2017

INFERENCE USING BHATTACHARYYA DISTANCE TO MODEL INTERACTION EFFECTS WHEN THE NUMBER OF PREDICTORS FAR EXCEEDS THE SAMPLE SIZE

Sarah A. Janse
University of Kentucky, sarah.janse@uky.edu
Digital Object Identifier: https://doi.org/10.13023/ETD.2017.455

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Sarah A. Janse, Student

Dr. Katherine Thompson, Major Professor

Dr. Constance L. Wood, Director of Graduate Studies

INFERENCE USING BHATTACHARYYA DISTANCE TO MODEL INTERACTION EFFECTS WHEN THE NUMBER OF PREDICTORS FAR EXCEEDS THE SAMPLE SIZE

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Arts and Sciences
at the University of Kentucky

By

Sarah A. Janse

Lexington, Kentucky

Co-Directors: Dr. Katherine Thompson, Professor of Statistics

and Dr. Arnold Stromberg, Professor of Statistics

Lexington, Kentucky

ABSTRACT OF DISSERTATION

INFERENCE USING BHATTACHARYYA DISTANCE TO MODEL INTERACTION EFFECTS WHEN THE NUMBER OF PREDICTORS FAR EXCEEDS THE SAMPLE SIZE

In recent years, statistical analyses, algorithms, and modeling of big data have been constrained due to computational complexity. Further, the added complexity of relationships among response and explanatory variables, such as higher-order interaction effects, make identifying predictors using standard statistical techniques difficult. These difficulties are only exacerbated in the case of small sample sizes in some studies. Recent analyses have targeted the identification of interaction effects in big data, but the development of methods to identify higher-order interaction effects has been limited by computational concerns. One recently studied method is the Feasible Solutions Algorithm (FSA), a fast, flexible method that aims to find a set of statistically optimal models via a stochastic search algorithm. Although FSA has shown promise, its current limits include that the user must choose the number of times to run the algorithm. Here, statistical guidance is provided for this number iterations by deriving a lower bound on the probability of obtaining the statistically optimal model in a number of iterations of FSA. Moreover, logistic regression is severely limited when two predictors can perfectly separate the two outcomes. In the case of small sample sizes, this occurs quite often by chance, especially in the case of a large number of predictors. Bhattacharyya distance is proposed as an alternative method to address this limitation. However, little is known about the theoretical properties or distribution of B-distance. Thus, properties and the distribution of this distance measure are derived here. A hypothesis test and confidence interval are developed and tested on both simulated and real data.

KEYWORDS: Bhattacharyya Distance, model selection, Feasible Solutions Algorithm, perfect separation, interaction effects, logistic regression

SARAH A. JANSE
Student's Signature

OCTOBER 10, 2017
Date

INFERENCE USING BHATTACHARYYA DISTANCE TO MODEL INTERACTION EFFECTS WHEN THE NUMBER OF PREDICTORS FAR EXCEEDS THE SAMPLE SIZE

By

Sarah A. Janse

KATHERINE THOMPSON
Co-Director of Dissertation

ARNOLD STROMBERG
Co-Director of Dissertation

CONSTANCE L. WOOD
Director of Graduate Studies

OCTOBER 10, 2017
Date

To my loving and supportive husband, Zaan.

ACKNOWLEDGEMENTS

I would like to thank all of the people who have helped, guided, and supported me throughout my graduate work. In particular, I would like to thank my committee members, Dr. Arnold Stromberg, Dr. Solomon Harrar, Dr. David Fardo, and Dr. William Griffith for their encouragement and feedback. Also, I cannot thank Dr. Katherine Thompson enough for her countless hours, enthusiastic support, and tireless effort throughout her time as my advisor. I am truly grateful for her encouragement, patience, and advice both within my research and outside of my graduate work. Lastly, I would like to express my gratitude for the unmeasurable amount of love and support I have received from Zaan Janse and Wendell and Susan Witt.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the current world of ever-growing data, there is a great need for statistical methods and algorithms that can handle the analysis of large data sets. As the number of predictors in a data set increases, so does the difficulty in effectively modeling the interactions between these variables. Many approaches exist to identify interaction effects in data sets of small to moderate size. Classical statistical methods suggest considering all pairwise combinations of possible explanatory variables in the proposed logistic regression or linear regression model, and selecting a set of variables based on either hypothesis tests or a model selection criterion. Although the theory supporting these techniques is developed, often the data sets of interest have an inordinate number of possible explanatory variables to consider in higher-order interactions using the conventional implementations of classical methods.

For example, genomic data (that often contains as many as thousands of individuals and millions of locations on the genome, or loci) is also unique in its complexity due to intricate dependencies among genes and traits, often in the form of information from external influences or genetic makeup that are unaccounted for during analysis. For instance, a recent study showed that although genome-wide association study (GWAS) data identified 71 loci associated with Crohn's disease risk, these contributed to only 21.5% of the estimated heritability in the disease. Heritability is the degree of variation in a trait in a population that is due to genetic variation between individuals in that population [Wray and Visscher, 2008]. However, use of a biologically-informed model for data analysis showed that interactions could account for as much as 80% of the missing heritability in Crohn's disease [Zuk et al., 2012]. Thus, a person's susceptibility to disease may depend more on the combined effect of all the genes in the background than on the specific disease genes in the forefront.

Epistasis is a specific relationship between genes where the effect of one gene is dependent on the presence of one or more other genes. In particular, statistical interaction effects among genes contribute to epistasis, which is especially difficult to identify in genomic data [Moore and Williams, 2009]. These interactions have been targets of recent analyses of genomic data, although development of methods to identify higher-order interaction effects has been more limited due to computational concerns [Gemperline, 1999]. The difficulties from the complex nature of interaction effects coupled with the size of genomic data sets cause computational issues when classical methods are applied

using standard implementations. For example, even using the second largest supercomputer at IBM (a 262,144 core machine), Goudey et al. performed analyses examining two-way interactions on a large data set of 1.1 million single nucleotide polymorphisms (SNPs), or locations along the genome, from 2000 samples in less than 10 minutes, and project that analyzing three-way interactions on the same computer would take approximately 5.8 years and on other computers could take more than a thousand years [Goudey et al., 2015]. Thus, exhaustive searches using anything less than a massive supercomputer are highly impractical for this type of data, especially in the case of higher-order interactions.

In contrast, stochastic search algorithms address the computational concern of the exhaustive search methods by employing some aspect of randomness in order to perform a non-exhaustive search over the possible explanatory variables. However, these methods may not produce the same result every time, and thus may fail to identify the truly statistically optimal model according to the criterion used. In addition, most exhaustive and stochastic searches produce a single "best" model or set of explanatory variables with respect to some hypothesis testing or model selection criterion. In any given data set, there may be another, nearly best model that is more biologically meaningful than the statistically best model. Only considering the single statistically optimal solution leaves little room to consider more practically meaningful combinations of variables without further experimentation.

Thus, room for improvement exists in implementing methods that are fast and flexible in their ability to detect both main and interaction effects in a model, and reliable in detecting effects that are not only statistically, but also practically, significant. One recently studied method is the Feasible Solutions Algorithm (FSA), a fast, flexible method that aims to find a set of statistically optimal models via a stochastic search algorithm [Lambert, 2015]. FSA may produce several nearly optimal solutions from different iterations of the algorithm, rather than a single optimal solution. This variability produces multiple results for consideration to glean practically reasonable conclusions from the data, rather than ending with a single (statistically) optimal solution. In the latter case, one solution may be optimal and practically nonsensical, while another nearly optimal solution exists that is biologically relevant. FSA will show the analyst both solutions for consideration during analysis.

Although FSA has shown promise, in its current implementation, the user must choose the number of times to run the algorithm. Each iteration of FSA is referred to as a random start and begins with an arbitrary model of a specific order. The number of random starts must be chosen by the user and to date there has not been any statistical guidance in implementing FSA. Thus, here I derive a

bound on the probability of obtaining the statistically optimal solution in a set number of random starts of FSA that can be used to select this number of iterations of FSA to run. This allows users to choose a bound such that they obtain the statistically optimal solution with a desired probability, prior to beginning data analysis.

The difficulties due to number of predictors described here are only exacerbated in the case of small sample sizes in some studies. For example, a recent analysis of gene expression data in nonhuman primates was conducted in order to examine the association of these biosignatures with periodontal disease. Traditionally, gene expression data have been analyzed using two-sample t-tests on the expression for each gene across groups to find genes whose mean expressions differ across disease/healthy groups. However, small sample sizes, combined with the number of tests adjusted for, make this analysis underpowered. Further, t-tests are designed to detect differences in mean expression between disease and healthy samples rather than treating the disease as the response variable. For example, the hypoxia pathway is expected to be informative due to previously-identified alterations in hypoxia gene expression (i.e., tissue responses to low oxygen). When looking at two genes in the hypoxia pathway, HIF1A and HIF3A, the two-sample t-tests for each gene showed no significant differences in expression across the healthy and diseased individuals. However, this analysis is limited by consideration of single genes, leading to a failure to identify differences in HIF1A and HIF3A expression linked to periodontitis.

The accuracy of this prediction can be improved by using multiple genes simultaneously in disease risk prediction. The standard statistical technique to do so is logistic regression, which predicts presence or absence of the disease using two genes. However, logistic regression is severely limited in the case of small sample sizes, and often even in the case of large sample sizes, if the two genes can perfectly separate the data. In this case, HIF1A and HIF3A can perfectly separate healthy and diseased individuals and logistic regression fails. In the analysis of the genes from three biological pathways including hypoxia, approximately 15% of gene pairs showed perfect discrimination. Any set of predictors in the data that shows linear separation of disease and healthy individuals will be a case when logistic regression is inapplicable. However, these are the cases that are arguably more helpful since the separation, if large, could be an accurate way to demarcate disease/healthy individuals. This occurs especially often by chance when sample sizes are small.

