[43] U. M. Braga-Neto and E. R. Dougherty, “Is cross-validation valid for small-sample microarray classification?,” *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[44] W. H. Highleyman, “The design and analysis of pattern recognition experiments,” *Bell Systems Technical Journal*, vol. 41, pp. 723–744, 1962.

[45] G. Hughes, “On the mean accuracy of statistical pattern recognizers,” *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 2002.

[46] D. C. Allais, “The problem of too many measurements in pattern recognition and prediction,” in *Proc. of IEEE Int. Convention Record P-7*. IEEE, 1966, pp. 124–124.

[47] M. Skurichina, “Stabilizing weak classifiers: Regularization and combining techniques in discriminant analysis,” Ph.D. dissertation, Technische Universiteit Delft, Netherlands, 2001.

[48] S. Raudys, "Determination of optimal dimensionality in statistical pattern classification," *Pattern Recognition*, vol. 11, no. 4, pp. 263–270, 1979.

[49] D. Berrar, I. Bradbury, and W. Dubitzky, "Avoiding model selection bias in small-sample genomic datasets," *Bioinformatics*, vol. 22, no. 10, pp. 1245–1250, 2006.

[50] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics*, vol. 19, no. 12, pp. 1484–1491, 2003.

[51] U. M. Braga-Neto, "Small-sample error estimation: mythology versus mathematics," in *Proc. of SPIE*, 2005, vol. 5916, pp. 304–314.

[52] E. B. Laber and S. A. Murphy, "Small sample inference for generalization error in classification using the CUD bound," in *Proc. of the Conference on Uncertainty in Artificial Intelligence*. NIH Public Access, 2008, vol. 2008, pp. 357–365.

[53] A. L. Oberg and O. Vitek, "Statistical design of quantitative mass spectrometry-based proteomic experiments," *Journal of Proteome Research*, vol. 8, no. 5, pp. 2144–2156, 2009.

[54] M. J. Duffy and J. Crown, "A personalized approach to cancer treatment: How biomarkers can help," *Clinical Chemistry*, vol. 54, no. 11, pp. 1770–1779, 2008.

[55] C. Auffray, Z. Chen, and L. Hood, "Systems medicine: the future of medical genomics and healthcare," *Genome*, vol. 1, no. 1, pp. 2, 2009.

[56] A. L. Boulesteix, C. Strobl, T. Augustin, and M. Daumer, "Evaluating microarray-based classifiers: An overview," *Cancer Informatics*, vol. 6, pp. 77–97, 2008.

[57] R. Simon, "The use of genomics in clinical trial design," *Clinical Cancer Research*, vol. 14, no. 19, pp. 5984–5993, 2008.

[58] K. Dobbin and R. Simon, “Sample size determination in microarray experiments for class comparison and prognostic classification,” *Biostatistics*, vol. 6, no. 1, pp. 27–38, 2005.

[59] F. Azuaje, *Bioinformatics and Biomarker Discovery: ”Omic” Data Analysis for Personalized Medicine*, Washington DC: Wiley, 2010.

[60] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, New York: Wiley-Interscience, 2 edition, 2000.

[61] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, New York: Springer, corrected edition, 1996.

[62] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, San Diego: Academic Press, 2 edition, 1990.

[63] G. Mclachlan, *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley-Interscience, 2004.

[64] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: A review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[65] Panel on Discriminant Analysis & Classification & Clustering, “Discriminant analysis and clustering,” *Statistical Science*, vol. 4, no. 1, pp. 34–69, 1989.

[66] T. W. Anderson, “Classification by multivariate analysis,” *Psychometrika*, vol. 16, no. 1, pp. 31–50, 1951.

[67] A. Bowke and R. Sitgreaves, “An asymptotic expansion for the distribution function of the classification statistic W,” *The Miner*, p. 293, 1988.

[68] R. Sitgreaves, “Some results on the distribution of the W-classification statistic,” *The Miner*, p. 241, 1988.

[69] M. Okamoto, “An asymptotic expansion for the distribution of the linear discriminant function,” *The Annals of Mathematical Statistics*, vol. 34, no. 4, pp. 1286–1301, 1963.

[70] F. J. Wyman, D. M. Young, and D. W. Turner, “A comparison of asymptotic error rate expansions for the sample linear discriminant function,” *Pattern Recognition*, vol. 23, no. 7, pp. 775–783, 1990.

[71] A. D. Deev, “Representation of statistics of discriminant analysis and asymptotic expansions when space dimensions are comparable with sample size,” in *Soviet Math. Dokl*, 1970, vol. 11, pp. 1547–1550.

[72] A. D. Deev, “Asymptotic expansions for distributions of statistics W, M, W* in discriminant analysis,” *Statistical Methods of Classification*, vol. 31, pp. 6–57, 1972.

[73] S. Raudys, “On the amount of a priori information in designing the classification algorithm,” *Engineering Cybernetics*, vol. 4, pp. 168–174, 1972.

[74] B. Efron, “The efficiency of logistic regression compared to normal discriminant analysis,” *Journal of the American Statistical Association*, vol. 70, no. 352, pp. 892–898, 1975.

[75] J. W. Sayre, “The distributions of the actual error rates in linear discriminant analysis,” *Journal of the American Statistical Association*, vol. 75, no. 369, pp. pp. 201–205, 1980.

[76] M. J. Schervish, “Asymptotic expansions for the means and variances of error rates,” *Biometrika*, vol. 68, no. 1, pp. 295–299, 1981.

[77] Y. S. Kharin, "The investigation of risk for statistical classifiers using minimum estimators," *Theory of Probability and its Applications*, vol. 28, pp. 623–630, 1984.

[78] G. J. McLachlan, "An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis," *Australian & New Zealand Journal of Statistics*, vol. 15, no. 3, pp. 210–214, 1973.

[79] G. J. McLachlan, "An asymptotic expansion for the variance of the errors of misclassification of the linear discriminant function," *Australian & New Zealand Journal of Statistics*, vol. 14, no. 1, pp. 68–72, 1972.

[80] T. T. Vu and U. M. Braga-Neto, "Is bagging effective in the classification of small-sample genomic and proteomic data?," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2009, pp. 173–202, 2009.

[81] G. Toussaint, "Bibliography on estimation of misclassification," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 472 – 479, jul 1974.

[82] S. Raudys and D. M. Young, "Results in statistical discriminant analysis: a review of the former soviet union literature," *Journal of Multivariate Analysis*, vol. 89, no. 1, pp. 1 – 35, 2004.

[83] D. J. Hand, "Recent advances in error rate estimation," *Pattern Recognition Letters*, vol. 4, no. 5, pp. 335 – 346, 1986.

[84] G. J. Mclachlan, "Error rate estimation in discriminant analysis: Recent advances," *Advances in Multivariate Statistical Analysis*, pp. 233–252, 1987.

[85] R. A. Schiavo and D. J. Hand, "Ten more years of error rate research," *International Statistical Review / Revue Internationale de Statistique*, vol. 68, no. 3, pp. 295–310, 2000.

[86] A. Zollanvari, "Analytic study of performance of error estimators for linear discriminant analysis with applications in genomics," Ph.D. dissertation, Texas A&M University, College Station, TX, 2010.

[87] M. Pawlak, "On the asymptotic properties of smoothed estimators of the classification error rate," *Pattern Recognition*, vol. 21, no. 5, pp. 515–524, 1988.

[88] M. Pawlak and X. Liao, "Estimation of error rates using smoothed estimators," in *Proc. of IEEE ninth International Conference on Pattern Recognition, 1988*. IEEE, 2002, pp. 954–956.

[89] S. M. Snapinn and J. D. Knoke, "An evaluation of smoothed classification error-rate estimators," *Technometrics*, vol. 27, no. 2, pp. 199–206, 1985.

[90] G. E. Tutz, "Smoothed additive estimators for non-error rates in multiple discriminant analysis," *Pattern Recognition*, vol. 18, no. 2, pp. 151–159, 1985.

[91] M. H. Quenouille, "Approximate tests of correlation in time-series," in *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge Univ Press, 1949, vol. 45, pp. 483–484.

[92] J. W. Tukey, "Bias and confidence in not quite large samples," *Annals of Mathematical Statistics*, vol. 29, no. 2, pp. 614, 1958.

[93] J. A. Hartigan, "Using subsample values as typical values," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1303–1317, 1969.

[94] J. A. Hartigan, "Error analysis by replaced samples," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 33, no. 1, pp. 98–110, 1971.

[95] J. A. Hartigan, "Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values," *The Annals of Statistics*, vol. 3, no. 3, pp. 573–580, 1975.

[96] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, pp. 1–26, 1979.

[97] B. Efron, "Computers and the theory of statistics: Thinking the unthinkable," *SIAM Review*, vol. 21, no. 4, pp. 460–480, 1979.

[98] B. Efron, "Nonparametric standard errors and confidence intervals," *Canadian Journal of Statistics*, vol. 9, no. 2, pp. 139–158, 1981.

[99] B. Efron, "Estimating the error rate of a prediction rule: Improvement on cross-validation," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 316–331, 1983.

[100] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, vol. 37, no. 1, pp. 36–48, 1983.

[101] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, New York: Chapman & Hall, 1993.

[102] B. Efron and R. Tibshirani, "Improvements on cross-validation: The .632+ bootstrap method," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 548–560, 1997.

[103] R. B. Kline, *Principles and Practice of Structural Equation Modeling*, New York: The Guilford Press, 2010.

[104] L. Anselin, *Spatial Econometrics: Methods and Models*, New York: Springer, 1988.

[105] B. F. J. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology*, New York: Chapman & Hall, 2007.

[106] K. Singh, "On the asymptotic accuracy of Efron's bootstrap," *The Annals of Statistics*, vol. 9, pp. 1187–1195, 1981.

[107] P. J. Bickel and D. Freedman, "Some asymptotic theory for the bootstrap.," *The Annals of Statistics*, vol. 9, pp. 1196–1217, 1981.

[108] R. Beran, "Estimated sampling distributions: The bootstrap and competitors," *The Annals of Statistics*, vol. 10, no. 1, pp. 212–225, 1982.

[109] G. A. Young, "Bootstrap: more than a stab in the dark? With discussion and a rejoinder by the author," *Stat. Sci.*, vol. 9, no. 3, pp. 382–415, 1994.

[110] M. R. Chernick and V. K. Murthy, "Properties of bootstrap samples," *American Journal of Mathematical and Management Sciences*, vol. 3, no. 5, pp. 161–170, 1985.

[111] J. Shao and D. Tu, *The Jackknife and Bootstrap*, New York: Springer, 1995.

[112] C. Sima and E. R. Dougherty, "Optimal convex error estimators for classification," *Pattern Recognition*, vol. 39, no. 9, pp. 1763–1780, 2006.

[113] M. R. Chernick, V. K. Murthy, and C. D. Nealy, "Application of bootstrap and other resampling techniques: Evaluation of classifier performance," *Pattern Recognition Letters*, vol. 3, no. 3, pp. 167 – 178, 1985.

[114] K. Fukunaga and R. R. Hayes, "Estimation of classifier performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 10, pp. 1087–1101, 1989.

[115] A. C. Davison and P. Hall, "On the bias and variability of bootstrap and cross-validation estimates of error rate in discrimination problems," *Biometrika*, vol. 79, no. 2, pp. 279–284, 1992.

[116] M. R. Chernick, *Bootstrap Methods: A Guide for Practitioners and Researchers*, New Jersey: Wiley-Interscience, 2 edition, 2007.

[117] C. Samprit and C. Sangit, "Estimation of misclassification probabilities by bootstrap methods," *Journal Communications in Statistics - Simulation and Computation*, vol. 12, pp. 645 – 656, 1983.