Thus, Bhattacharyya's distance (B-distance) is proposed as an alternative to logistic regression. B-distance addresses the severe limitation of logistic regression in that it can measure the distance between healthy and disease individuals when there is linear separation. In fact, the distance also

increases as the separation becomes larger and decreases as the separation is smaller, while logistic regression fails in either case. Thus, B-distance has been proposed to identify interaction effects among genes in gene expression data. This allows the quantification of differences between healthy/disease groups using multiple genes simultaneously, rather than using single genes. However, little is known about the theoretical properties or distribution of B-distance. These properties will be examined here and used to derive inference methods for B-distance.

This work focused on developing theoretical properties of two different methods for identifying interactions in big data. Chapter 2 includes a review of current methods related to the techniques relevant to this work, as well as their advantages and limitations. In Chapter 3, I will discuss the selection of number of iterations of FSA necessary to achieve the statistically optimal solution with a given probability. This selection is based on a lower bound to this probability, which I prove exists and test through simulations and real data analysis. Chapter 4 is dedicated to the development of properties of the sample B-distance, including the derivation of its distribution under various assumptions. The usefulness of these results is then presented through simulation studies and real data analysis. Chapter 5 focuses on the novelty of these results and the future direction of work associated with these findings.

Chapter 2

Literature Review

Here I discuss current methods that relate to identifying interaction effects in big data. I will start with an overview of general model selection techniques and highlight both advantages and limitations of the current methods. Next, I move on to discuss model selection methods specific to the analysis of genetic data since there are many applications of big data modeling methods in this area. Then I focus on discussing the version of FSA that is currently used to identify interaction effects in big data and describe the limitations that I address later in this work. Other issues of modeling big data occur when logistic regression is used to model the probability of being in one of two groups and one or more predictors can perfectly separate the data. Thus, I will discuss this problem and the limitations of other classical approaches when modeling a binary response. In order to address these concerns, I propose using Bhattacharyya distance and thus discuss this distance measure, as well as other distance measures as background information. Specifically, I am interested in B-distance in the multivariate normal case, which is defined here, but I also present some of the previous uses of this distance measure in other contexts.

2.1 General Model Selection

Common algorithms and methods exist to find the best subset of predictors that adequately explain the response variable. The simplest data driven modeling approaches are stepwise selection methods. Forward selection starts with the model containing the most significant predictor and continues adding the most significant predictors one at a time until no other variables pass some preset threshold of significance [Neter et al., 1996]. Backward selection is a another common method related to forward selection. This method starts with the model including all of the predictors of interest and then removes the least significant variables one at a time as long as they aren't below the preset threshold [Neter et al., 1996]. The method continues in this way refitting reduced models until no other variables can be removed. The advantage of these approaches is that they are simple to implement and most statistical software already have options to use these methods, i.e. the leaps package in R [Lumley and Miller, 2004]. However, with forward selection each addition of a new variable may render one or more of the already included variables non-significant and with backward selection, sometimes variables are dropped that would be significant when added to the final reduced

models. These stepwise selection methods also suffer from high variability in the resulting model and low prediction accuracy, especially when there are many predictors or correlated predictor variables (or both).

Penalized regression techniques are common approaches that have been proposed to address these issues of variability. Penalized regression addresses this instability by decreasing the variance involved in coefficient estimation. Similar to ordinary least squares estimation, penalized regression methods estimate the regression coefficients by minimizing the residual sum of squares. However, penalized regression methods place a constraint on the size of the regression coefficients, also known as a penalty, that causes coefficient estimates to be biased, but that improves the overall prediction error of the model by decreasing the variance of the coefficient estimates. Common penalized regression techniques include LASSO regression, adaptive LASSO regression, elastic net, and ridge regression [Tibshirani, 1996, Zou, 2006, Zou and Hastie, 2005, Hoerl and Kennard, 1970]. Advantages of these methods include stability of estimates, higher prediction accuracy, and computational efficiency, as well as the ability to easily implement these in R [Friedman et al., 2009]. Disadvantages of these methods include that the user must choose the tuning parameter used for penalization and that interpretability of these results can be difficult.

Some other existing analyses only consider effects of single predictors or genes or are computationally infeasible. For example, exhaustive search is a brute-force search method that consists of systematically enumerating all possible combinations of candidate explanatory variables and choosing those predictors that optimize a specified objective function. Advantages of exhaustive search are that it can identify the statistically optimal solution and it is easy to implement. However, one of the largest disadvantages of exhaustive search is the computational complexity required, especially when the number of predictors is large. Another disadvantage is the method's inability to identify other important predictors outside of the statistically optimal solution. There may be one set of predictors that is the best statistical combination, but other pairs of predictors with strong interactions may exist that provide important information about the response. Exhaustive search does not typically provide these answers.

2.2 Model Selection in Genetic Data

As the ability to collect large amounts of genetic data has increased, so has the need for statistical models that can handle the complexities of this type of data. These complexities arise from not

only the size of the data, but also from the intricate dependencies between predictors. In particular, gene-gene and gene-environment interactions contribute to epistasis and these interaction effects are especially difficult to identify in genomic data [Ma et al., 2013]. Thus, several tools and methods have been proposed to analyze these specific types of genomic data sets.

PLINK is an open-source C/C++ whole genome association study (WGAS) tool set with five main domains of function: data management, summary statistics, population stratification, association analysis, and identity-by-descent estimation [Purcell et al., 2007]. The software can implement several types of association tests for genetic data and general linear and logistic regression models that allow for multiple binary or continuous covariates having both main effects and interactions. One can test for joint effects or perform a scan conditional on a given SNP or set of SNPs, for example. Also, gene-gene and gene-environment interaction tests for quantitative and binary disease traits can be performed. These are done through an exhaustive search strategy. Limitations of PLINK are that as mentioned earlier, exhaustive search is computationally intensive and it is not able to handle three-way interactions.

BOOST (BOolean Operation-based Screening and Testing) is a method for the discovery of unknown potential gene-gene interactions that underlie complex diseases [Wan et al., 2010]. BOOST allows examination of all pairwise interactions in genome-wide case-control studies and maintains computational efficiency. It is a computationally and statistically useful tool in the coming era of large-scale interaction mapping in genome-wide case-control studies. It consists of both a screening stage and a testing stage. In the screening stage, a non-iterative method is used to approximate the likelihood ratio statistic in evaluating all pairs of SNPs and those that pass a specified threshold are kept for testing. In the testing stage, the classical likelihood ratio test is employed to measure the interaction effects of selected SNP pairs. It handles covariates in two ways; if the covariate is discrete or can be discretized, the method can be directly extended to handle it. If not, logistic regression can be used in the post-processing step to adjust for the covariate. However, in the current stage this method cannot be applied to GWAS data involving continuous phenotypes unless those continuous phenotypes can be discretized. Other limitations exist with respect to statistical power and its flexibility, i.e., covariate by gene interactions may be missed by adjusting for covariates post-analysis.

GBOOST is a method that extends BOOST through the use of graphic processing units (GPUs), which are highly parallel hardware that provide massive computing resources [Yung et al., 2011]. This method further speeds up the analysis of gene-gene interactions by implementing the BOOST method based on a GPU framework. The computational burden of BOOST lies in the screening stage.

Thus, GBOOST modifies input data structures and parallelizes computations in the screening stage. GBOOST achieves a 40-fold speedup compared with BOOST. However, the statistical limitations of BOOST remain applicable here.

Random Jungle is a random forest method used to deal with the complex dependencies of genetic data [Schwarz et al., 2010]. This freely available software package detects important SNPs by permutation importance measures, a method of scoring based on how shuffling random features in the data affects the the performance of a model. It offers different permutation importance measures and includes options such as the backward elimination method. However, it is aimed at testing for associations between a single SNP and the outcome while allowing for interactions, instead of specifically testing for interactions between SNPs. This method of testing is easier than testing for interactions and has difficulty in finding interacting SNP pairs displaying weak main effects. This is because trees built in Random Jungle rely on the main effects of SNPs. Thus, gene-gene interaction with no marginal effects might be left unrealized when the random forest algorithm is applied.

The methods listed here are a subset of those that have been developed to specifically handle model selection for genetic data. Many of these methods are easy to implement, but have large drawbacks, such as computational complexity, not being able to handle continuous responses, or the inability to test for higher-order interaction effects. Next, I review a general model selection method, FSA, along with its advantages and limitations.

2.3 Feasible Solutions Algorithm (FSA)

Issues from the complex nature of interaction effects, coupled with the size of data, cause theoretical and computational problems when classical methods are applied using standard implementations. To address these limitations, some recent work has been focused on revisiting versions of the Feasible Solutions Algorithm (FSA) first popularized by Doug Hawkins at the University of Minnesota in the early 1990's [Miller, 1984, Hawkins and Olive, 1999]. Several versions of FSA exist [Hawkins, 1993b,a, 1994b,a], but this work focuses on feasible solutions algorithms for specific criteria. These criteria include the least median of squares (LMS) and least trimmed squares (LTS) that are standard high breakdown criteria for linear regression and the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) for the estimation of the location vector and scatter matrix of multivariate data. However, I am focused on the following version of FSA, which is used for subset selection.

Here, the version of FSA under consideration is specifically the version designed to find interactions when the number of predictors is large [Lambert, 2015]. FSA searches the set of all possible interaction effects to identify those that improve the predictive model for a given response. Issues from the complex nature of interaction effects, coupled with the size of big data, cause theoretical and computational problems when classical methods are applied using standard implementations. An advantage of FSA is that by foregoing exhaustive search and not checking every single possible model, computational time is improved. Another advantage of applying FSA in these cases is that it provides more than one feasible solution, or candidate set of explanatory variables, for a particular data analysis. Providing a set of solutions increases the likelihood of finding practically significant associations rather than solely statistically significant associations.

Specifically, FSA is carried out as follows:

1. Randomly choose m variables from the possible p predictors and compute a specified objective function, e.g. R^2.
2. Consider exchanging one of the m selected explanatory variables from the current model with another explanatory variable in the data set.
3. If an improvement exists, make the exchange that improves the objective function the most.
4. Keep making exchanges until the objective function does not improve. The explanatory variables included in the resulting model are called a feasible solution.
5. Repeat steps (1)-(4) for the number of random starts specified to find additional feasible solutions.

These steps are shown in an example in the flowchart in Figure 2.1. In this example, predictors X_1, X_2, ..., X_5 are considered to model some response Y. FSA randomly selects two of these predictors (when considering a two-way interaction model) and calculates R^2 for the associated model. Then, all possible exchanges a of a single variable are considered and R^2 is computed for each model associated with a pair of predictors. The combination of predictors associated with the highest R^2 will be chosen. FSA continues in this manner, considering all possible exchanges that have not been considered previously, until no exchange can improve the objective function. The algorithm stops when this occurs and the resulting combination of predictors is called a feasible solution. Each iteration of FSA follows these steps and provides a single feasible solution.

Consider an example with 5 predictors of interest, X_1, X_2, ..., X_5. For a single random start, when considering $m = 2$ way interactions, FSA will randomly choose 2 of these predictors.

Random Start

X_2,X_4 $\longrightarrow$ $R^2 = 0.26$ Compute R^2

$\Downarrow$

Consider all possible exchanges for each variable, computing R^2 for each of these.

X_2,X_1 $\longrightarrow$ $R^2 = 0.22$
X_2,X_3 $\longrightarrow$ $R^2 = 0.14$
X_2,X_5 $\longrightarrow$ $R^2 = 0.82$
X_1,X_4 $\longrightarrow$ $R^2 = 0.35$
X_3,X_4 $\longrightarrow$ $R^2 = 0.09$
X_5,X_4 $\longrightarrow$ $R^2 = 0.17$

$\Downarrow$

Make the exchange that most improves the objective function.

X_2,X_5

$\Downarrow$

Consider the possible exchanges that have not been considered previously, computing R^2 for each of these.

X_1,X_5 $\longrightarrow$ $R^2 = 0.64$
X_3,X_5 $\longrightarrow$ $R^2 = 0.13$

$\Downarrow$

Stop when the objective function can no longer be improved.
X_2,X_5 is a
Feasible Solution

Figure 2.1: This chart contains a short example to illustrate the steps of FSA for a single iteration. Each iteration starts at a random pair of variables (when $m = 2$) and calculates the specified objective function, i.e., R^2. Then all possible exchanges of a single variable are considered and the respective R^2 calculated. FSA will make the exchange that optimizes the criterion. It will proceed in this way until no swap can improve the objective function. The pair of variables that cannot be improved upon in a single step is called a feasible solution. Each iteration of FSA provides a single feasible solution.

A feasible solution is optimal in that no one exchange of a variable in the model for another outside of the model can improve the selected criterion function. Not only does FSA provide a set of feasible solutions, but it is often more computationally efficient than standard exhaustive approaches due to its stochastic nature. Other advantages of the algorithm include the ability to analyze both linear and logistic regression models, as well as being able to implement several different optimization criteria. However, FSA is limited in that the number of iterations to be performed is user-chosen

and currently there is no statistical guidance to an appropriate way to choose this number. In Chapter 3, I focus on the required number of iterations, or replications, required by FSA to produce the statistically optimal model.

2.4 Logistic Regression

Logistic regression is a regression model that is a special case of the generalized linear model used in the presence of a categorical response variable. Although it can accommodate more than two categories, I will specifically be discussing the case of modeling a binary outcome. This response is usually modeled as having values of "0" or "1" that represent outcomes such as pass/fail, alive/dead, or healthy/sick. Logistic regression was developed by statistician David Cox in 1958 [Cox, 1958] and is used to estimate the probability of a binary response based on one or more predictors. It allows one to say that the presence of a risk factor increases the probability of a given outcome by a certain percentage.

Consider the logistic regression model. The outcome of the regression is not a prediction of a Y value, as in linear regression, but a probability of belonging to one of two conditions of Y, which can take on any value between 0 and 1 rather than just 0 and 1. Because probability is bounded, what will actually be modeled is the logit of the probability given by

$$log(\frac{\pi_j}{1-\pi_j}) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ...\beta_p x_{pj}$$

where π_j indicates the probability of an event for observation j, β_i are the regression coefficients associated with the reference group and the explanatory variables. That is, for observation j,

$$\pi_j = \frac{e^{\boldsymbol{X}_j{}^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_j{}^T\boldsymbol{\beta}}}$$

where $\boldsymbol{X}_j{}^T = (X_{1j}, X_{2j}, ..., X_{pj})$ are the p explanatory variables for individual j.

Advantages of logistic regression include that it is well documented and understood. Interpretation of the logistic regression model is convenient in that it allows for coefficients of predictors to be interpreted as the relationship to the odds of being in one response group over the other. One drawback of logistic regression is that the logistic regression model can only be fit numerically, meaning each logistic regression requires multiple computations. Moreover, logistic regression has

an issue when perfect separation occurs. This is a severe limitation of using logistic regression to consider possible gene sets whose expressions will improve clinicians ability to identify patients with a given disease. Any set of predictors in the data that shows linear separation of disease and healthy individuals will be a scenario when logistic regression is inapplicable. However, these are the combinations of predictors that are arguably more helpful in personalizing treatment since the separation, if large, could be an accurate way to demarcate disease/healthy individuals. This occurs especially often by chance when sample sizes are small. Next, perfect separation and its complications are discussed in detail.

2.5 Perfect Separation

Separation occurs in models of binary outcome data when one or more explanatory variables can perfectly predict the outcome variable. For explanatory variables, complete separation occurs when the variables can perfectly predict both zeros and ones. Quasicomplete separation occurs when predictors perfectly predict either zeros or ones, but not both. Overlap, the ideal case for getting accurate parameter estimates in logistic regression, occurs when there are no such predictors. With overlap, the usual maximum likelihood estimates exist and provide reasonable estimates of parameters. However, under complete or quasicomplete separation, finite maximum likelihood estimates do not exist and the usual method of calculating standard errors fails.

For example, consider Figure 2.2. Notice that the dashed line perfectly separates the individuals in groups 1 and 2. In this case, logistic regression fails. Although perfect separation causes logistic regression to fail, it is not a problem with the data, but instead a problem with the model.

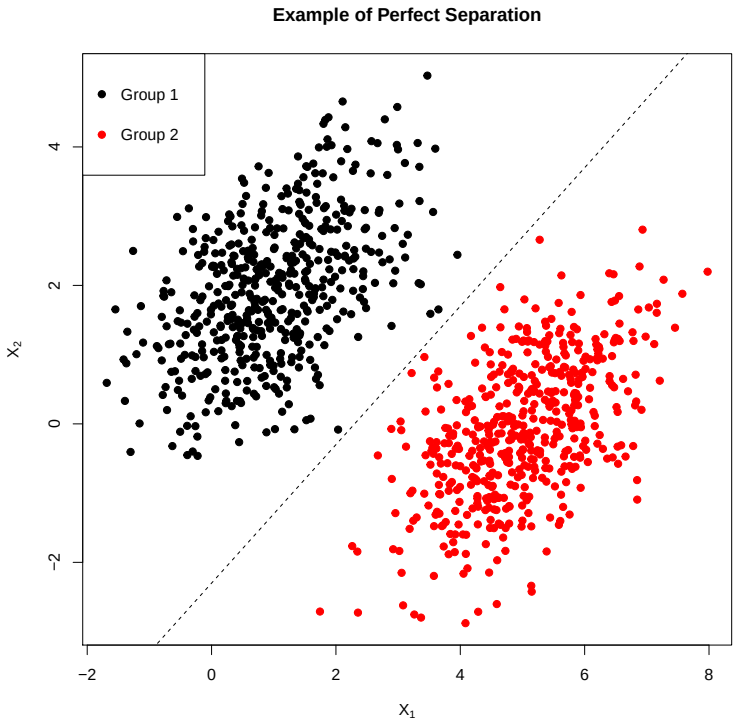


Figure 2.2: This plots shows the relationship between two predictors and the response groups. Black dots denote observations from Group 1 and red dots denote observations from Group 2. In this case, the predictors can perfectly separate observations in group 1 from group 2 (by the dashed line) and logistic regression fails. It is clear, though, that this combination of variables could be useful in determining group membership.

Finding predictors that perfectly separate the data is not necessarily a poor result, since these predictors are arguably the most important in helping to model the outcome. However, perfect separation occurs quite often just by chance in the case of small sample sizes. Thus, an alternative method is needed to model data in the case of small sample sizes where perfect separation causes logistic regression to fail.

One alternative model proposed to allow for perfect separation is called the hidden logistic regression model. This method is a slightly more general model proposed under which the observed response is strongly related but not equal to the unobservable true response. It is given its name because the unobservable true responses are comparable to a hidden layer in a feedforward neural network [Rousseeuw and Christmann, 2003]. However, a limitation of this method is its requirement of user-chosen tuning parameters and accurate estimation of these parameters from the data itself is very difficult, if not impossible, unless the sample size is extremely large.

Another method for addressing the separation problem is proposed by Zorn[Zorn, 2005], but first

suggested by Firth [Firth, 1993] and is based on a penalized likelihood correction to the standard binomial GLM score function [Zorn, 2005]. He applied this method to data from a study on the postwar fate of leaders. However, inference is difficult here and the method entails the use of Jeffreys' prior so caution is urged, particularly in instances where the model in question has a large numbers of nuisance parameters. For instance, others have noted that Jeffreys prior should not be used for conditional logit models, due to the inclusion of a large numbers of nuisance parameters [Poirier, 1994]. Another approach incorporates a form of prior information into the model to stabilize the estimates by introducing the concept of a partial prior distribution to deal with the issues that arise from use of the Jeffreys invariant prior [Rainey, 2016].