[118] A. K. Jain, R. C. Dubes, and C. C. Chen, "Bootstrap techniques for error estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 5, pp. 628–633, 1987.

[119] S. Raudys, "On the accuracy of a bootstrap estimate of the classification error," *Proceedings of Ninth International Joint Conference on Pattern Recognition*, pp. 1230–1232, 1988.

[120] U. M. Braga-Neto and E. Dougherty, "Bolstered error estimation," *Pattern Recognition*, vol. 37, no. 6, pp. 1267 – 1281, 2004.

[121] U. M. Braga-Neto, R. Hashimoto, E. R. Dougherty, D. V. Nguyen, and R. J. Carroll, "Is cross-validation better than re-substitution for ranking genes?," *Bioinformatics*, vol. 20, no. 2, pp. 253–258, 2004.

[122] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. of International Joint Conference on Artificial Intelligence*. Citeseer, 1995, vol. 14, pp. 1137–1145.

[123] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics*, vol. 21, no. 15, pp. 3301–3307, 2005.

[124] W. J. Fu, R. J. Carroll, and S. Wang, "Estimating misclassification error with small samples via bootstrap cross-validation," *Bioinformatics*, vol. 21, no. 9, pp. 1979–1986, 2005.

[125] M. R. Chernick, V. K. Murthy, and C. D. Nealy, "Estimation of error rate for linear discriminant functions by resampling: Non-Gaussian populations," *Computers & Mathematics with Applications*, vol. 15, no. 1, pp. 29–37, 1988.

[126] K. Yamada, H. Sakurai, H. Imai, and Y. Sato, "Effects of kurtosis for the error rate estimators using resampling methods in two class discrimination," *Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 340–347, 2009.

[127] P. I. Good, *Resampling Methods: A Practical Guide to Data Analysis*, Boston: Birkhauser, 2001.

[128] G. M. Fitzmaurice, W. J. Krzanowski, and D. J. Hand, "A Monte Carlo study of the 632 bootstrap estimator of error rate," *Journal of Classification*, vol. 8, no. 2, pp. 239–250, 1991.

[129] J. M. P. Sanchez and X. L. O. Cepeda, "The use of Smooth Bootstrap Techniques for Estimating the Error Rate of a Prediction Rule," *Communications in Statistics-Simulation and Computation*, vol. 18, no. 3, pp. 1169–1186, 1989.

[130] K. D. Wernecke and G. Kalb, "Estimation of error rates by means of simulated bootstrap distributions," *Biometrical Journal*, vol. 29, no. 3, pp. 287–292, 1987.

[131] P. Hall and Y. Maesono, "A weighted bootstrap approach to bootstrap iteration," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 1, pp. 137–144, 2000.

[132] R. Tibshirani, "Some applications of the bootstrap in complex problems," in *Exploring the Limits of Bootstrap*, R. LePage and L. Billard, Eds., pp. 271–277. New York: Wiley, 1992.

[133] N. I. Fisher and P. Hall, "Bootstrap algorithms for small samples," *Journal of Statistical Planning and Inference*, vol. 27, no. 2, pp. 157 – 169, 1991.

[134] P. Diaconis and S. Holmes, "Gray codes for randomization procedures," *Statistics and Computing*, vol. 4, no. 4, pp. 287–302, December 1994.

[135] F. Scholz, "The bootstrap small sample properties," Research report, University of Washington, Seattle, WA, June 2007.

[136] P. S. Porter, S. T. Rao, J. Y. Ku, R. L. Poirot, and M. Dakins, "Small sample properties of nonparametric bootstrap t confidence intervals," *Journal of the Air & Waste Management Association*, vol. 47, no. 11, pp. 1197–1203, 1997.

[137] K. Y. F. Chan and S. M. S. Lee, "An exact iterated bootstrap algorithm for small-sample bias reduction," *Comput. Stat. Data Anal.*, vol. 36, no. 1, pp. 1–13, 2001.

[138] E. R. Dougherty, "Validation of inference procedures for gene regulatory networks," *Current Genomics*, vol. 8, no. 6, pp. 351–359, 2007.

[139] I. Shmulevich, E. R. Dougherty, and W. Zhang, "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks," *Proceedings of the IEEE*, vol. 90, no. 11, pp. 1778–1792, 2002.

[140] H. de Jong, “Modeling and simulation of genetic regulatory systems: a literature review,” *Journal of Computational Biology*, vol. 9, no. 1, pp. 67–103, 2002.

[141] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, et al., “Assessing computational tools for the discovery of transcription factor binding sites,” *Nature Biotechnology*, vol. 23, no. 1, pp. 137–144, 2005.

[142] K.T. Hemachandra, “A mathematical framework for expressing multivariate distributions useful in wireless communications,” M.S. thesis, University of Alberta, Edmonton, Alberta, August 2010.

[143] M. A. Moran, “On the expectation of errors of allocation associated with a linear discriminant function,” *Biometrika*, vol. 62, no. 1, pp. 141–148, 1975.

[144] R. Price, “Some non-central F-distributions expressed in closed form,” *Biometrika*, vol. 51, no. 1-2, pp. 107–122, 1964.

[145] J. P. Imhof, “Computing the distribution of quadratic forms in normal variables,” *Biometrika*, vol. 48, no. 3-4, pp. 419–426, 1961.

[146] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, “On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers,” *Pattern Recognition*, vol. 42, no. 11, pp. 2705–2723, 2009.

[147] R. E. Schapire, “The strength of weak learnability,” *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[148] Y. Freund, “Boosting a weak learning algorithm by majority,” *Information and computation*, vol. 121, no. 2, pp. 256–285, 1995.

[149] L. Xu, A. Krzyzak, and C. Y. Suen, “Methods of combining multiple classifiers and their applications to handwriting recognition,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418–435, 2002.

[150] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[151] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[152] L. Lam and S. Y. Suen, “Application of majority voting to pattern recognition: an analysis of its behavior and performance,” *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 27, no. 5, pp. 553–568, 2002.

[153] B. Efron, “The jackknife, the bootstrap and other resampling plans,” in *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, 1982, vol. 38.

[154] S. Alvarez, R. Diaz-Uriarte, A. Osorio, A. Barroso, L. Melchor, M. F. Paz, E. Honrado, R. Rodríguez, M. Urioste, L. Valle, et al., “A predictor based on the somatic genomic changes of the BRCA1/BRCA2 breast cancer tumors identifies the non-BRCA1/BRCA2 tumors with BRCA1 promoter hypermethylation,” *Clinical Cancer Research*, vol. 11, no. 3, pp. 1146, 2005.

[155] E. C. Gunther, D. J. Stone, R. W. Gerwien, P. Bento, and M. P. Heyes, “Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro,” *Science’s STKE*, vol. 100, no. 16, pp. 9608–9613, 2003.

[156] R. Díaz-Uriarte and A. de Andrés, “Gene selection and classification of microarray data using random forest,” *BMC Bioinformatics*, vol. 7, no. 1, pp. 3, 2006.

[157] A. Statnikov, L. Wang, and C. F. Aliferis, “A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification,”

BMC Bioinformatics, vol. 9, no. 1, pp. 319, 2008.

[158] G. Izmirlian, "Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial," *Annals of the New York Academy of Sciences*, vol. 1020, no. The Applications of Bioinformatics in Cancer Detection, pp. 154–174, 2004.

[159] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.

[160] P. Geurts, M. Fillet, D. De Seny, M. A. Meuwis, M. Malaise, M. P. Merville, and L. Wehenkel, "Proteomic mass spectra classification using decision tree based ensemble methods," *Bioinformatics*, vol. 21, no. 14, pp. 3138–3145, 2005.

[161] B. Zhang, T. D. Pham, and Y. Zhang, "Bagging support vector machine for classification of SELDI-TOF mass spectra of ovarian cancer serum samples," in *Proc. of the 20th Australian Joint Conference on Advances in Artificial Intelligence*. Springer, 2007, pp. 820–826.

[162] A. Assareh, M. H. Moradi, and V. Esmaeili, "A novel ensemble strategy for classification of prostate cancer protein mass spectra," in *Proc. of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 5987–5990.

[163] W. Tong, Q. Xie, H. Hong, H. Fang, L. Shi, R. Perkins, and E. F. Petricoin, "Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: assessing chance correlation and prediction confidence," *Environmental Health Perspectives*, vol. 112, no. 16, pp. 1622, 2004.

[164] T. G. Dietterich, “An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization,” *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[165] S. Dudoit, J. Fridlyand, and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” *JAMA*, vol. 97, no. 457, pp. 77–87, 2002.

[166] L. K. Hansen and P. Salamon, “Neural network ensembles,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 2002.

[167] E. Bauer and R. Kohavi, “An empirical comparison of voting classification algorithms: Bagging, boosting, and variants,” *Machine Learning*, vol. 36, no. 1, pp. 105–139, 1999.

[168] S. Y. Sohn and H. W. Shin, “Experimental study for the comparison of classifier combination methods,” *Pattern Recognition*, vol. 40, no. 1, pp. 33–40, 2007.

[169] J. Hua, W. D. Tembe, and E. R. Dougherty, “Performance of feature-selection methods in the classification of high-dimension data,” *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.

[170] E. R. Dougherty, I. Shmulevich, J. Chen, and Z. J. Wang, *Genomic Signal Processing and Statistics*, New York: Hindawi Publishing Corporation, 2005.

[171] M. R. Chernick, “Bootstrap methods: A practitioner’s guide,” *IIE Transactions*, vol. 35, pp. 583–587, 2003.

[172] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, New York: Springer, New York, 2005.

[173] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. of the National Conference on Artificial Intelligence*. John Wiley & Sons Ltd, 1992, pp. 129–129.

[174] V. Van't, J. Laura, D. Hongyue, M. J. Van de Vijver, Y. D. He, A. A. M. Hart, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[175] M. J. Van De Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, et al., "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, no. 25, pp. 1999, 2002.

[176] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13790, 2001.

[177] H. J. Issaq, T. D. Veenstra, T. P. Conrads, and D. Felschow, "The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification," *Biochemical and Biophysical Research Communications*, vol. 292, no. 3, pp. 587–592, 2002.

[178] B. L. Adam, Y. Qu, J. W. Davis, M. D. Ward, M. A. Clements, L. H. Cazares, O. J. Semmes, P. F. Schellhammer, Y. Yasui, Z. Feng, et al., "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men," *Cancer Research*, vol. 62, no. 13, pp. 3609, 2002.

[179] U. M. Braga-Neto, R. Hashimoto, E. R. Dougherty, D. V. Nguyen, and R. J. Carroll, “Is cross-validation better than resubstitution for ranking genes?,” *Bioinformatics*, vol. 20, no. 2, pp. 253–258, 2004.

[180] U. M. Braga-Neto and E. Dougherty, “Exact performance of error estimators for discrete classifiers,” *Pattern Recognition*, vol. 38, no. 11, pp. 1799 – 1814, 2005.

[181] Leo Breiman, “Out-of-bag estimation,” Research report, Statistics Department, University of California, Los Angeles, CA, 1996.

[182] T. Bylander, “Estimating generalization error on two-class datasets using out-of-bag estimates,” *Machine Learning*, vol. 48, no. 1-3, pp. 287–297, 2002.

[183] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, “A comparison of decision tree ensemble creation techniques,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 173–180, 2007.

[184] G. Martínez-Muñoz and A. Suárez, “Out-of-bag estimation of the optimal sample size in bagging,” *Pattern Recognition*, vol. 43, no. 1, pp. 143–152, 2010.