2.6 Other Classical Approaches

Often when working with binary outcome data, the classical approach to take is to perform multiple t-tests, testing for differences in mean expression between the two response groups. However, these are done one at a time and are unable to consider interactions. Also, small sample sizes, combined with the number of tests adjusted for, make this analysis very underpowered. Further, t-tests are designed to detect differences in mean expression between response groups rather than treating the group outcome as the response variable, which is usually the goal when working with gene expression data, for example, the data examined in Chapter 4.

Another classical approach for handling a large amount of predictors when trying to model a dichotomous outcome is known as principal components analysis (PCA). PCA combines information from all predictors or genes, but does not use disease status information [Reyes-Aldasoro and Bhalerao, 2006]. The method maps the dimensions of the full data set to a lower dimensional space, with new features that contain the useful information and ignores redundant and irrelevant information. Here, the new features are uncorrelated and are the projections onto axes that maximize the variances of the data. PCA creates new features that are linearly independent, as well as allows the ranking of features according to the size of the global covariance in each principal axis from which a 'subspace' of features can be presented to a classifier. However, while this eigenspace method is effective in many cases, it requires the computation of all the features for a give data set, which is computationally inefficient. Moreover, interpretability is a large issue here as often it is extremely difficult to interpret the new features created, as they are combinations of many features.

2.7 Distance Measures and Divergences

One particular distance, Bhattacharyya's distance, is being proposed here as a method to address perfect separation and to identify interaction effects in big data with the presences of a binary outcome. Thus, it is useful to discuss other distance measures and divergences.

2.7.1 Mahalanobis Distance

Mahalanobis Distance is a measure of the distance between a point and a distribution that was introduced by P. C. Mahalanobis in 1936 [Mahalanobis, 1936]. The Mahalanobis distance of an observation $\vec{x} = (x_1, x_2, x_3, \ldots, x_N)^T$ from a set of observations with mean $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$ and covariance matrix S is defined as

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$

[De Maesschalck et al., 2000].

Mahalanobis distance can also be defined as a dissimilarity measure between two random vectors $\vec{x}$ and $\vec{y}$ of the same distribution with the covariance matrix S

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}.$$

It is typically used to compare a single point to a distribution rather than to compare two distributions by computing the distance between that point and the mean of the distribution. Another limitation here is that the distance formula only incorporates differences in means, but does not consider differences due to covariance. Thus, it would pick up differences in distributions that are centered in different locations, but would most likely miss interaction terms of interest that share the same mean but have different covariances resulting in a "criss-cross" shape of a traditional interaction.

2.7.2 Kullback-Leibler Divergence

The Kullback-Liebler Divergence is another measure of how one probability distribution diverges from a second expected probability distribution [Kullback and Leibler, 1951]. For discrete probability

distributions P and Q, the Kullback-Leibler divergence from Q to P is defined as

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\, \log \frac{P(i)}{Q(i)}$$

[MacKay, 2003]. For continuous probability distributions P and Q, the Kullback-Leibler divergence is defined as

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\, \log \frac{p(x)}{q(x)}\, dx,$$

where p and q denote the densities of P and Q [Bishop, 2006].

Notice that the Kullback-Leibler divergence requires a reference distribution and thus, lacks symmetry in its basic form. This is not a desirable property when considering differences in distributions where one is not obviously the reference.

2.7.3 Hellinger Distance

Hellinger distance, sometimes also referred to as Matusita distance, is another distance measure used to quantify the similarity between two probability distributions [Nikulin, 2001]. The Hellinger distance is defined as

$$H^2(P,Q) = \frac{1}{2} \int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx = 1 - \int \sqrt{f(x)g(x)}\, dx,$$

where f and g are probability density functions of the two probability distributions, P and Q, being compared [Nikulin, 2001]. Previously, Hellinger distance has been used in minimum distance estimation, which is a statistical method for fitting a mathematical model to data, usually the empirical distribution [Beran, 1977]. Hellinger distance is related to B-distance and this relationship is discussed in a later section.

2.8 Bhattacharyya distance

Here, I propose B-distance as an alternative to logistic regression to address the problems that arise when perfect separation causes logistic regression to fail. In doing so, this measure can be used to identify important combinations of predictors in large data sets. Bhattacharyya Distance (B-distance) is a distance measure that measures the similarity of two discrete or two continuous distributions. Bhattacharyya derived a measure of distance between two populations defined in any

way and having the same number of variates in which a one-to-one correspondence can be established [Bhattacharyya, 1943].

For two discrete probability distributions p and q over the same domain X, B-distance is defined as

$$D_B(p,q) = -ln(BC(p,q))$$

where

$$BC(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)}$$

is the Bhattacharyya coefficient.

For continuous probability distributions, the Bhattacharyya coefficient is defined as

$$BC(p,q) = \int \sqrt{p(x)q(x)}dx,$$

where p and q are continuous probability distributions.

The Bhattacharyya coefficient is an approximate measurement of the amount of overlap between two statistical distributions. In either case, $0 \leq BC \leq 1$ and $0 \leq D_B \leq \infty$.

B-distance has previously been used to measure the separability of classes in classification and it is considered to be better than the Mahalanobis distance, as the Mahalanobis distance is a particular case of the Bhattacharyya distance when the standard deviations of the two classes are the same. Consequently, when two classes have similar means but different variances, the Mahalanobis distance would tend to zero, whereas B-distance increases depending on the difference between the variances. B-distance is also preferred to the Kullback-Leibler divergence since it does not require a reference distribution and is therefore symmetric.

There are several properties known about B-distance, including that it is in the class of f-divergences. An f-divergence is a function $D_f(P||Q)$ that measures the difference between two probability distributions P and Q. These divergences were introduced and studied independently by Csiszar [1964], Morimoto [1963] and Ali and Silvey [1966] and are sometimes known as Csiszr f-divergences, Csiszr-Morimoto divergences or Ali-Silvey distances. Some statistical properties of f-divergences include non-negativity, monotonicity, and joint convexity.

2.8.1 Bhattacharyya distance in the Multivariate Normal Case

For the purposes of this work, I will be looking at the distance between two populations that are assumed to follow multivariate normal distributions. For multivariate normal distributions $p_i = \mathcal{N}(\mu_i, \sigma_i)$, B-distance is defined as

$$D_B = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} ln\left(\frac{\det\boldsymbol{\Sigma}}{\sqrt{\det\boldsymbol{\Sigma}_1 \det\boldsymbol{\Sigma}_2}}\right)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the means and covariances of the distributions and

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}$$

The first term here relates to the location of the two distributions, where the second term helps to account for differences in shape or direction of the two populations.

2.8.2 Uses of Bhattacharyya Distance

Previous uses of B-distance include signal selection [Kailath, 1967], feature extraction and selection [Ray, 1989][Guorong et al., 1996] [Choi and Lee, 2003], phone clustering [Mak and Barnard, 1996], pattern recognition [Fukunaga, 2013] [Bishop, 2006], image processing [Goudail et al., 2004], and speaker recognition [You et al., 2009]. A "Bhattacharyya space" has been proposed as a feature selection technique that can be applied to texture segmentation, a process of identifying important features that can help to distinguish between different textures [Reyes-Aldasoro and Bhalerao, 2006].

B-distance has often been used as a class separability measure for feature selection and extraction. Feature extraction is generally known as the process of transforming high dimensional data into a low dimensional feature space based on an optimization criterion. The key to feature extraction is reducing dimensionality without serious loss of class separability. In feature selection, a set of the original measurements is discarded and the most useful ones are kept. Those variables selected constitute the feature space.

Kailath discusses using B-distance in signal selection and compares it to an often used divergence. This divergence is now known as the Kullback-Leibler divergence. The Bayesian probability error is an optimum measure of effectiveness of a set of features selected for the purpose of classification. In communication and radar problems, the optimum signals are those that minimize the probability of

error. The computation or estimation of this error is very difficult and thus various indirect measures of feature effectiveness have been suggested. This minimization of probability error is difficult to carry out and is often impossible. He proposes using B-distance, which is often much easier to evaluate than the divergence and gives results that are at least as good as, and that are often better, than those given by the divergence [Kailath, 1967].

Thus, the Bhattacharrya coefficient and B-distance have become popular feature evaluation criteria. This is because the lower and upper bounds to the Bayesian probability error can be expressed in terms of the coefficient and also because closed-form expressions are available for the coefficient in the case of the exponential family of distributions.

As mentioned previously, B-distance has been used in signal and radar selection. The ability to make decisions about these classifications relies on the distances between stochastic Gaussian processes. Schweppe developed a new expression for the Bhattacharyya distance that expresses the distance in terms of the effects of physically-realizable linear systems (filters) acting on the Gaussian processes [Schweppe, 1967]. The distance is given by time integrals of the mean values and variances of the outputs of filters designed to generate the conditional expectations of certain processes.

In the context of feature evaluation in a two-class pattern recognition problem, Ray shows that irrespective of the values of the a priori probabilities of the two classes, the maximum difference between the existing lower and upper bounds to Bayesian probability error in terms of the B-distance coefficient is approximately equal to 0.2071 [Ray, 1989]. Others have studied using the classification-based B-distance measure to guide biphone clustering [Mak and Barnard, 1996]. In this case, B-distance is used in a data driven approach together with a 2-Level Agglomerative Hierarchical Biphone Clustering algorithm to derive generalized left/right biphones (BGBs), which are subword phonetic units used in speech recognition for classification.

A recursive algorithm called Bhattacharyya distance feature selection has been proposed for selecting a real-optimum feature under the specific case of multivariate normal distribution [Guorong et al., 1996]. This optimization is done by minimizing the upper bound of error probability that was mentioned previously. Another feature extraction method based on B-distance is proposed by Choi and Lee. They also provide an error estimation equation based on this classification error to predict the minimum number of features required for classification without serious information loss [Choi and Lee, 2003]. They also consider multiclass problems by introducing the Bhattacharyya distance feature matrix. Users can then choose the desired M number of features based off eigenvectors

corresponding to the largest eigenvalues of this matrix.