APPENDIX A

PROOFS IN CHAPTER III

Proof of theorem 1:

According to (3.14),

$$E[\hat{\varepsilon}_C] = \frac{m_0(C)}{m(C)} P\{\psi_C(X_1) = 1\} + \frac{m_1(C)}{m(C)} P\{\psi_C(X_{n_0+1}) = 0\}$$

$$\begin{aligned}
P\{\psi_C(X_1) = 1\} &= P\left\{\left(X_1 - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\right)(\hat{\mu}_0^C - \hat{\mu}_1^C) < 0\right\} \\
&= P\left\{X_1 - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} < 0, \hat{\mu}_0^C - \hat{\mu}_1^C > 0\right\} + \\
&\quad + P\left\{X_1 - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} > 0, \hat{\mu}_0^C - \hat{\mu}_1^C < 0\right\} \\
&= P\{B_0 < 0\} + P\{B_0 > 0\}
\end{aligned}$$

$$\begin{aligned}
P\{\psi_C(X_{n_0+1}) = 0\} &= P\left\{\left(X_{n_0+1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\right)(\hat{\mu}_0^C - \hat{\mu}_1^C) > 0\right\} \\
&= P\left\{X_{n_0+1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} < 0, \hat{\mu}_0^C - \hat{\mu}_1^C < 0\right\} + \\
&\quad + P\left\{X_{n_0+1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} > 0, \hat{\mu}_0^C - \hat{\mu}_1^C > 0\right\} \\
&= P\{B_1 < 0\} + P\{B_1 > 0\}
\end{aligned}$$

B_0 and B_1 are two bivariate Gaussian random vectors with the means and covariance matrices as follows:

$$E\left[B_0\right]=\begin{bmatrix}\frac{\mu_0-\mu_1}{2}\\ -\mu_0+\mu_1\end{bmatrix},\ \Sigma_{B_0}=\begin{pmatrix}\left(1+\frac{s_0(C)}{4}\right)\sigma_0^2+\frac{s_1(C)}{4}\sigma_1^2 & \left(-s_1(C)\sigma_1^2+s_0(C)\sigma_0^2\right)/2\\ . & s_0(C)\sigma_0^2+s_1(C)\sigma_1^2\end{pmatrix}. \tag{A.1}$$

$$E\left[B_1\right]=\begin{bmatrix}\frac{\mu_1-\mu_0}{2}\\ \mu_0-\mu_1\end{bmatrix},\ \Sigma_{B_1}=\begin{pmatrix}\left(1+\frac{s_1(C)}{4}\right)\sigma_1^2+\frac{s_0(C)}{4}\sigma_0^2 & \left(-s_1(C)\sigma_1^2+s_0(C)\sigma_0^2\right)/2\\ . & s_0(C)\sigma_0^2+s_1(C)\sigma_1^2\end{pmatrix}. \tag{A.2}$$

Proof of theorem 2:

Following (3.15), we have:

$$\begin{aligned}E[\hat{\varepsilon}_C^2]=&\underbrace{\frac{m_0(C)}{m^2(C)}P\{\psi_C(X)=1|\,X\in\Pi_0\}+\frac{m_1(C)}{m^2(C)}P\{\psi_C(X)=0|\,X\in\Pi_1\}}_{(1)}+\\ &+\sum_{i\neq j=1}^{n_0}I_{C(i)=0}I_{C(j)=0}\underbrace{P\{\psi_C(X_i)=1,\psi_C(X_j)=1\}}_{(2)}+\\ &+\sum_{i\neq j=n_0+1}^{n_0+n_1}I_{C(i)=0}I_{C(j)=0}\underbrace{P\{\psi_C(X_i)=0,\psi_C(X_j)=0\}}_{(3)}+\\ &+\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n_0+n_1}I_{C(i)=0}I_{C(j)=0}\underbrace{P\{\psi_C(X_i)=1,\psi_C(X_j)=0\}}_{(4)}+\\ &+\sum_{j=1}^{n_0}\sum_{i=n_0+1}^{n_0+n_1}I_{C(i)=0}I_{C(j)=0}\underbrace{P\{\psi_C(X_i)=0,\psi_C(X_j)=1\}}_{(5)}.\end{aligned}$$

According to Theorem 1, we have:

$$(1)=\frac{m_0(C)}{m^2(C)}\Big(P\{B_0(C)\geq 0\}+P\{B_0(C)<0\}\Big)+\frac{m_1(C)}{m^2(C)}\Big(P\{B_1(C)\geq 0\}+P\{B_1(C)<0\}\Big).$$

Also,

$$\begin{aligned}(2) = P\Bigg\{ & \left(X_i - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\right)\left(\hat{\mu}_0^C - \hat{\mu}_1^C\right) \; < \; 0, \\ & \left(X_j - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\right)\left(\hat{\mu}_0^C - \hat{\mu}_1^C\right) \; < 0 \mid X_i, X_j \in \Pi_0 \Bigg\} \\ = & P\left\{X_i - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} > 0, \, \hat{\mu}_0^C - \hat{\mu}_1^C < 0, \, X_j - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} > 0\right\} + \\ & + P\left\{X_i - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} < 0, \, \hat{\mu}_0^C - \hat{\mu}_1^C > 0, \, X_j - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} < 0\right\} \\ = & P\{T_{00}(i,j,C) \geq 0\} + P\{T_{00}(i,j,C) < 0\},\end{aligned}$$

where $T_{00}(i,j,C)$ is defined as in the statement of the theorem. It is clear that the mean vector and covariance matrix of $T_{00}(i,j,C)$ are the same for all pairs (i,j). So denote $T_{00}(i,j,C) = T_{00}(C)$.

Similarly,

$$(3) = P\{T_{11}(i,j,C) \geq 0\} + P\{T_{11}(i,j,C) < 0\} = P\{T_{11}(C) \geq 0\} + P\{T_{11}(C) < 0\}$$

$$(4) = P\{T_{01}(i,j,C) \geq 0\} + P\{T_{01}(i,j,C) < 0\} = P\{T_{01}(C) \geq 0\} + P\{T_{01}(C) < 0\}$$

$$(5) = P\{T_{01}(j,i,C) \geq 0\} + P\{T_{01}(j,i,C) < 0\} = P\{T_{01}(C) \geq 0\} + P\{T_{01}(C) < 0\}$$

Theorem 2 follows immediately with $m(C), m_0(C)$ and $m_1(C)$ defined as in (3.32).

Proof of theorem 3:

Following (3.16), we have:

$$\begin{aligned}
E[\hat{\varepsilon}_{C_1}\hat{\varepsilon}_{C_2}] = & \sum_{i,j=1}^{n_0} I_{C_1(i)=0} I_{C_2(j)=0} \underbrace{P\{\psi_{C_1}(X_i)=1,\psi_{C_2}(X_j)=1\}}_{(1)} + \\
& + \sum_{i,j=n_0+1}^{n_0+n_1} I_{C_1(i)=0} I_{C_2(j)=0} \underbrace{P\{\psi_{C_1}(X_i)=0,\psi_{C_2}(X_j)=0\}}_{(2)} + \\
& + \sum_{i=1}^{n_0}\sum_{j=n_0}^{n_0+n_1} I_{C_1(i)=0} I_{C_2(j)=0} \underbrace{P\{\psi_{C_1}(X_i)=1,\psi_{C_2}(X_j)=0\}}_{(3)} + \\
& + \sum_{i=1}^{n_0}\sum_{j=n_0}^{n_0+n_1} I_{C_2(i)=0} I_{C_1(j)=0} \underbrace{P\{\psi_{C_2}(X_i)=1,\psi_{C_1}(X_j)=0\}}_{(4)}.
\end{aligned}$$

$$\begin{aligned}
(1) = & P\Bigg\{ \left(X_i - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2}\right)\left(\hat{\mu}_0^{C_1}-\hat{\mu}_1^{C_1}\right) < 0, \\
& \left(X_j - \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2}\right)\left(\hat{\mu}_0^{C_2}-\hat{\mu}_1^{C_2}\right) < 0 \mid X_i, X_j \in \Pi_0 \Bigg\} \\
= & P\left\{X_i - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2} > 0,\, \hat{\mu}_0^{C_1}-\hat{\mu}_1^{C_1} < 0,\, X_j - \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2} > 0,\, \hat{\mu}_0^{C_2}-\hat{\mu}_1^{C_2} < 0\right\} + \\
& + P\left\{X_i - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2} > 0,\, \hat{\mu}_0^{C_1}-\hat{\mu}_1^{C_1} < 0,\, X_j - \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2} < 0,\, \hat{\mu}_0^{C_2}-\hat{\mu}_1^{C_2} > 0\right\} + \\
& + P\left\{X_i - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2} < 0,\, \hat{\mu}_0^{C_1}-\hat{\mu}_1^{C_1} > 0,\, X_j - \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2} < 0,\, \hat{\mu}_0^{C_2}-\hat{\mu}_1^{C_2} > 0\right\} + \\
& + P\left\{X_i - \frac{\hat{\mu}_0^{C_1}+\hat{\mu}_1^{C_1}}{2} < 0,\, \hat{\mu}_0^{C_1}-\hat{\mu}_1^{C_1} > 0,\, X_j - \frac{\hat{\mu}_0^{C_2}+\hat{\mu}_1^{C_2}}{2} > 0,\, \hat{\mu}_0^{C_2}-\hat{\mu}_1^{C_2} < 0\right\} \\
= & P\{F_{00}^{I}(i,j,C_1,C_2) \geq 0\} + P\{F_{00}^{II}(i,j,C_1,C_2) \geq 0\} + \\
& + P\{F_{00}^{I}(i,j,C_1,C_2) < 0\} + P\{F_{00}^{II}(i,j,C_1,C_2) < 0\},
\end{aligned}$$

where $F_{00}^{I}(i,j,C_1,C_2)$ and $F_{00}^{II}(i,j,C_1,C_2)$ are defined as in (3.33) and (3.34).

Similarly,

$$\begin{aligned}
(2) = F^{I}_{11}(i,j,C_1,C_2) > 0\} + P\{F^{I}_{11}(i,j,C_1,C_2) < 0\} + \\
+ P\{F^{II}_{11}(i,j,C_1,C_2) > 0\} + P\{F^{II}_{11}(i,j,C_1,C_2) < 0\} \\
(3) = P\{F^{I}_{01}(i,j,C_1,C_2) \geq 0\} + P\{F^{I}_{01}(i,j,C_1,C_2) < 0\} + \\
+ P\{F^{II}_{01}(i,j,C_1,C_2) \geq 0\} + P\{F^{II}_{01}(i,j,C_1,C_2) < 0\} \\
(4) = P\{F^{I}_{01}(j,i,C_2,C_1) > 0\} + F^{I}_{01}(j,i,C_2,C_1) < 0\} + \\
+ P\{F^{II}_{01}(j,i,C_2,C_1) \geq 0\} + P\{F^{II}_{01}(j,i,C_1,C_1) < 0\}
\end{aligned}$$

where F^{I}_{11} and F^{II}_{11} are defined as in (3.35) and (3.36) and F^{I}_{01} and F^{II}_{01} are defined as in (3.37) and (3.38). Theorem 3 follows immediately.