Reyes-Aldasoro and Bhalerao propose a feature selection method based on a Bhattacharyya space. This space is constructed from the B-distances of different variables extracted with sub-band filters from training samples. Then the marginal distributions of the Bhattacharyya space present a sequence of the most discriminant sub-bands that can be used as a path for a wrapper algorithm, such as forward or backward selection [Reyes-Aldasoro and Bhalerao, 2006]. This method was used along with a multiresolution classification algorithm to address a texture segmentation problem. When used on a standard set of texture mosaics, it produced the lowest misclassification errors reported. The authors propose another application of the Bhattacharyya space for detecting which pairs of classes would be particularly difficult to discriminate over all the measurement space. They suggest that in some cases, the individual use of one point of the space can be of interest. The use of this space also implies that the number of classes is previously known and is thus not presented as a method to determine the presence or absence of a number of clusters in a certain space. If this is required, they suggest other methods such as the two-point correlation function or the distance histogram proposed by Fatemi-Ghomi could be used.

Speaker recognition is the process of validating a claimed identity by evaluating the extent to which a test sample matches the claimant's model. In text-independent speaker recognition, both Gaussian mixture models (GMM) and support vector machines (SVM) have been proven to be effective classifiers and most popularly used for many years. A GMM-supervector characterizes a speaker's voice with the parameters of a GMM, which include mean vectors, covariance matrices, and mixture weights. GMM-supervector SVM benefits from both GMM and SVM frameworks to achieve high performance. The conventional Kullback-Leibler kernel in GMM-supervector SVM classifier limits the adaptation of GMM to mean value and leaves covariance unchanged, as has been noted previously. You et. al introduced the GMM-UBM mean interval (GUMI) concept based on using B-distance instead of KL distance, which leads to a novel kernel for SVM classifier [You et al., 2009]. The new kernel allows for exploitation of information not only from the mean, but also from the covariance term. They demonstrated the effectiveness of the new kernel on the 2006 National Institute of Standards and Technology (NIST) speaker recognition evaluation dataset.

Other authors note that Bhattacharyya's concepts have found wide applications in diverse fields, including genetics. Evolutionary geneticists commonly use various distance measures, like Nei's standard genetic distance, Cavalli-Sforza's arc or chord distance and Balakrishnan and Sanghvi's distance, all of which are explicitly or implicitly contained within Bhattacharyya's work

[Chattopadhyay et al., 2004]. It is noted, though, that Bhattacharyya's work, which precede the others by two to five decades, is rarely cited in any of the most prominent and visible works in phylogenetic analysis. Thus, B-distance is a precursor to many genetic distance measures, but has not been given credit as so previously.

In a recent study, a number of measures, including Bhattacharyya, Euclidean, Kullback-Leibler, and Fisher, were studied for image discrimination and it was concluded that B-distance is the most effective texture discrimination for sub-band filtering schemes [Fukunaga, 2013],which are methods for breaking a signal or image into a number of different frequency bands for analysis.

Lastly, B-distance has also been used to identify genes that are differentially expressed between healthy and diseased individuals in a colon cancer experiment. B-distance is used as the gene selection method to identify the small number of differentially expressed genes for a colon cancer analysis. Univariate B-distance was calculated for all genes and the 100 largest distances were chosen as the feature set. These genes were used as input to a fuzzy neural network to classify subjects as having colon cancer or not. Then genes were sequentially eliminated until 7 differentially expressed genes were identified that classified normal and colon cancer with 95.15% accuracy [Tian and Lim, 2013]. My work also considers using B-distance to analyze gene expression data. However, this study differs from the work here, though, in that it uses B-distance to first identify a feature set and then uses these features as input to a classification method.

It is clear from these current uses of B-distance that it can be used as an effective way of performing feature selection and extraction, signal selection, phone clustering, speaker recognition, and pattern recognition. This is mainly due to the fact that it can be used to provide bounds on the probability of error and is easier to compute than other methods. Comparisons of B-distance with other measures show that B-distance is the preferred measure and thus I will consider using it as a way to handle perfect separation and to identify interaction effects in big data. In Chapter 4, I return to B-distance and derive some interesting properties that can be used in hypothesis testing for identifying both main effects and interaction effects.

Chapter 3

FSA

Many approaches exist to identify interaction effects in data sets with small to moderate size. Classical statistical methods suggest considering all pairwise combinations of possible explanatory variables in the proposed logistic regression or linear regression model, and selecting a set of variables based on either hypothesis tests or a model selection criterion. Although the theory supporting these techniques is developed, often the data sets of interest have an inordinate number of possible explanatory variables to consider in higher-order interactions using the conventional implementations of classical methods.

For example, genomic data is also unique in its complexity due to intricate dependencies among genes and traits, often in the form of information from external influences or genetic makeup that are unaccounted for during analysis. In particular, interaction effects among genes contribute to epistasis, which is especially difficult to identify in genomic data [Moore and Williams, 2009]. These interactions have been targets of recent analyses of genomic data although development of methods to identify higher-order interaction effects has been more limited due to computational concerns [Gemperline, 1999]. For example, even using the second largest supercomputer at IBM (a 262,144 core machine), Goudey et al. performed analyses examining two-way interactions on a large data set of 1.1 million single nucleotide polymorphisms (SNPs) from 2000 samples in less than 10 minutes, and project that analyzing three-way interactions on the same computer would take approximately 5.8 years [Goudey et al., 2015]. Thus, exhaustive searches using anything less than a massive supercomputer are highly impractical for this type of data, especially in the case of higher-order interactions.

In contrast, stochastic search algorithms address the computational concern of the exhaustive search methods by employing some aspect of randomness in order to perform a non-exhaustive search over the possible explanatory variables. However, these methods may not produce the same result every time, and thus may fail to identify the truly optimal model according to the criterion used. In addition, most exhaustive and stochastic searches produce a single "best" model or set of explanatory variables with respect to some hypothesis testing or model selection criterion. In any given data set, there may be another, nearly best model that is more practical than the statistically best model. Only considering the single statistically optimal solution leaves little room to consider more

practically meaningful combinations of variables without further experimentation.

Thus, room for improvement exists in implementing methods that are fast and flexible in their ability to detect both main and interaction effects in a model, and reliable in detecting effects that are not only statistically, but also practically, significant. FSA may produce several nearly optimal solutions from different iterations of the algorithm, rather than a single optimal solution. This variability in FSA produces multiple results for consideration to glean practically reasonable conclusions from the data, rather than ending with a single (statistically) optimal solution. In the latter case, one solution may be optimal and practically nonsensical, while another nearly optimal solution exists that is biologically relevant. FSA will show the analyst both solutions for consideration during analysis.

Due to the stochastic nature of FSA, one issue that arises when implementing the algorithm is the choice of number of iterations. Here, each replication is referred to as a random start and begins with an arbitrary model based on the desired m^{th}-order interaction. The number of random starts must be chosen by the user. Thus, I derive a bound on the probability of obtaining the statistically optimal solution in a set number of random starts of FSA that can be used to select this number. This allows users to choose a bound such that they obtain the statistically optimal solution with a desired probability, prior to beginning data analysis.

3.1 Methods

In FSA, each random start begins with an arbitrary model with a fixed number of predictors and proceeds by taking steps to better models based on some optimization criteria, e.g. R^2. The algorithm proceeds until it reaches a statistically optimal model for a given random start. Thus, each random start, or iteration of FSA, will have at least one step, but often times will have several more. Although FSA provides users with a set of potentially interesting feasible solutions, the user may be interested in obtaining the statistically optimal solution. FSA is not guaranteed to identify the statistically optimal solution, but as the analyst increases the number of random starts, FSA is more likely to do so. Thus, some number of random starts is needed to obtain the statistically optimal solution with some probability. However, the more random starts, the longer FSA will take to run. Therefore, it would be highly useful to have information regarding how many random starts to choose in order to obtain the statistically optimal solution with some probability, while still maintaining computational efficiency.

As the number of explanatory variables, p, in a data set increases, it is more difficult to identify the statistically optimal solution and will require more random starts. I propose choosing the number of random starts as a function of p. As p goes to infinity, the probability that the statistically optimal solution is identified by FSA is bounded below. The limit described in Theorem 1 holds for FSA in the case of considering m-way interactions.

Theorem 1: In the case of using FSA to find a statistically significant m-way interaction in a predictive model, as the number of potential explanatory variables, p, goes to infinity, a lower bound on the probability of identifying the statistically optimal model in cp random starts, where $0 < c < 1$, is $1 - e^{-cm^2}$.