Proof of theorem 4:

Following (3.17), we have:

$$\begin{aligned}
E[\hat{\varepsilon}_C \hat{\varepsilon}_r] = \frac{1}{nm(C)} \Bigg[\sum_{i,j=1}^{n_0} I_{C(i)=0} \underbrace{P\{\psi_C(X_i) = 1, \psi(X_j) = 1\}}_{(1)} + \\
+ \sum_{i,j=n+0+1}^{n_0+n_1} I_{C(i)=0} \underbrace{P\{\psi_C(X_i) = 0, \psi(X_j) = 0\}}_{(2)} + \\
+ \sum_{i=1}^{n_0} \sum_{j=n_0}^{n_0+n_1} I_{C(i)=0} \underbrace{P\{\psi_C(X_i) = 1, \psi(X_j) = 0\}}_{(3)} + \\
+ \sum_{i=1}^{n_0} \sum_{j=n_0}^{n_0+n_1} I_{C(j)=0} \underbrace{P\{\psi_C(X_j) = 0, \psi(X_i) = 1\}}_{(4)} \Bigg]
\end{aligned}$$

$$\begin{aligned}
(1) = P\Bigg\{ & \left(X_i - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\right)\left(\hat{\mu}_0^C - \hat{\mu}_1^C\right) \, < \, 0, \\
& \left(X_j - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right)(\hat{\mu}_0 - \hat{\mu}_1) \, < \, 0 \mid X_i, X_j \in \Pi_0 \Bigg\} \\
= P & \left\{ X_i - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} > 0, \, \hat{\mu}_0^C - \hat{\mu}_1^C < 0, X_j - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} > 0, \, \hat{\mu}_0 - \hat{\mu}_1 < 0 \right\} + \\
& + P\left\{ X_i - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} > 0, \, \hat{\mu}_0^C - \hat{\mu}_1^C < 0, X_j - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} < 0, \, \hat{\mu}_0 - \hat{\mu}_1 > 0 \right\} + \\
& + P\left\{ X_i - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} < 0, \, \hat{\mu}_0^C - \hat{\mu}_1^C > 0, X_j - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} < 0, \, \hat{\mu}_0 - \hat{\mu}_1 > 0 \right\} + \\
& + P\left\{ X_i - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} < 0, \, \hat{\mu}_0^C - \hat{\mu}_1^C > 0, X_j - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} > 0, \, \hat{\mu}_0 - \hat{\mu}_1 < 0 \right\} \\
= P & \{G_{00}^{I}(i,j,C) \geq 0\} + P\{G_{00}^{II}(i,j,C) \geq 0\} + \\
& + P\{G_{00}^{I}(i,j,C) < 0\} + P\{G_{00}^{II}(i,j,C) < 0\},
\end{aligned}$$

where $G_{00}^{I}(i,j,C)$ and $G_{00}^{I}(i,j,C)$ are defined as in (3.39) and (3.40).

Similarly,

$$\begin{aligned}
(2) &= G_{11}^{I}(i,j,C) > 0\} + P\{G_{11}^{I}(i,j,C) < 0\} + \\
&\quad + P\{G_{11}^{II}(i,j,C) > 0\} + P\{G_{11}^{II}(i,j,C) < 0\} \\
(3) &= P\{G_{01}^{I}(i,j,C) \geq 0\} + P\{G_{01}^{I}(i,j,C) < 0\} + \\
&\quad + P\{G_{01}^{II}(i,j,C) \geq 0\} + P\{G_{01}^{II}(i,j,C) < 0\} \\
(4) &= P\{G_{10}^{I}(i,j,C) > 0\} + G_{10}^{I}(i,j,C) < 0\} + \\
&\quad + P\{G_{10}^{II}(i,j,C) \geq 0\} + P\{G_{10}^{II}(i,j,C) < 0\}
\end{aligned}$$

where G_{11}^{I} and G_{11}^{II} are defined as in (3.41) and (3.42) and G_{01}^{I} and G_{01}^{II} are defined as in (3.43) and (3.44). Theorem 4 follows immediately.

Proof of theorem 5:

Following (3.19), we have:

$$E[\varepsilon\hat{\varepsilon}_C] = \frac{m_0(C)(1-\gamma)}{m(C)}\underbrace{P\{\psi(X)=1,\psi_C(X_1)=1|X\in\Pi_0\}}_{(1)}+$$
$$+\frac{m_1(C)(1-\gamma)}{m(C)}\underbrace{P\{\psi(X)=1,\psi_C(X_{n_0+1})=0|X\in\Pi_0\}}_{(2)}+$$
$$+\frac{m_0(C)\gamma}{m(C)}\underbrace{P\{\psi(X)=0,\psi_C(X_1)=1|X\in\Pi_1\}}_{(3)}+$$
$$+\frac{m_1(C)\gamma}{m(C)}\underbrace{P\{\psi(X)=0,\psi_C(X_{n_0+1})=0|X\in\Pi_1\}}_{(4)}.$$

$$\begin{aligned}
(1) &= P\{\psi_C(X_1)=1,\psi(X)=1\ |X_1,X\in\Pi_0\} \\
&= P\Big\{\left(X_1-\frac{\hat{\mu}_0^C+\hat{\mu}_1^C}{2}\right)\left(\hat{\mu}_0^C-\hat{\mu}_1^C\right)\ <0,\left(X-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)(\hat{\mu}_0-\hat{\mu}_1)\ <0\ |\ X_1,X\in\Pi_0\Big\} \\
&= P\left\{X_1-\frac{\hat{\mu}_0^C+\hat{\mu}_1^C}{2}>0,\,\hat{\mu}_0^C-\hat{\mu}_1^C<0,X-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}>0,\,\hat{\mu}_0-\hat{\mu}_1<0\right\}+ \\
&\quad +P\left\{X_1-\frac{\hat{\mu}_0^C+\hat{\mu}_1^C}{2}>0,\,\hat{\mu}_0^C-\hat{\mu}_1^C<0,X-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}<0,\,\hat{\mu}_0-\hat{\mu}_1>0\right\}+ \\
&\quad +P\left\{X_1-\frac{\hat{\mu}_0^C+\hat{\mu}_1^C}{2}<0,\,\hat{\mu}_0^C-\hat{\mu}_1^C>0,X-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}<0,\,\hat{\mu}_0-\hat{\mu}_1>0\right\}+ \\
&\quad +P\left\{X_1-\frac{\hat{\mu}_0^C+\hat{\mu}_1^C}{2}<0,\,\hat{\mu}_0^C-\hat{\mu}_1^C>0,X-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}>0,\,\hat{\mu}_0-\hat{\mu}_1<0\right\} \\
&= P\{K_{00}^I(C)\geq 0\}+P\{K_{00}^{II}(C)\geq 0\}++P\{K_{00}^I(C)<0\}+P\{K_{00}^{II}(C)<0\},
\end{aligned}$$

Similarly,

$$
\begin{aligned}
(2) &= K^{I}_{11}(C) > 0\} + P\{K^{I}_{11}(C) < 0\}+ \\
&\quad + P\{K^{II}_{11}(C) > 0\} + P\{K^{II}_{11}(C) < 0\} \\
(3) &= P\{K^{I}_{01}(C) \geq 0\} + P\{K^{I}_{01}(C) < 0\}+ \\
&\quad + P\{K^{II}_{01}(C) \geq 0\} + P\{K^{II}_{01}(C) < 0\} \\
(4) &= P\{K^{I}_{10}(C) > 0\} + K^{I}_{10}(C) < 0\}+ \\
&\quad + P\{K^{II}_{10}(C) \geq 0\} + P\{K^{II}_{10}(C) < 0\}
\end{aligned}
$$

where K^{I}_{11} and K^{II}_{11} are defined as in (3.49) and (3.50) and K^{I}_{01} and K^{II}_{01} are defined as in (3.51) and (3.52). Theorem 5 follows immediately.

Algorithm to compute $P(s_0, s_1)$

In the full bootstrap sampling case, $P(s_0, s_1) = P(s)$, where

$$
s = \frac{1}{n^2} \sum_{i=1}^{n} C(i)^2 \,, \tag{A.3}
$$

whereas in the stratified bootstrap sampling case, $P(s_0, s_1) = P(s_0)P(s_1)$. We limit ourselves therefore to describe the algorithm to compute $P(s)$ for a generic bootstrap vector of size n. Let

$$
S_n(x,y) = \frac{n!}{n^n} \sum_{\substack{i_1+\cdots+i_n=x \\ i_1^2+\cdots+i_n^2=y}} \frac{1}{i_1!\ldots i_n!} \,, \quad x,y \in Z^+, \, x \geq 1, \, \frac{x^2}{n} \leq y \leq x^2 \,. \tag{A.4}
$$

Clearly, $P(s) = S_n(n, n^2 s)$. Now, notice that

$$\begin{aligned} S_n(x,y) &= \frac{n!}{n^n} \sum_{\substack{i_1+\cdots+i_n=x \\ i_1^2+\cdots+i_n^2=y}} \frac{1}{i_1!\dots i_n!} = \frac{n!}{n^n} \sum_{j=0}^{n} \sum_{\substack{i_2+\cdots+i_n=x-j \\ i_2^2+\cdots+i_n^2=y-j^2}} \frac{1}{j!\,i_2!\dots i_n!} \\ &= \left(\frac{n-1}{n}\right)^{n-1} \sum_{j=0}^{x} \frac{1}{j!} \frac{(n-1)!}{(n-1)^{n-1}} \sum_{\substack{i_2+\cdots+i_n=x-j \\ i_2^2+\cdots+i_n^2=y-j^2}} \frac{1}{i_2!\dots i_n!} \qquad \text{(A.5)} \\ &= \left(\frac{n-1}{n}\right)^{n-1} \sum_{j=0}^{x} \frac{1}{j!} S_{n-1}(x-j, y-j^2)\,. \end{aligned}$$

This, together with the fact that

$$S_1(x,y) = \begin{cases} 1/x!\,, & \text{if } y = x^2 \\ 0\,, & \text{otherwise} \end{cases}\,, \qquad \text{(A.6)}$$

provides an efficient recursive algorithm to compute $P(s)$ up to moderate sample size n. The details of the computation for the purposes of this paper were as follows: we set the maximum value of n to 200 and stored values of $S_n(x,y)$ as a matrix of size 200×200, for each n. For $n = 1$, $S_1(x,y)$ has nonzero values at the positions (i, i^2), for $i = 1, 2, \dots 200$ only, c.f. (A.6). Then we compute $S_n(x,y)$ based on the value of $S_{n-1}(x,y)$ recursively through (A.5). Each matrix of size 200×200, corresponding to one value of n, took around three minutes to compute on a state-of-the-art computer*. In all, it took less than twelve hours for compute all the values of $S_n(x,y)$ up to $n = 200$. For each value of n, the probabilities $P(s) = S_n(n, n^2 s)$ were extracted from the table for $S_n(x,y)$ and saved separately to be used in the numerical examples.

*An I-Mac Intel Core 2 Duo 2.4 GHz with 2GB RAM.

APPENDIX B

PROOFS IN CHAPTER IV

Proof of Theorem 13:

From (3.14),

$$
\begin{aligned}
E[\hat{\varepsilon}_C] &= \frac{m_0(C)}{m(C)} P\left\{\psi_C(X_1) = 1\right\} + \frac{m_1(C)}{m(C)} P\{\psi_C(X_{n_0+1}) = 0\} \\
&= \frac{m_0(C)}{m(C)} \underbrace{P\Big\{ \left(\hat{\mu}_0^C - \hat{\mu}_1^C\right)^T \Sigma^{-1} \left(X_1 - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\right) < 0 \Big\}}_{(1)} + \\
&\quad + \frac{m_1(C)}{m(C)} \underbrace{P\Big\{ \left(\hat{\mu}_0^C - \hat{\mu}_1^C\right)^T \Sigma^{-1} \left(X_{n_0+1} - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\right) > 0 \Big\}}_{(2)}
\end{aligned}
$$

$$
\begin{aligned}
(1) &= P\{U^T V < 0\} \\
&= P\big\{(U+V)^T(U+V) - (U-V)^T(U-V) < 0\big\} \\
&= P\big\{(Z_0^1)^T Z_0^1 - (Z_0^2)^T Z_0^2 < 0\big\} \\
&= G_0\left(Z_0\right)
\end{aligned}
$$

where

$$
\begin{aligned}
U &= s^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\widehat{\mu}_0^C - \widehat{\mu}_1^C), \qquad & V &= 2(s+4)^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \left(X_1 - \frac{\widehat{\mu}_0^C + \widehat{\mu}_1^C}{2}\right), \\
Z_0^1 &= U + V, \;\; Z_0^2 = U - V, & Z_0 &= [(Z_0^1)^T (Z_0^2)^T]^T,
\end{aligned}
$$

where s is defined as in (3.12). It is clear that U and V are p-dimensional Gaussian random variables with dispersion matrix I_p. As a results, Z_0^1 and Z_0^2 are also p-dimensional Gaussian random variables with the means and covariance matrices as followings:

$$E\left[Z_0^1\right] = \left(s^{-\frac{1}{2}} + (s+4)^{-\frac{1}{2}}\right)\Sigma^{-\frac{1}{2}}(\mu_0 - \mu_1),$$
$$E\left[Z_0^2\right] = \left(s^{-\frac{1}{2}} - (s+4)^{-\frac{1}{2}}\right)\Sigma^{-\frac{1}{2}}(\mu_0 - \mu_1),$$
$$\Sigma_{Z_0^1, Z_0^2} = \mathbf{0_{p\times p}},\ \Sigma_{Z_0^1} = 2(1+\rho)I_p,\ \ \Sigma_{Z_0^2} = 2(1-\rho)I_p, \text{ where } \rho = \frac{s_0 - s_1}{\sqrt{s(s+4)}}.$$

Similarly for (2)

$$(2) = P\left\{(Z_1^1)^T Z_1^1 - (Z_1^2)^T Z_1^2 > 0\right\} = G_1(Z_1).$$

Theorem 13 follows immediately.

Proof of Theorem 14:

From (3.15), we have

$$E[\hat{\varepsilon}_C^2] = \underbrace{\frac{m_0(C)}{m^2(C)} P\{\psi_C(X) = 1 |\, X \in \Pi_0\} + \frac{m_1(C)}{m^2(C)} P\{\psi_C(X) = 0 |\, X \in \Pi_1\}}_{(1)} +$$
$$+ \frac{1}{m^2(C)} \left[\sum_{i=1}^{n_0} \sum_{j\neq i}^{n_0} I_{C(i)=0} I_{C(j)=0} \underbrace{P\{\psi_C(X_i) = 1, \psi_C(X_j) = 1\}}_{(2)} + \right.$$
$$+ \sum_{i=n_0+1}^{n_0+n_1} \sum_{j\neq i}^{n_0+n_1} I_{C(i)=0} I_{C(j)=0} \underbrace{P\{\psi_C(X_i) = 0, \psi_C(X_j) = 0\}}_{(3)} +$$
$$+ \sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n_0+n_1} I_{C(i)=0} I_{C(j)=0} \underbrace{P\{\psi_C(X_i) = 1, \psi_C(X_j) = 0\}}_{(4)} +$$
$$\left. + \sum_{j=1}^{n_0} \sum_{i=n_0+1}^{n_0+n_1} I_{C(i)=0} I_{C(j)=0} \underbrace{P\{\psi_C(X_i) = 0, \psi_C(X_j) = 1\}}_{(5)} \right].$$

The term (1) is obtained using Theorem 13. Consider (2)

$$\begin{aligned}
(2) &= P\{\psi_C(X_i) = 1, \psi_C(X_j) = 1 | X_i, X_j \in \Pi_0\} \text{ with } C(i) = C(j) = 0 \\
&= P\Big\{(\hat{\mu}_0^C - \hat{\mu}_1^C)^T \Sigma^{-1}\Big(X_i - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\Big) < 0, \\
&\quad (\hat{\mu}_0^C - \hat{\mu}_1^C)^T \Sigma^{-1}\Big(X_j - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2}\Big) < 0 | X_i, X_j \in \Pi_0\Big\} \\
&= P\big\{U^T V_i < 0, U^T V_j < 0\big\} \\
&= P\big\{(U + V_i)^T (U + V_i) - (U - V_i)^T (U - V_i) < 0, \\
&\quad (U + V_j)^T (U + V_j) - (U - V_j)^T (U - V_j) < 0\big\} \\
&= P\big\{(T_{00}^1)^T T_{00}^1 - (T_{00}^2)^T T_{00}^2 < 0, (T_{00}^3)^T T_{00}^3 - (T_{00}^4)^T T_{00}^4 < 0\big\} \\
&= G_{00}(T_{00})
\end{aligned}$$

where

$$\begin{aligned}
U &= s^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\widehat{\mu}_0^C - \widehat{\mu}_1^C), \\
V_i &= 2(s+4)^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \Big(X_i - \frac{\widehat{\mu}_0^C + \widehat{\mu}_1^C}{2}\Big), X_i \in \Pi_0, \\
V_j &= 2(s+4)^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \Big(X_j - \frac{\widehat{\mu}_0^C + \widehat{\mu}_1^C}{2}\Big), X_j \in \Pi_0, \\
T_{00} &= [(T_{00}^1)^T \, (T_{00}^2)^T \, (T_{00}^3)^T \, (T_{00}^4)^T]^T,
\end{aligned}$$

where s is defined as in (3.12), and $T_{00}^1 = U + V_i$, $T_{00}^2 = U - V_i$, $T_{00}^3 = U + V_j$, $T_{00}^4 = U - V_j$. It is clear that U, V_i, and V_j are p-dimensional Gaussian random variables with dispersion matric I_p. As a results, T_{00}^i, $i = 1, 2, 3, 4$ are also p-dimensional Gaussian random variables.

Basic algebra gives us

$$
\begin{aligned}
E\left[T_{00}^{1}\right] &= E\left[T_{00}^{3}\right] = \big(s^{-\frac{1}{2}} + (s+4)^{-\frac{1}{2}}\big)\Sigma^{-\frac{1}{2}}(\mu_0-\mu_1),\\
E\left[T_{00}^{2}\right] &= E\left[T_{00}^{4}\right] = \big(s^{-\frac{1}{2}} - (s+4)^{-\frac{1}{2}}\big)\Sigma^{-\frac{1}{2}}(\mu_0-\mu_1),\\
\Sigma_{T_{00}^{1}} &= \Sigma_{T_{00}^{3}} = 2\left(1+\rho(C)\right)I_p,\\
\Sigma_{T_{00}^{2}} &= \Sigma_{T_{00}^{4}} = 2\left(1-\rho(C)\right)I_p, \text{ where } \rho(C) = \frac{s_0-s_1}{\sqrt{s(s+4)}}.\\
\Sigma_{T_{00}^{1},T_{00}^{2}} &= \Sigma_{T_{00}^{3},T_{00}^{4}} = \mathbf{0_{p\times p}}
\end{aligned}
$$

We are to compute $\Sigma_{T_{00}^{1},T_{00}^{3}}$.

$$
\begin{aligned}
\Sigma_{V_i,V_j} &= E\Bigg[\left(2(s+4)^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}\left(X_i - \frac{\widehat{\mu}_0^C+\widehat{\mu}_1^C}{2}\right) - EV_i\right)\times\\
&\quad\times\left(2(s+4)^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}\left(X_j - \frac{\widehat{\mu}_0^C+\widehat{\mu}_1^C}{2}\right) - EV_j\right)^T\Bigg]\\
&= \frac{s}{s+4}I_p,\\
\Sigma_{U,V_i} &= E\Bigg[\left(s^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(\widehat{\mu}_0^C-\widehat{\mu}_1^C) - EU\right)\times\\
&\quad\times\left(2(s+4)^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}\left(X_i - \frac{\widehat{\mu}_0^C+\widehat{\mu}_1^C}{2}\right) - EV_i\right)^T |X_i\in\Pi_0\Bigg]\\
&= 2\frac{1}{\sqrt{s(s+4)}}\Sigma^{-\frac{1}{2}}\left(-\frac{1}{2}\Sigma_{\hat{\mu}_0^C} + \frac{1}{2}\Sigma_{\hat{\mu}_1^C}\right)\Sigma^{-\frac{1}{2}}\\
&= \frac{s_1-s_0}{\sqrt{s(s+4)}}I_p\\
&= \Sigma_{U,V_j}.
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_{T_{00}^{1},T_{00}^{3}} &= Cov(U+V_i, U+V_j) = \Sigma_U + \Sigma_{U,V_i} + \Sigma_{U,V_j} + \Sigma_{V_i,V_j}\\
&= \left(1 + \frac{2(s_1-s_0)}{\sqrt{s(s+4)}} + \frac{s}{s+4}\right)I_p,
\end{aligned}
$$

$$\begin{aligned}\Sigma_{T_{00}^1,T_{00}^4} &= Cov(U+V_i,U-V_j) = \Sigma_U + \Sigma_{U,V_i} - \Sigma_{U,V_j} - \Sigma_{V_i,V_j} \\ &= \frac{2s+4}{s+4} I_p = \Sigma_{T_{00}^2,T_{00}^3},\end{aligned}$$

$$\begin{aligned}\Sigma_{T_{00}^2,T_{00}^4} &= Cov(U-V_i,U-V_j) = \Sigma_U - \Sigma_{U,V_i} - \Sigma_{U,V_j} + \Sigma_{V_i,V_j} \\ &= \left(\frac{s}{s+4} - \frac{2(s_1-s_0)}{\sqrt{s(s+4)}}\right) I_p\end{aligned}$$

So,

$$\Sigma_{T_{00}} = \begin{pmatrix} 2(1+\rho)I_p & \mathbf{0_{p\times p}} & \left(\frac{2s+4}{s+4} + \frac{2(s_1-s_0)}{\sqrt{s(s+4)}}\right)I_p & \frac{2s+4}{s+4}I_p \\ . & 2(1-\rho)I_p & \frac{2s+4}{s+4}I_p & \left(\frac{2s+4}{s+4} - \frac{2(s_1-s_0)}{\sqrt{s(s+4)}}\right)I_p \\ . & . & 2(1+\rho)I_p & \mathbf{0_{p\times p}} \\ . & . & . & 2(1-\rho)I_p \end{pmatrix}.$$

Similarly for (3), (4), and (5):

$$\begin{aligned}(3) &= P\{(T_{11}^1)^T T_{11}^1 - (T_{11}^2)^T T_{11}^2 > 0, (T_{11}^3)^T T_{11}^3 - (T_{11}^4)^T T_{11}^4 > 0\} = G_{11}(T_{11}), \\ (4) &= P\{(T_{01}^1)^T T_{01}^1 - (T_{01}^2)^T T_{01}^2 < 0, (T_{01}^3)^T T_{01}^3 - (T_{01}^4)^T T_{01}^4 > 0\} = G_{01}(T_{01}), \\ (5) &= P\{(T_{01}^1)^T T_{01}^1 - (T_{01}^2)^T T_{01}^2 > 0, (T_{01}^3)^T T_{01}^3 - (T_{01}^4)^T T_{01}^4 < 0\} = G_{10}(T_{01}).\end{aligned}$$

with

$$\begin{aligned}E\left[T_{11}^1\right] = E\left[T_{11}^3\right] = E\left[T_{01}^2\right] = E\left[T_{01}^3\right] &= \left(s^{-\frac{1}{2}} - (s+4)^{-\frac{1}{2}}\right)\Sigma^{-\frac{1}{2}}(\mu_0-\mu_1), \\ E\left[T_{11}^2\right] = E\left[T_{11}^4\right] = E\left[T_{01}^1\right] = E\left[T_{01}^4\right] &= \left(s^{-\frac{1}{2}} + (s+4)^{-\frac{1}{2}}\right)\Sigma^{-\frac{1}{2}}(\mu_0-\mu_1),\end{aligned}$$

$$\Sigma_{T_{11}} = \Sigma_{T_{01}} = \Sigma_{T_{00}}.$$

Theorem 14 follows immediately.