Lemma:

$$\lim_{x \to \infty} \left[1 + \frac{k}{x}\right]^{tx} = e^{tk}$$

Proof of Theorem 1: Let p be the number of possible explanatory variables that are chosen from, c be a constant such that $0 < c < 1$, and cp be the number of random starts. Then there are $\binom{p-m}{m}$ pairs of variables out of the total $\binom{p}{m}$ possible pairs that do not contain any of the variables in the optimal solution, consisting of m variables. Note that, if you randomly start with $m-1$ out of the m variables in the statistically optimal solution, you are guaranteed to obtain the optimal solution. Then, the probability of not identifying the optimal solution in the first step of a given random start is

$$\frac{\binom{p-m}{m}}{\binom{p}{m}} \tag{3.1}$$

and so the probability of obtaining the optimal solution in the first step of a given random start is

$$1 - \frac{\binom{p-m}{m}}{\binom{p}{m}}. \tag{3.2}$$

For a given random start, FSA completes at least one step, and often more than one step, before reaching a feasible solution. Since I am only considering finding the statistically optimal solution after the first step and not considering the cases where the optimal solution could be found in later steps, equation (3.2) will be a lower bound on the probability of identifying the statistically optimal solution in a single random start. So, the probability of obtaining the statistically optimal solution in at least one of the cp random starts is greater than $1 - \left[\frac{\binom{p-m}{m}}{\binom{p}{m}}\right]^{cp}$, where $\left[\frac{\binom{p-m}{m}}{\binom{p}{m}}\right]^{cp}$ is the probability that none of the random starts identify the optimal solution in the first step of FSA. So consider

$$\lim_{p\to\infty}\left(1-\left[\frac{\binom{p-m}{m}}{\binom{p}{m}}\right]^{cp}\right)$$
$$=\lim_{p\to\infty}1-\lim_{p\to\infty}\left[\frac{\binom{p-m}{m}}{\binom{p}{m}}\right]^{cp}$$
$$=1-\lim_{p\to\infty}\left[\frac{(p-m)!}{m!(p-2m)!}\frac{m!(p-m)!}{p!}\right]^{cp}$$
$$=1-\lim_{p\to\infty}\left[\frac{(p-m)!(p-m)!}{p!(p-2m)!}\right]^{cp}$$
$$=1-\lim_{p\to\infty}\left[\frac{(p-m)!}{(p-2m)!p(p-1)\cdots(p-m+1)}\right]^{cp}$$
$$=1-\lim_{p\to\infty}\left[\frac{(p-m)(p-m-1)\cdots(p-2m+1)}{p(p-1)\cdots(p-m+1)}\right]^{cp}$$

Notice that both the numerator and denominator in the limit statement contain m quantities. The last line can be written as

$$=1-\lim_{p\to\infty}\left[\frac{p-m}{p}\right]^{cp}\left[\frac{p-m-1}{p-1}\right]^{cp}\cdots\left[\frac{p-2m+1}{p-m+1}\right]^{cp}$$
$$=1-\lim_{p\to\infty}\left[\frac{p-m}{p}\right]^{cp}\lim_{p\to\infty}\left[\frac{p-m-1}{p-1}\right]^{cp}\cdots\lim_{p\to\infty}\left[\frac{p-2m+1}{p-m+1}\right]^{cp}$$
$$=1-\lim_{p\to\infty}\left[1-\frac{m}{p}\right]^{cp}\lim_{p\to\infty}\left[1-\frac{m}{p-1}\right]^{cp}\cdots\lim_{p\to\infty}\left[1-\frac{m}{p-m+1}\right]^{cp}$$

Then

$$\lim_{p\to\infty}\left[1-\frac{m}{p}\right]^{cp}=e^{-cm}$$

by the lemma with $t=c$ and $k=-m$. Next,

$$\lim_{p\to\infty}\left[1-\frac{m}{p-1}\right]^{cp}$$
$$=\lim_{p\to\infty}\left[1-\frac{m}{p-1}\right]^{c(p-1)}\left[1-\frac{m}{p-1}\right]^{c}$$

Since $\lim_{p\to\infty}\left[1-\frac{m}{p-1}\right]^{c(p-1)}=e^{-cm}$ by the lemma with $t=c$ and $k=-m$ and $\lim_{p\to\infty}\left[1-\right.$

$\frac{m}{p-1}\right]^c = 1$,

$$\lim_{p\to\infty}\left[1-\frac{m}{p-1}\right]^{cp} = e^{-cm}.$$

Next,

$$\lim_{p\to\infty}\left[1-\frac{m}{p-m+1}\right]^{cp}$$
$$= \lim_{p\to\infty}\left[1-\frac{m}{p-m+1}\right]^{c(p-m+1)}\left[1-\frac{m}{p-m+1}\right]^{c(m-1)}.$$

Since $\lim_{p\to\infty}\left[1-\frac{m}{p-m+1}\right]^{c(p-m+1)} = e^{-cm}$ by the lemma with $t=c$ and $k=-m$ and $\lim_{p\to\infty}\left[1-\frac{m}{p-m+1}\right]^{c(m-1)} = 1$,

$$\lim_{p\to\infty}\left[1-\frac{m}{p-m+1}\right]^{cp} = e^{-cm}.$$

So,

$$1-\lim_{p\to\infty}\left[\frac{\binom{p-m}{m}}{\binom{p}{m}}\right]^{cp} = 1-e^{-cm}\times e^{-cm}\times\cdots e^{-cm}(m \text{ times})$$
$$= 1-e^{-c(m^2)}.$$

Thus,

$$\text{P(Obtaining the statistically optimal model in } cp \text{ random starts using FSA)} \geq 1-\left[\frac{\binom{p-m}{m}}{\binom{p}{m}}\right]^{cp},$$

and note that:

$$\lim_{p\to\infty}\left(1-\left[\frac{\binom{p-m}{m}}{\binom{p}{m}}\right]^{cp}\right) = 1-e^{-cm^2}.$$

Figures 3.1 and 3.2 show how the calculated probability of obtaining the optimal solution approaches the lower bound derived above for 5 values of c with $m=2$ and $m=3$, respectively. It can be seen that the lower bound is attained very quickly and thus is appropriate when considering data sets

with a large number of explanatory variables, p. It is also clear that the probability of obtaining the statistically optimal solution increases as the number of starts increases, as is expected.

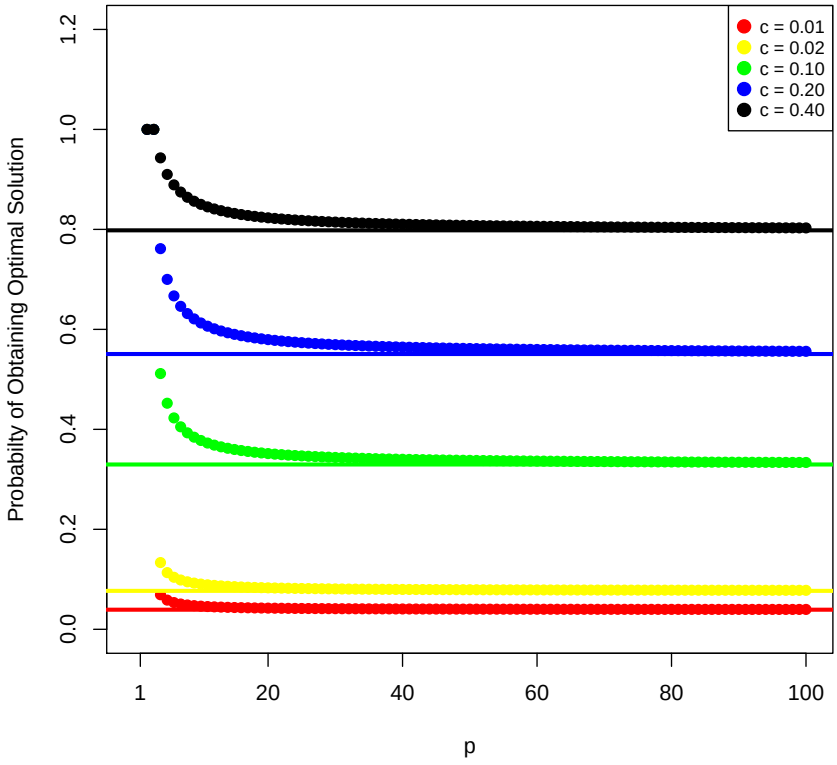


Figure 3.1: In this plot, the dots show the exact value of the lower bound for varied values of c, and the lines show the asymptotic lower bound on the probability of getting the statistically optimal solution with $m = 2$. The lower bound is attained very quickly and the probability of identifying the statistically optimal solution increases as the number of random starts increases, as expected.

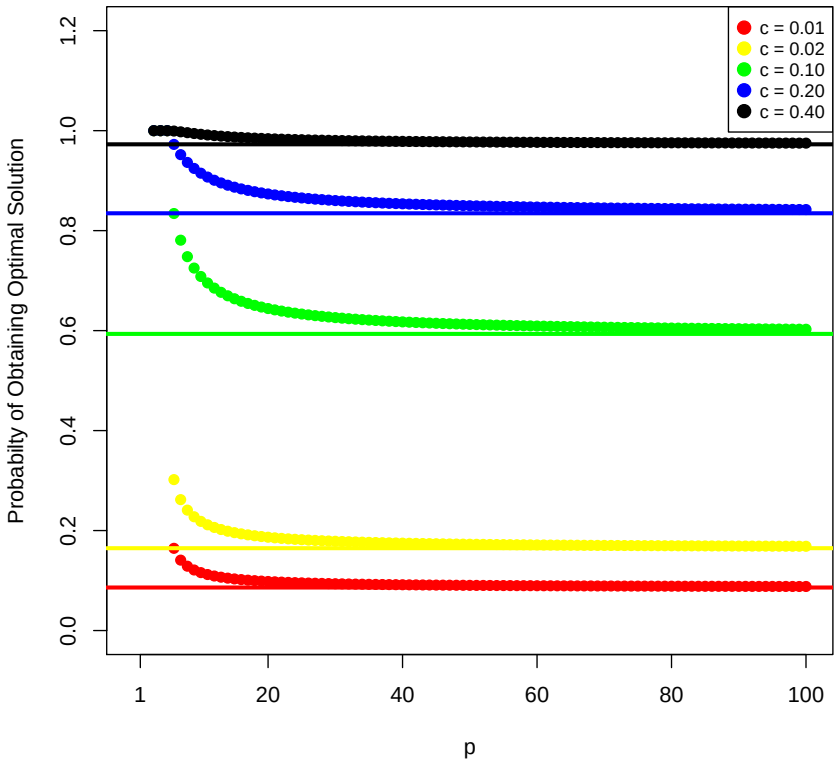


Figure 3.2: In this plot, the dots show the exact value of the lower bound for varied values of c, and the lines show the asymptotic lower bound on the probability of getting the statistically optimal solution with $m = 3$. The lower bound is attained very quickly and the probability of identifying the statistically optimal solution increases as the number of random starts increases, as expected.

Next, simulation studies are performed for both quantitative and binary response variables to examine the outcomes of utilizing the lower bound derived above. These simulations are followed by a real data analysis to demonstrate the use of the lower bound in practice.