Proof of Theorem 15:

The same technique in the proof of Theorem 14 is applied for Theorem 15. From (3.16), we have

$$
\begin{aligned}
E[\hat{\varepsilon}_{C_1}\hat{\varepsilon}_{C_2}] = \frac{1}{m(C_1)m(C_2)} \Bigg[& \sum_{i=1}^{n_0}\sum_{j=1}^{n_0} I_{C_1(i)=0} I_{C_2(j)=0} \underbrace{P\{\psi_{C_1}(X_i)=1, \psi_{C_2}(X_j)=1\}}_{(1)} + \\
& + \sum_{i=n_0+1}^{n_0+n_1}\sum_{j=n_0+1}^{n_0+n_1} I_{C_1(i)=0} I_{C_2(j)=0} \underbrace{P\{\psi_{C_1}(X_i)=0, \psi_{C_2}(X_j)=0\}}_{(2)} + \\
& + \sum_{i=1}^{n_0}\sum_{j=n_0}^{n_0+n_1} I_{C_1(i)=0} I_{C_2(j)=0} \underbrace{P\{\psi_{C_1}(X_i)=1, \psi_{C_2}(X_j)=0\}}_{(3)} + \\
& + \sum_{i=1}^{n_0}\sum_{j=n_0}^{n_0+n_1} I_{C_2(i)=0} I_{C_1(j)=0} \underbrace{P\{\psi_{C_2}(X_i)=1, \psi_{C_1}(X_j)=0\}}_{(4)} \Bigg].
\end{aligned}
$$

Consider (1)

$$
\begin{aligned}
(1) &= P\{\psi_{C_1}(X_i)=1, \psi_{C_2}(X_j)=1 | X_i, X_j \in \Pi_0\} \text{ with } C_1(i)=C_2(j)=0 \\
&= P\Big\{ \left(\hat{\mu}_0^{C_1} - \hat{\mu}_1^{C_1}\right)^T \Sigma^{-1} \left(X_i - \frac{\hat{\mu}_0^{C_1} + \hat{\mu}_1^{C_1}}{2} \right) < 0, \\
&\quad \left(\hat{\mu}_0^{C_2} - \hat{\mu}_1^{C_2}\right)^T \Sigma^{-1} \left(X_j - \frac{\hat{\mu}_0^{C_2} + \hat{\mu}_1^{C_2}}{2} \right) < 0 | X_i, X_j \in \Pi_0 \Big\} \\
&= P\{U_1^T V_1 < 0, U_2^T V_2 < 0\} \\
&= P\{(U_1+V_1)^T(U_1+V_1) - (U_1-V_1)^T(U_1-V_1) < 0, \\
&\quad (U_2+V_2)^T(U_2+V_2) - (U_2-V_2)^T(U_2-V_2) < 0\} \\
&= P\big\{ (F_{00}^1)^T F_{00}^1 - (F_{00}^2)^T F_{00}^2 < 0, (F_{00}^3)^T F_{00}^3 - (F_{00}^4)^T F_{00}^4 < 0 \big\} \\
&= G_{00}\big(F_{00}(C_1, C_2, i, j)\big)
\end{aligned}
$$

where

$$\begin{aligned}
U_i &= s(C_i)^{-1/2}\Sigma^{-\frac{1}{2}}(\widehat{\mu}_0^{C_i} - \widehat{\mu}_1^{C_i}),\, i \in (1,2)\\
V_1 &= 2(s(C_1)+4)^{-1/2}\Sigma^{-\frac{1}{2}}\left(X_i - \frac{\widehat{\mu}_0^{C_1} + \widehat{\mu}_1^{C_1}}{2}\right), X_i \in \Pi_0,\\
V_2 &= 2(s(C_2)+4)^{-1/2}\Sigma^{-\frac{1}{2}}\left(X_j - \frac{\widehat{\mu}_0^{C_2} + \widehat{\mu}_1^{C_2}}{2}\right), X_j \in \Pi_0,\\
F_{00} &= [(F_{00}^1)^T\,(F_{00}^2)^T\,(F_{00}^3)^T\,(F_{00}^4)^T]^T,
\end{aligned}$$

and, $F_{00}^1 = U_1 + V_1, F_{00}^2 = U_1 - V_1, F_{00}^3 = U_2 + V_2, F_{00}^4 = U_2 - V_2$. Basic algebra gives us

$$\begin{aligned}
E\left[F_{00}^1\right] &= \left(s(C_1)^{-1/2} + (s(C_1)+4)^{-1/2}\right)\Sigma^{-\frac{1}{2}}(\mu_0 - \mu_1),\\
E\left[F_{00}^2\right] &= \left(s(C_1)^{-1/2} - (s(C_1)+4)^{-1/2}\right)\Sigma^{-\frac{1}{2}}(\mu_0 - \mu_1),\\
E\left[F_{00}^3\right] &= \left(s(C_2)^{-1/2} + (s(C_2)+4)^{-1/2}\right)\Sigma^{-\frac{1}{2}}(\mu_0 - \mu_1),\\
E\left[F_{00}^4\right] &= \left(s(C_2)^{-1/2} - (s(C_2)+4)^{-1/2}\right)\Sigma^{-\frac{1}{2}}(\mu_0 - \mu_1).
\end{aligned}$$

$$\begin{aligned}
\Sigma_{F_{00}^1} &= 2\,(1+\rho(C_1))\,I_p & \Sigma_{F_{00}^2} &= 2\,(1-\rho(C_1))\,I_p,\\
\Sigma_{F_{00}^3} &= 2\,(1+\rho(C_2))\,I_p & \Sigma_{F_{00}^4} &= 2\,(1-\rho(C_2))\,I_p,\\
\Sigma_{F_{00}^1,F_{00}^2} &= \Sigma_{F_{00}^3,F_{00}^4} = \mathbf{0_{p\times p}}
\end{aligned}$$

Basic algebra give us:

$$\begin{aligned}
\Sigma_{U_1,U_2} &= \frac{r_0 + r_1}{\sqrt{s(C_1)s(C_2)}} I_p, & \Sigma_{U_2,V_1} &= \frac{\frac{2C_2(i)}{n_0} - r_0 + r_1}{\sqrt{s(C_2)(s(C_1)+4)}} I_p,\\
\Sigma_{U_1,V_2} &= \frac{\frac{2C_1(j)}{n_0} - r_0 + r_1}{\sqrt{s(C_1)(s(C_2)+4)}} I_p, & \Sigma_{V_1,V_2} &= \frac{r_0 + r_1 - \frac{2C_1(j)+2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}} I_p.
\end{aligned}$$

$$\begin{aligned}\Sigma_{F_{00}^1,F_{00}^3} &= \Sigma_{U_1,U_2}+\Sigma_{U_1,V_2}+\Sigma_{U_2,V_1}+\Sigma_{V_1,V_2}\\ &= \Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}+\frac{\frac{2C_1(j)}{n_0}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}+\\ &\quad+\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}+\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big)I_p\\ &= \kappa_{001}(C_1,C_2,i,j)I_p.\end{aligned}$$

Similarly,

$$\begin{aligned}\Sigma_{F_{00}^1,F_{00}^4} &= \Sigma_{U_1,U_2}-\Sigma_{U_1,V_2}+\Sigma_{U_2,V_1}-\Sigma_{V_1,V_2}\\ &= \Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}-\frac{\frac{2C_1(j)}{n_0}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}+\\ &\quad+\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}-\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big)I_p\\ &= \kappa_{002}(C_1,C_2,i,j)I_p,\\ \Sigma_{F_{00}^2,F_{00}^3} &= \Sigma_{U_1,U_2}+\Sigma_{U_1,V_2}-\Sigma_{U_2,V_1}-\Sigma_{V_1,V_2}\\ &= \Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}+\frac{\frac{2C_1(j)}{n_0}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}-\\ &\quad-\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}-\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big)I_p\\ &= \kappa_{003}(C_1,C_2,i,j)I_p,\\ \Sigma_{F_{00}^2,F_{00}^4} &= \Sigma_{U_1,U_2}-\Sigma_{U_1,V_2}-\Sigma_{U_2,V_1}+\Sigma_{V_1,V_2}\\ &= \Big(\frac{r_0+r_1}{\sqrt{s(C_1)s(C_2)}}-\frac{\frac{2C_1(j)}{n_0}-r_0+r_1}{\sqrt{s(C_1)(s(C_2)+4)}}-\\ &\quad-\frac{\frac{2C_2(i)}{n_0}-r_0+r_1}{\sqrt{s(C_2)(s(C_1)+4)}}+\frac{r_0+r_1-\frac{2C_1(j)+2C_2(i)}{n_0}}{\sqrt{(s(C_1)+4)(s(C_2)+4)}}\Big)I_p\\ &= \kappa_{004}(C_1,C_2,i,j)I_p.\end{aligned}$$

So,

$$\Sigma_{F_{00}(C_1,C_2,i,j)} = \begin{pmatrix} (1+\rho(C_1))I_p & \mathbf{0_{p\times p}} & \kappa_{001}I_p & \kappa_{002}I_p \\ . & (1-\rho(C_1))I_p & \kappa_{003}I_p & \kappa_{004}I_p \\ . & . & (1+\rho(C_2))I_p & \mathbf{0_{p\times p}} \\ . & . & . & (1-\rho(C_2))I_p \end{pmatrix}.$$

Similarly for (3), (4), and (5).

$$(3) = P\{(F_{11}^1)^T F_{11}^1 - (F_{11}^2)^F F_{11}^2 > 0, (F_{11}^3)^T F_{11}^3 - (F_{11}^4)^F F_{11}^4 > 0\} = G_{11}(F_{11}(C_1,C_2,i,j)),$$

$$(4) = P\{(F_{01}^1)^T F_{01}^1 - (F_{01}^2)^F F_{01}^2 < 0, (F_{01}^3)^T F_{01}^3 - (F_{01}^4)^F F_{01}^4 > 0\} = G_{01}(F_{01}(C_1,C_2,i,j)),$$

$$(5) = P\{(F_{00}^1)^T F_{00}^1 - (F_{00}^2)^F F_{00}^2 > 0, (F_{00}^3)^T F_{00}^3 - (F_{00}^4)^F F_{00}^4 < 0\} = G_{00}(F_{01}(C_2,C_1,j,i)).$$

F_{11} and F_{01} are 4p-dimensional Gaussian random variables specified as in Theorem 15.

Proof of Theorem 16:

The same technique in the proof of Theorem 14 is applied for Theorem 16. From (3.17), we have

$$\begin{aligned} E[\hat{\varepsilon}_C\hat{\varepsilon}_r] = \frac{1}{nm(C)} \Bigg[& \sum_{i=1}^{n_0}\sum_{j=1}^{n_0} I_{C(i)=0}\, \underbrace{P\{\psi_C(X_i)=1, \psi(X_j)=1\}}_{(1)} + \\ & + \sum_{i=n_0+1}^{n_0+n_1}\sum_{j=n_0+1}^{n_0+n_1} I_{C(i)=0}\, \underbrace{P\{\psi_C(X_i)=0, \psi(X_j)=0\}}_{(2)} + \\ & + \sum_{i=1}^{n_0}\sum_{j=n_0}^{n_0+n_1} I_{C(i)=0}\, \underbrace{P\{\psi_C(X_i)=1, \psi(X_j)=0\}}_{(3)} + \\ & + \sum_{i=1}^{n_0}\sum_{j=n_0}^{n_0+n_1} I_{C(j)=0}\, \underbrace{P\{\psi_C(X_j)=0, \psi(X_i)=1\}}_{(4)} \Bigg]. \end{aligned}$$

Consider (1)

$$\begin{aligned}
(1) &= P\{\psi_C(X_i) = 1, \psi(X_j) = 1 | X_i, X_j \in \Pi_0\}, \quad C(l) = 0 \\
&= P\{(\hat{\mu}_0^C - \hat{\mu}_1^C)^T \Sigma^{-1} \left(X_i - \frac{\hat{\mu}_0^C + \hat{\mu}_1^C}{2} \right) < 0, \\
&\quad (\hat{\mu}_0 - \hat{\mu}_1)^T \Sigma^{-1} \left(X_j - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} \right) < 0 | X_i, X_j \in \Pi_0\} \\
&= P\{U_1^T V_1 < 0, U_2^T V_2 < 0 | X_i, X_j \in \Pi_0\} \\
&= P\{(U_1 + V_1)^T (U_1 + V_1) - (U_1 - V_1)^T (U_1 - V_1) < 0, \\
&\quad (U_2 + V_2)^T (U_2 + V_2) - (U_2 - V_2)^T (U_2 - V_2) < 0 | X_i, X_j \in \Pi_0\} \\
&= P\{(M_{00}^1)^T M_{00}^1 - (M_{00}^2)^T M_{00}^2 < 0, \, (M_{00}^3)^T M_{00}^3 - (M_{00}^4)^T M_{00}^4 < 0\} \\
&= G_{00}(M_{00})
\end{aligned}$$

where $X_i \in \Pi_0, X_j \in \Pi_0, C(i) = 0$, and

$$\begin{aligned}
U_1 &= s^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\widehat{\mu}_0^C - \widehat{\mu}_1^C), & V_1 &= 2(s+4)^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \left(X_i - \frac{\widehat{\mu}_0^C + \widehat{\mu}_1^C}{2} \right), \\
U_2 &= \left(\frac{1}{n_0} + \frac{1}{n_1} \right)^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\widehat{\mu}_0 - \widehat{\mu}_1), & V_2 &= \left(1 - \frac{3}{4n_0} + \frac{1}{4n_1} \right)^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \left(X_j - \frac{\widehat{\mu}_0 + \widehat{\mu}_1}{2} \right),
\end{aligned}$$

and, $M_{00} = [(M_{00}^1)^T \, (M_{00}^2)^T \, (M_{00}^3)^T \, (M_{00}^4)^T]^T$,
$M_{00}^1 = U_1 + V_1, M_{00}^2 = U_1 - V_1, M_{00}^3 = U_2 + V_2, M_{00}^4 = U_2 - V_2$. Basic algebra gives us

$$\begin{aligned}
E\left[M_{00}^1\right] &= \left[s^{-\frac{1}{2}} + (s+4)^{-\frac{1}{2}} \right] \Sigma^{-\frac{1}{2}} (\mu_0 - \mu_1), \\
E\left[M_{00}^2\right] &= \left[s^{-\frac{1}{2}} - (s+4)^{-\frac{1}{2}} \right] \Sigma^{-\frac{1}{2}} (\mu_0 - \mu_1), \\
E\left[M_{00}^3\right] &= \left[\left(\frac{1}{n_0} + \frac{1}{n_1} \right)^{-\frac{1}{2}} + \left(1 - \frac{3}{4n_0} + \frac{1}{4n_1} \right)^{-\frac{1}{2}} \right] \Sigma^{-\frac{1}{2}} (\mu_0 - \mu_1), \\
E\left[M_{00}^4\right] &= \left[\left(\frac{1}{n_0} + \frac{1}{n_1} \right)^{-\frac{1}{2}} - \left(1 - \frac{3}{4n_0} + \frac{1}{4n_1} \right)^{-\frac{1}{2}} \right] \Sigma^{-\frac{1}{2}} (\mu_0 - \mu_1).
\end{aligned}$$

$$\Sigma_{U_1,V_1} = \rho(C) I_p, \quad \rho(C) = \frac{s_0 - s_1}{\sqrt{s(s+4)}}$$

$$\Sigma_{U_2,V_2} = \sqrt{\frac{n}{4n_0n_1 - 3n_1 + n_0}} I_p = \rho_0 I_p$$

$$\Sigma_{M_{00}^1, M_{00}^3} = Cov(U_1 + V_1, U_2 + V_2) = \Sigma_{U_1,U_2} + \Sigma_{U_1,V_2} + \Sigma_{U_2,V_1} + \Sigma_{V_1,V_2}$$

$$\begin{aligned}
\Sigma_{U_1,U_2} &= E\left[\left((s_0+s_1)^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(\widehat{\mu}_0^C - \widehat{\mu}_1^C) - EU_1\right)\left(\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(\widehat{\mu}_0 - \widehat{\mu}_1) - E_{U_2}\right)\right] \\
&= \sqrt{\frac{n_0 + n_1}{n_0 n_1 (s_0 + s_1)}} I_p
\end{aligned}$$

$$\begin{aligned}
\Sigma_{U_1,V_2} = E\Big[&\left((s_0+s_1)^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(\widehat{\mu}_0^C - \widehat{\mu}_1^C) - EU_1\right) \times \\
&\left(\left(1 - \frac{3}{4n_0} + \frac{1}{4n_1}\right)^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}\left(X_j - \frac{\widehat{\mu}_0 + \widehat{\mu}_1}{2}\right) - E_{V_2}\right)|X_m \in \Pi_0\Big] \\
= &\left(\frac{2C(j) - 1}{2n_0} + \frac{1}{2n_1}\right)\left[(s_0 + s_1)\left(1 - \frac{3}{4n_0} + \frac{1}{4n_1}\right)\right]^{-\frac{1}{2}} I_p \\
= &\frac{n_0 + 2n_1 C(j) - n_1}{\sqrt{n_0 n_1 (4n_0 n_1 - 3n_1 + n_0)s}}
\end{aligned}$$

$$\begin{aligned}
\Sigma_{V_1,U_2} = E\Big[&\left(2(s+4)^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}\left(X_l - \frac{\widehat{\mu}_0^C + \widehat{\mu}_1^C}{2}\right) - E_{V_1}\right) \times \\
&\left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}(\widehat{\mu}_0 - \widehat{\mu}_1)\Big] \\
= &\left(\frac{1}{n_0} + \frac{1}{n_1}\right)\left[(s+4)\left(\frac{1}{n_0} + \frac{1}{n_1}\right)\right]^{-\frac{1}{2}} I_p \\
= &\sqrt{\frac{n_0 + n_1}{n_0 n_1 (s_0 + s_1 + 4)}} I_p
\end{aligned}$$

$$\begin{aligned}\Sigma_{V_1,V_2} = E\Big[&\left(2(s+4)^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}\left(X_i - \frac{\widehat{\mu}_0^C + \widehat{\mu}_1^C}{2}\right) - E_{V_1}\right) \times \\ &\left(\left(1 - \frac{3}{4n_0} + \frac{1}{4n_1}\right)^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \left(X_j - \frac{\widehat{\mu}_0 + \widehat{\mu}_1}{2}\right) - E_{V_2}\right) |X_i, X_j \in \Pi_0\Big] \\ &= \left(4I_{i=j} - \frac{2C(j)+1}{n_0} + \frac{1}{n_1}\right)\left[(s+4)\left(1 - \frac{3}{4n_0} + \frac{1}{4n_1}\right)\right]^{-\frac{1}{2}} I_p \\ &= \frac{4n_0n_1I_{i=j} - 2n_1C(j) - n_1 + n_0}{\sqrt{n_0n_1(4n_0n_1 - 3n_1 + n_0)(s+4)}} I_p\end{aligned}$$

$$\begin{aligned}\Sigma_{M_{00}^1, M_{00}^3} &= \left(\sqrt{\frac{n_0+n_1}{n_0n_1(s_0+s_1)}} + \frac{2n_1C(j) - n_1 + n_0}{\sqrt{n_0n_1s(4n_0n_1 - 3n_1 + n_0)}} + \right. \\ &\left. + \sqrt{\frac{n_0+n_1}{n_0n_1(s+4)}} + \frac{4n_0n_1I_{i=j} - 2n_1C(j) - n_1 + n_0}{\sqrt{n_0n_1(s+4)(4n_0n_1 - 3n_1 + n_0)}}\right) I_p \\ &= \eta_{001} I_p\end{aligned}$$

$$\begin{aligned}\Sigma_{M_{00}^1, M_{00}^4} &= \left(\sqrt{\frac{n_0n_1}{(n_0+n_1)(s_0+s_1)}} - \frac{2n_1C(j) - n_1 + n_0}{\sqrt{n_0n_1s(4n_0n_1 - 3n_1 + n_0)}} + \right. \\ &\left. + \sqrt{\frac{n_0+n_1}{n_0n_1(s+4)}} - \frac{2n_0n_1I_{i=j} - 2n_1C(j) - n_1 + n_0}{\sqrt{n_0n_1(s+4)(4n_0n_1 - 3n_1 + n_0)}}\right) I_p \\ &= \eta_{002} I_p\end{aligned}$$

$$\begin{aligned}\Sigma_{M_{00}^2, M_{00}^3} &= \left(\sqrt{\frac{n_0+n_1}{n_0n_1(s_0+s_1)}} - \frac{2n_1C(j) - n_1 + n_0}{\sqrt{n_0n_1s(4n_0n_1 - 3n_1 + n_0)}} - \right. \\ &\left. - \sqrt{\frac{n_0+n_1}{n_0n_1(s+4)}} + \frac{2n_0n_1I_{i=j} - 2n_1C(j) - n_1 + n_0}{\sqrt{n_0n_1(s+4)(4n_0n_1 - 3n_1 + n_0)}}\right) I_p \\ &= \eta_{003} I_p\end{aligned}$$

$$
\begin{aligned}
\Sigma_{M_{00}^2, M_{00}^4} &= \left(\sqrt{\frac{n_0 + n_1}{n_0 n_1 (s_0 + s_1)}} - \frac{2n_1 C(j) - n_1 + n_0}{\sqrt{n_0 n_1 s (4n_0 n_1 - 3n_1 + n_0)}} - \right. \\
&\quad \left. - \sqrt{\frac{n_0 + n_1}{n_0 n_1 (s+4)}} + \frac{2n_0 n_1 I_{i=j} - 2n_1 C(j) - n_1 + n_0}{\sqrt{n_0 n_1 (s+4)(4n_0 n_1 - 3n_1 + n_0)}} \right) I_p \\
&= \eta_{004} I_p
\end{aligned}
$$

$$
\Sigma_{M_{00}} = \begin{pmatrix}
2(1+\rho(C))I_p & \mathbf{0_{p\times p}} & \eta_{001} & \eta_{002} \\
. & 2(1-\rho(C))I_p & \eta_{003} & \eta_{004} \\
. & . & 2(1+\rho_0)I_p & \mathbf{0_{p\times p}} \\
. & . & . & 2(1-\rho_0)I_p
\end{pmatrix}.
$$

Similarly for (2), (3), and (4):

$$
(2) = P\{(M_{11}^1)^T M_{11}^1 - (M_{11}^2)^M M_{11}^2 > 0, (M_{11}^3)^M M_{11}^3 - (M_{11}^4)^M M_{11}^4 > 0\} = G_{11}(M_{11}(C,i,j))
$$

$$
(3) = P\{(M_{01}^1)^M M_{01}^1 - (M_{01}^2)^M M_{01}^2 < 0, (M_{01}^3)^M M_{01}^3 - (M_{01}^4)^M M_{01}^4 > 0\} = G_{01}(M_{01}(C,i,j))
$$

$$
(4) = P\{(M_{10}^1)^M M_{10}^1 - (M_{10}^2)^M M_{10}^2 > 0, (M_{10}^3)^M M_{10}^3 - (M_{10}^4)^M M_{10}^4 < 0\} = G_{10}(M_{10}(C,j,i))
$$

M_{11}, M_{01}, and M_{10} are specified as in Theorem 16. Theorem 16 follows immediately.