3.2 FSA Simulations

Quantitative trait data are simulated as the sum of two covariates and their interaction under the typical regression model for values of p of 50, 100, 1000, and 2500. One hundred data sets are simulated for each value of p. Binary trait data are simulated in an analogous manner. Simulation parameters are as follows:

- Quantitative response variable
  - $X_{ij} \sim U(0,1)$ for $i = 1, ..., n$ and $j = 1, .., p$
  - $Y_i = 5 + X_{i1} + X_{i2} + 2X_{i1}X_{i2} + \epsilon_i$ where $\epsilon_i \sim N(0,1)$
- Binary response variable
  - $X_{ij} \sim U(0,1)$ for $i = 1, ..., n$ and $j = 1, ..., p$
  - $\pi_i = \frac{e^{X_{i1}+X_{i2}+2X_{i1}X_{i2}}}{(1+e^{X_{i1}+X_{i2}+2X_{i1}X_{i2}})}$
  - $Y_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1 - \pi_i \end{cases}$

FSA is used to provide a set of feasible solutions for every simulated data set via the implementation in [Lambert, 2015]. Exhaustive search is then performed to find the statistically optimal solution using R^2 and AIC as the criteria functions for the quantitative and binary response variables, respectively. The numbers of random starts chosen for FSA are values of c including 0.01, 0.02, 0.1, 0.2, and 0.4 with each value of p. Then, for each simulation setting, the percentage of simulated data sets producing the statistically optimal solution using FSA is calculated. These percentages, along with the lower bound from Section 3, are plotted in Figures 3.3 and 3.4.

3.3 FSA Simulation Results

Figures 3.3 and 3.4 show results from 100 simulated data sets for both linear and logistic regression methods in FSA for four values of p and five values of c. Note that the asymptotic lower bound proposed here depends only on m and c. The red dots represent these lower bounds for each value of c. The yellow diamonds, green squares, blue dots, and black triangles represent the percentage of 100 simulations with $p = 50$, 100, 1000, and 2500 respectively, when $m = 2$ (where FSA was able to identify the statistically optimal solution). It is clear from these figures that the lower bound

is often much lower than the observed probability and is thus very conservative, but does provide good guidance as to the number of random starts needed to produce at least one feasible solution containing the statistically optimal solution.

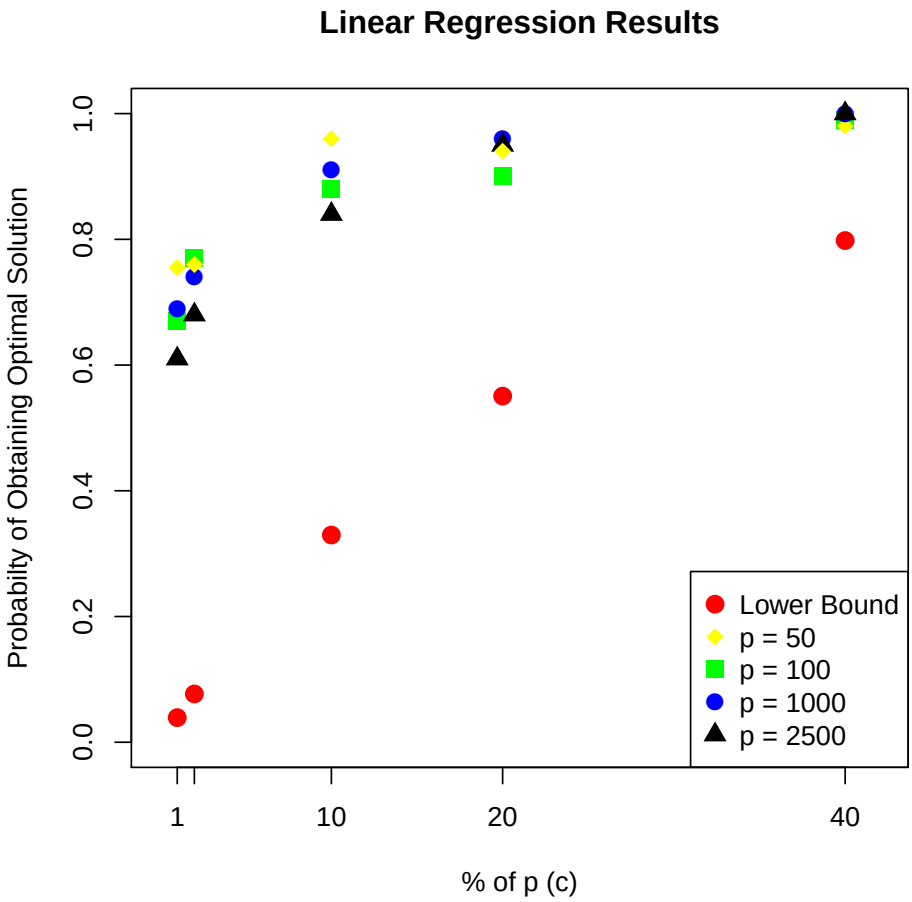


Figure 3.3: Simulation results for the probability of getting the optimal solution in 100 simulations with a quantitative response variable: For each value of c, the lower bounds are represented by the red dots and the probability of identifying the statistically optimal solution for the four values of p are represented by yellow diamonds ($p = 50$), green squares ($p = 100$), blue dots ($p = 1000$), and black triangles ($p = 2500$). This plot shows that the lower bound is valid for all values of p in the simulation study.

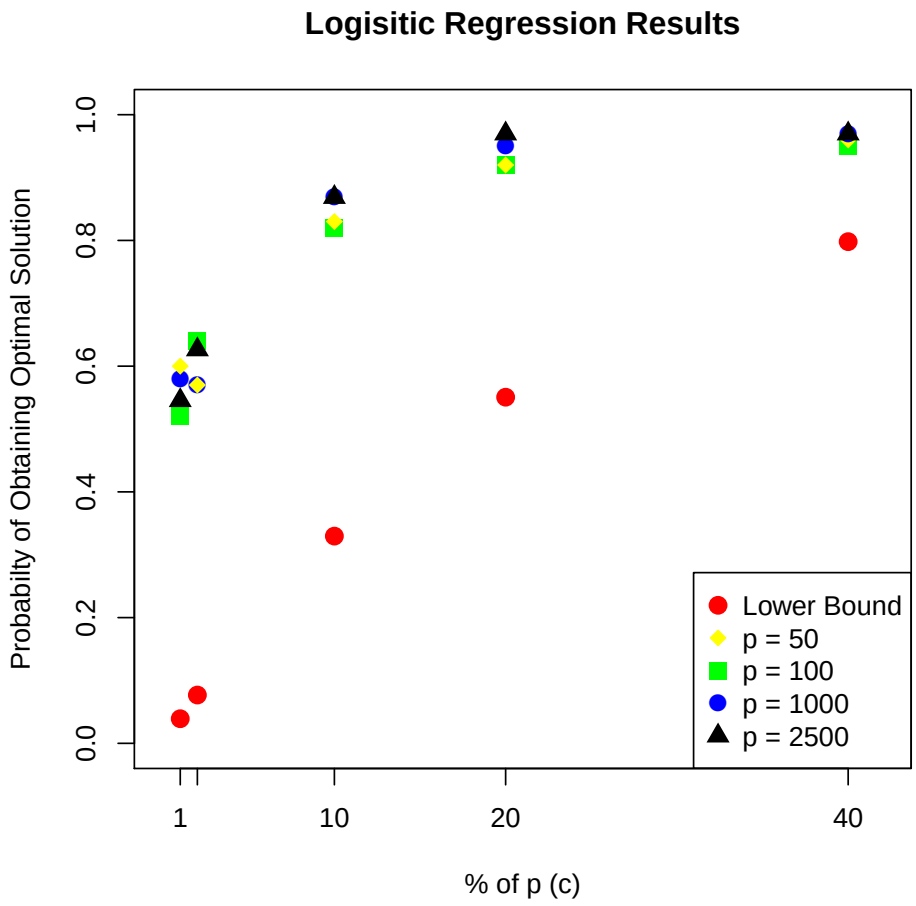


Figure 3.4: Simulation results for the probability of getting the optimal solution in 100 simulations with a binary response variable: For each value of c, the lower bounds are represented by the red dots and the probability of identifying the statistically optimal solution for the four values of p are represented by yellow diamonds ($p = 50$), green squares ($p = 100$), blue dots ($p = 1000$), and black triangles ($p = 2500$). This plot also shows that the lower bound is valid for all values of p in the simulation study.

3.4 FSA Real Data Example

Data for this analysis were collected in a genome-wide association study using 288 outbred mice in a study that aimed to identify, or map, locations along the genome called SNPs that influence HDL cholesterol, systolic blood pressure, triglyceride levels, glucose, or urinary albumin-to-creatinine ratios [Zhang et al., 2012]. The goal is to determine if SNPs or interactions of SNPs are associated with HDL levels. Information from 3045 SNPs on chromosome 11 are analyzed in this real data analysis.

Using the lower bound in Theorem 1, if the desired probability of obtaining the statistically optimal

solution including a two-way interaction is to be at least 95%, then the following equation needs to be solved for c:

$$1 - e^{-c2^2} = 0.95$$

$$\iff c = 0.7489331$$

Since the number of possible predictors is $p = 3045$, the number of random starts needed is $0.75(3045) = 2283.75$, or 2284 random starts.

Table 3.1: The exhaustive search produced the single single statistically optimal solution with $R^2 = 0.1256308$ (Column 3). Columns 1 and 2 show the SNPs that are identified in this model.

Variable 1	Variable 2	R^2
mb2863979	mb87344525	0.1256308

Table 3.2: FSA produced 33 feasible solutions and a subset of those are shown here, including the statistically optimal solution denoted in bold with $R^2 = 0.1256308$ (Column 3). Columns 1 and 2 show the SNPs that are identified in each of the models.