Proof of Theorem 17:

The same technique in the proof of Theorem 14 is applied for Theorem 17. From (3.19), we have

$$\begin{aligned}E[\varepsilon\hat{\varepsilon}_C] = {} & \frac{m_0(C)(1-p)}{m(C)}\underbrace{P\{\psi(X)=1,\psi_C(X_1)=1|X\in\Pi_0\}}_{(1)}+\\ & +\frac{m_1(C)(1-p)}{m(C)}\underbrace{P\{\psi(X)=1,\psi_C(X_{n_0+1})=1|X\in\Pi_0\}}_{(2)}+\\ & +\frac{m_0(C)p}{m(C)}\underbrace{P\{\psi(X)=0,\psi_C(X_1)=1|X\in\Pi_1\}}_{(3)}+\\ & +\frac{m_1(C)p}{m(C)}\underbrace{P\{\psi(X)=0,\psi_C(X_{n_0+1})=1|X\in\Pi_1\}}_{(4)}.\end{aligned} \tag{B.1}$$

Consider (1)

$$\begin{aligned}(1) &= P\{\psi(X)=1,\psi_C(X_l)=1|X,X_l\in\Pi_0\},\quad C(l)=0\\ &= P\{(\hat{\mu}_0-\hat{\mu}_1)^T\Sigma^{-1}\left(X-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)<0,\\ &\qquad (\hat{\mu}_0^C-\hat{\mu}_1^C)^T\Sigma^{-1}\left(X_l-\frac{\hat{\mu}_0^C+\hat{\mu}_1^C}{2}\right)<0|X,X_l\in\Pi_0\}\\ &= P\{U_1^TV_1<0,U_2^TV_2<0|X,X_l\in\Pi_0\}\\ &= P\{(K_{00}^1)^TK_{00}^1-(K_{00}^2)^TK_{00}^2<0,\,(K_{00}^3)^TK_{00}^3-(K_{00}^4)^TK_{00}^4<0\}\\ &= G_{00}(K_{00})\end{aligned}$$

where

$$\begin{aligned}
U_1 &= \left(\frac{1}{n_0}+\frac{1}{n_1}\right)^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}}(\widehat{\mu}_0-\widehat{\mu}_1),\\
V_1 &= \left(1+\frac{1}{4n_0}+\frac{1}{4n_1}\right)^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}}\left(X-\frac{\widehat{\mu}_0+\widehat{\mu}_1}{2}\right),\\
U_2 &= s^{-1/2}\Sigma^{-\frac{1}{2}}(\widehat{\mu}_0^C-\widehat{\mu}_1^C),\\
V_2 &= 2(s+4)^{-1/2}\Sigma^{-\frac{1}{2}}\left(X_l-\frac{\widehat{\mu}_0^C+\widehat{\mu}_1^C}{2}\right), \text{where } C(l)=0.
\end{aligned}$$

$$\begin{aligned}
K_{00}^1 &= U_1+V_1, & K_{00}^2 &= U_1-V_1,\\
K_{00}^3 &= U_2+V_2, & K_{00}^4 &= U_2-V_2.
\end{aligned}$$

$$\begin{aligned}
\Sigma_{K_{00}^1,K_{00}^3} &= Cov\left[U_1,U_2\right]+Cov\left[U_1,V_2\right]+Cov\left[V_1,U_2\right]+Cov\left[V_1,V_2\right],\\
\Sigma_{K_{00}^1,K_{00}^4} &= Cov\left[U_1,U_2\right]-Cov\left[U_1,V_2\right]+Cov\left[V_1,U_2\right]-Cov\left[V_1,V_2\right],\\
\Sigma_{K_{00}^2,K_{00}^3} &= Cov\left[U_1,U_2\right]+Cov\left[U_1,V_2\right]-Cov\left[V_1,U_2\right]-Cov\left[V_1,V_2\right],\\
\Sigma_{K_{00}^2,K_{00}^4} &= Cov\left[U_1,U_2\right]-Cov\left[U_1,V_2\right]-Cov\left[V_1,U_2\right]+Cov\left[V_1,V_2\right].
\end{aligned}$$

$$\begin{aligned}
Cov\left[U_1,U_2\right] &= \left[s^{-1/2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\right]^{-\frac{1}{2}} Cov\left[(\hat{\mu}_0-\hat{\mu}_1)(\hat{\mu}_0^C-\hat{\mu}_1^C)^T\right]\\
&= \left[s^{-1/2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\right]^{\frac{1}{2}} \left(\frac{1}{n_0}+\frac{1}{n_1}\right) I_p\\
&= \sqrt{\frac{n_0+n_1}{n_0 n_1 s}} I_p
\end{aligned}$$

$$
\begin{aligned}
&Cov\,[U_1, V_2] = \\
&= 2\left[(s+4)^{-1/2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\right]^{-\frac{1}{2}} Cov\left[(\hat{\mu}_0-\hat{\mu}_1)\left(X_m-\frac{\hat{\mu}_1^C+\hat{\mu}_0^C}{2}\right)^T |X_m \in \Pi_0\right] \\
&= 2\left[(s+4)^{-1/2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\right]^{-\frac{1}{2}} \left(\frac{1}{2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\right) I_p \\
&= \sqrt{\frac{n_0+n_1}{n_0 n_1 (s+4)}}\, I_p
\end{aligned}
$$

$$
\begin{aligned}
&Cov\,[V_1, U_2] = \\
&= \left[s^{-1/2}\left(1+\frac{1}{4n_0}+\frac{1}{4n_1}\right)\right]^{-\frac{1}{2}} Cov\left[(\hat{\mu}_0^C-\hat{\mu}_1^C)\left(X-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)^T |X \in \Pi_0\right] \\
&= \left[s^{-1/2}\left(1+\frac{1}{4n_0}+\frac{1}{4n_1}\right)\right]^{-\frac{1}{2}} \left(\frac{1}{2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)\right) I_p \\
&= \frac{n_0-n_1}{\sqrt{n_0 n_1 (4n_0 n_1 + n_0 + n_1) s}}\, I_p
\end{aligned}
$$

$$
\begin{aligned}
&Cov\,[V_1, V_2] = \\
&= 2\left[(4+s)^{-1/2}\left(1+\frac{1}{4n_0}+\frac{1}{4n_1}\right)\right]^{-\frac{1}{2}} Cov\left[\left(X_m-\frac{\hat{\mu}_1^C+\hat{\mu}_0^C}{2}\right)\left(X-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)^T\right] \\
&= 2\left[(4+s)^{-1/2}\left(1+\frac{1}{4n_0}+\frac{1}{4n_1}\right)\right]^{-\frac{1}{2}} \left(-\frac{1}{4n_0}+\frac{1}{4n_1}\right) I_p \\
&= \frac{n_0-n_1}{\sqrt{n_0 n_1 (4n_0 n_1 + n_0 + n_1)(s+4)}}\, I_p
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_{K^1_{00},K^3_{00}} &= \left(\sqrt{\frac{n_0+n_1}{n_0n_1s}}+\sqrt{\frac{n_0+n_1}{n_0n_1(s+4)}}+\right.\\
&+ \quad \left.\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)s}}+\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)(s+4)}}\right)I_p\\
&= \left(\frac{1}{\sqrt{s}}+\frac{1}{\sqrt{s+4}}\right)\left(\sqrt{\frac{n_0+n_1}{n_0n_1}}+\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)}}\right)I_p\\
&= \zeta_{001}I_p
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_{K^1_{00},K^4_{00}} &= \left(\sqrt{\frac{n_0+n_1}{n_0n_1s}}-\sqrt{\frac{n_0+n_1}{n_0n_1(s+4)}}+\right.\\
&+ \quad \left.\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)s}}-\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)(s+4)}}\right)I_p\\
&= \left(\frac{1}{\sqrt{s}}-\frac{1}{\sqrt{s+4}}\right)\left(\sqrt{\frac{n_0+n_1}{n_0n_1}}+\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)}}\right)I_p\\
&= \zeta_{002}I_p
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_{K^2_{00},K^3_{00}} &= \left(\sqrt{\frac{n_0+n_1}{n_0n_1s}}+\sqrt{\frac{n_0+n_1}{n_0n_1(s+4)}}+\right.\\
&- \quad \left.\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)s}}-\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)(s+4)}}\right)I_p\\
&= \left(\frac{1}{\sqrt{s}}+\frac{1}{\sqrt{s+4}}\right)\left(\sqrt{\frac{n_0+n_1}{n_0n_1}}-\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)}}\right)I_p\\
&= \zeta_{003}I_p
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_{K_{00}^2,K_{00}^4} &= \left(\sqrt{\frac{n_0+n_1}{n_0n_1s}}-\sqrt{\frac{n_0+n_1}{n_0n_1(s+4)}}+\right.\\
&- \quad \left.\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)s}}+\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)(s+4)}}\right)I_p\\
&= \left(\frac{1}{\sqrt{s}}-\frac{1}{\sqrt{s+4}}\right)\left(\sqrt{\frac{n_0+n_1}{n_0n_1}}-\frac{n_0-n_1}{\sqrt{n_0n_1(4n_0n_1+n_0+n_1)}}\right)I_p\\
&= \zeta_{004}I_p
\end{aligned}
$$

Similarly for (2), (3), and (4).

$$
\begin{aligned}
(2) &= P\{(K_{11}^1)^TK_{11}^1-(K_{11}^2)^TK_{11}^2>0,(K_{11}^3)^TK_{11}^3-(K_{11}^4)^TK_{11}^4>0\}=G_{11}(K_{11}(C)),\\
(3) &= P\{(K_{01}^1)^TK_{01}^1-(K_{01}^2)^TK_{01}^2<0,(K_{01}^3)^TK_{01}^3-(K_{01}^4)^TK_{01}^4>0\}=G_{01}(K_{01}(C)),\\
(4) &= P\{(K_{10}^1)^TK_{10}^1-(K_{10}^2)^TK_{10}^2>0,(K_{10}^3)^TK_{10}^3-(K_{10}^4)^TK_{10}^4<0\}=G_{10}(K_{10}(C)).
\end{aligned}
$$

K_{11}, K_{01}, and K_{10} are 4-dimensional Gaussian random variables as specified in Theorem 17.

VITA

Thang Vu is from Ha Noi, Viet Nam. He is a Graduate of Hanoi-Amsterdam High School, Ha Noi, Viet Nam for gifted students Class of 1998 and a Vietnam Education Foundation Fellow, cohort 2004.

Vu received the B.S degree in electronics and telecommunications from Hanoi University of Technology, Ha Noi, Viet Nam in 2003, the M.S.E degree in electrical engineering: systems from the University of Michigan at Ann Arbor, M.I, in 2006 and the Ph.D. degree in electrical engineering from Texas A&M University, College Station, TX, U.S in 2011.

Vu's current research topic is concerned with statistical pattern recognition and its applications in genomics/proteomics. His research interests include quantitative methods and their applications in biomedical research.

The typist for this dissertation was Thang Vu.