Variable 1	Variable 2	Times Chosen by FSA	R^2
mb104327194	mb91638370	42	0.0957401
mb13136127	mb31255782	898	0.1245719
mb28636979	**mb87344525**	**107**	**0.1256308**
mb111935889	mb43233761	25	0.1065257
mb62443411	mb99541026	23	0.1088855
mb112250554	mb96331482	56	0.1123864

The exhaustive search of the 3045 SNPs on chromosome 11 took approximately 11 hours in total on a large cluster without parallelization. Exhaustive search identified the statistically optimal solution that includes an interaction between mb2863979 and mb87344525 and corresponds to a value of $R^2 = 0.1256$. This result can be found in 3.1. FSA took approximately half an hour to perform 2284 random starts when parallelized on a larger cluster, using 16 cores. There are a total of 33 feasible solutions identified through FSA, including the statistically optimal solution, which is presented in bold in the subset of FSA results in 3.2. One of the SNPs identified here located at mb87344525 is located very close to a location identified in a previous study [Su et al., 2009]. A full table of the FSA results can be found in the Appendix. Out of the 2284 replications of FSA, the statistically optimal solution was identified in 107 of the replications. This illustrates that the number of random starts used here was sufficient to identify the statistically optimal solution. Thus, this work provides

a way for analysts to determine the number of iterations of FSA required to obtain the statistically optimal solution with a specified probability.

Chapter 4

Bhattacharyya Distance

4.1 Methods

I propose using B-distance as an alternative to logistic regression to measure the distance between the two response groups. It addresses the severe limitation of logistic regression in that it still measures the overall distance between individuals in the two response groups when there is linear separation. In fact, the distance also increases as the separation becomes larger and decreases as the separation is smaller, while logistic regression fails in either case.

Advantages of B-distance include that it is much faster than traditional logistic regression methods and it is better at classification than other distance measures, such as Mahalanobis distance or Kullback-Leibler distance. This is due to the fact that it incorporates covariances, as well as means, and that it does not require a reference distribution. Although some properties of B-distance have been studied previously, understanding of its estimator is limited. The distribution of the sample B-distance has yet to be studied. In this section, I will consider this distribution under strict assumptions about the covariances of the two groups and then will also attempt to relax some of those conditions in order to examine the distribution. I will also look at using this information to perform hypothesis testing and to create confidence intervals for the sample B-distance, $\hat{B}$.

4.1.1 The Sample Bhattacharyya Distance

Recall that for multivariate normal distributions $p_i = \mathcal{N}(\mu_i, \sigma_i)$, B-distance is defined as

$$D_B = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\ln\left(\frac{\det\boldsymbol{\Sigma}}{\sqrt{\det\boldsymbol{\Sigma}_1 \det\boldsymbol{\Sigma}_2}}\right)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the means and covariances of the distributions of the two groups and

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}.$$

It is clear that B-distance can be viewed as having two terms. The first term mostly incorporates differences in groups due to means, while the second term incorporates differences in the groups due

to covariances. Now, define the sample B-distance as

$$\hat{B} = \frac{1}{8}(\bar{\boldsymbol{X}}_{\mathbf{1}} - \bar{\boldsymbol{X}}_{\mathbf{2}})^T \hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}_{\mathbf{1}} - \bar{\boldsymbol{X}}_{\mathbf{2}}) + \frac{1}{2}\ln\left(\frac{\det\hat{\boldsymbol{\Sigma}}}{\sqrt{\det\hat{\boldsymbol{\Sigma}}_1 \det\hat{\boldsymbol{\Sigma}}_2}}\right)$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are replaced with their maximum likelihood estimates of $\bar{\boldsymbol{X}}_{\mathbf{1}}$ and $\bar{\boldsymbol{X}}_{\mathbf{2}}$, respectively and $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{\mathbf{1}}$ and $\boldsymbol{\Sigma}_2$ with the sample covariances $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}$, and $\hat{\boldsymbol{\Sigma}}_{\mathbf{2}}$, respectively.

4.1.2 Distribution of the Sample Bhattacharyya Distance Assuming Known Equal Covariances

B-distance can be viewed as having two terms. The first gives information about the differences in means or locations of the two groups and the second term provides information about the differences in variance-covariance or directions of the two groups. First the distribution of $\hat{B}$ is considered under some strict assumptions. Assume $\boldsymbol{\Sigma}_{\mathbf{1}} = \boldsymbol{\Sigma}_{\mathbf{2}} = \boldsymbol{\Sigma}$ and that this quantity is known. Recall that for multivariate normal distributions $p_i = \mathcal{N}(\mu_i, \sigma_i)$, B-distance is defined as

$$B \equiv D_B = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\ln\left(\frac{\det\boldsymbol{\Sigma}}{\sqrt{\det\boldsymbol{\Sigma}_1 \det\boldsymbol{\Sigma}_2}}\right)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the means and covariances of the distributions.

Assuming that $\boldsymbol{\Sigma}_{\mathbf{1}} = \boldsymbol{\Sigma}_{\mathbf{2}} = \boldsymbol{\Sigma}$, then the second term of the distance is equal to zero, leaving

$$B = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

So, under these assumptions,

$$\hat{B} = \frac{1}{8}(\bar{\boldsymbol{X}}_{\mathbf{1}} - \bar{\boldsymbol{X}}_{\mathbf{2}})^T \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{X}}_{\mathbf{1}} - \bar{\boldsymbol{X}}_{\mathbf{2}}).$$

Since the maximum likelihood estimates are being used and it is assumed that $\boldsymbol{\Sigma}_{\mathbf{1}} = \boldsymbol{\Sigma}_{\mathbf{2}} = \boldsymbol{\Sigma}$,

$$\bar{\boldsymbol{X}}_{\boldsymbol{i}} \sim \mathcal{N}_p(\boldsymbol{\mu}_{\boldsymbol{i}}, \frac{1}{n_i}\boldsymbol{\Sigma})$$

for $i = 1, 2$. Letting $\mathbf{X}^* = \bar{\boldsymbol{X}}_{\mathbf{1}} - \bar{\boldsymbol{X}}_{\mathbf{2}}$,

$$\mathbf{X}^* \sim \mathcal{N}_p\left(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*\right).$$

where $\boldsymbol{\mu}^* = \boldsymbol{\mu_1} - \boldsymbol{\mu_2}$ and $\boldsymbol{\Sigma}^* = \frac{n_1+n_2}{n_1 n_2}\boldsymbol{\Sigma}$.

Note the following theorem in Ravishanker and Dey [2001].

Theorem: *If* $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\boldsymbol{\Sigma}$ *is positive definite, then*

$$\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} \sim {\chi_p}^2(\lambda = \mu^T\boldsymbol{\Sigma}^{-1}\mu)$$

Using this theorem,

$$\mathbf{X}^{*T}\boldsymbol{\Sigma}^{-1}\mathbf{X}^* \sim {\chi_2}^2(\lambda = \boldsymbol{\mu}^{*T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^*).$$

Thus,

$$\frac{8n_1n_2}{n_1+n_2}\hat{B} \sim {\chi_2}^2(\lambda)$$

where

$$\lambda = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}).$$

Notice that the noncentrality parameter increases as the differences in means of the two groups increases. It would be useful to understand how quickly the distribution of this quantity converges to the χ^2 distribution. Thus, simulations are conducted to examine the asymptotic convergence of $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$ to this distribution as the sample sizes of the two groups, n_1 and n_2, increase. This provides an idea about what sample sizes are needed to use this distribution to perform hypothesis testing.

Simulations.— Simulations are conducted to examine the asymptotic convergence of $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$ to a ${\chi_2}^2(\lambda)$ distribution with $\lambda = \boldsymbol{\mu}^{*T}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*$) as the sample sizes of the two groups, n_1 and n_2, increase. Two different mean settings and two different covariances are considered to provide a combination of four different simulation parameter settings. Thus, the distribution can be examined when means are far apart, as well as close together, for covariances that include a nonzero term between the two predictors, as well as zero covariance between the two predictors. This is important so that the distribution can be considered under various conditions. For each combination of mean and covariance choices, sample sizes from $n = 5$ per group up to $n = 100$ per group are considered. These simulation settings can be seen in Table 4.1.

Table 4.1: Simulation parameters for comparing the distribution of $\hat{B}$ and the ${\chi_2}^2(\lambda)$ distribution are shown here. Column 1 denotes the number of the simulation setting. Columns 2 and 3 contain the means for the two distributions, while Column 4 displays the covariance of the two distributions. The different sample sizes that are used in the simulations are found in Column 5.

Simulation	$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\Sigma_1 = \Sigma_2 = \Sigma}$	n
1	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5 10 25 50 100
2	$\begin{pmatrix}\frac{1}{4}\\\frac{1}{4}\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5 10 25 50 100
3	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5 10 25 50 100
4	$\begin{pmatrix}\frac{1}{4}\\\frac{1}{4}\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5 10 25 50 100

Ten thousand data sets are simulated according to each of these parameter settings. For each data set, $\frac{8n_1n_2}{(n_1+n_2)}f(\hat{B})$ is calcualted and then plotted in a histogram. Then the distribution of $\frac{8n_1n_2}{(n_1+n_2)}f(\hat{B})$ from these simulations can be compared to the target distribution of ${\chi_2}^2(\lambda = {\boldsymbol{\mu}^*}^T{\boldsymbol{\Sigma}^*}^{-1}\boldsymbol{\mu}^*)$.

Simulation Results.— Simulations are conducted to examine the asymptotic convergence of $\frac{8n_1n_2}{n_1+n_2}\hat{B}$ to a ${\chi_2}^2(\lambda)$ distribution, where $\lambda = {\boldsymbol{\mu}^*}^T{\boldsymbol{\Sigma}^*}^{-1}\boldsymbol{\mu}^*$, as the sample sizes of the two groups, n_1 and n_2, increase. Figures 4.1 - 4.4 show results from 10000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group. In each figure, the red line represents the distribution of the target distribution and the black line represents the kernel density estimate of the empirical distribution of the simulated quantity $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$.

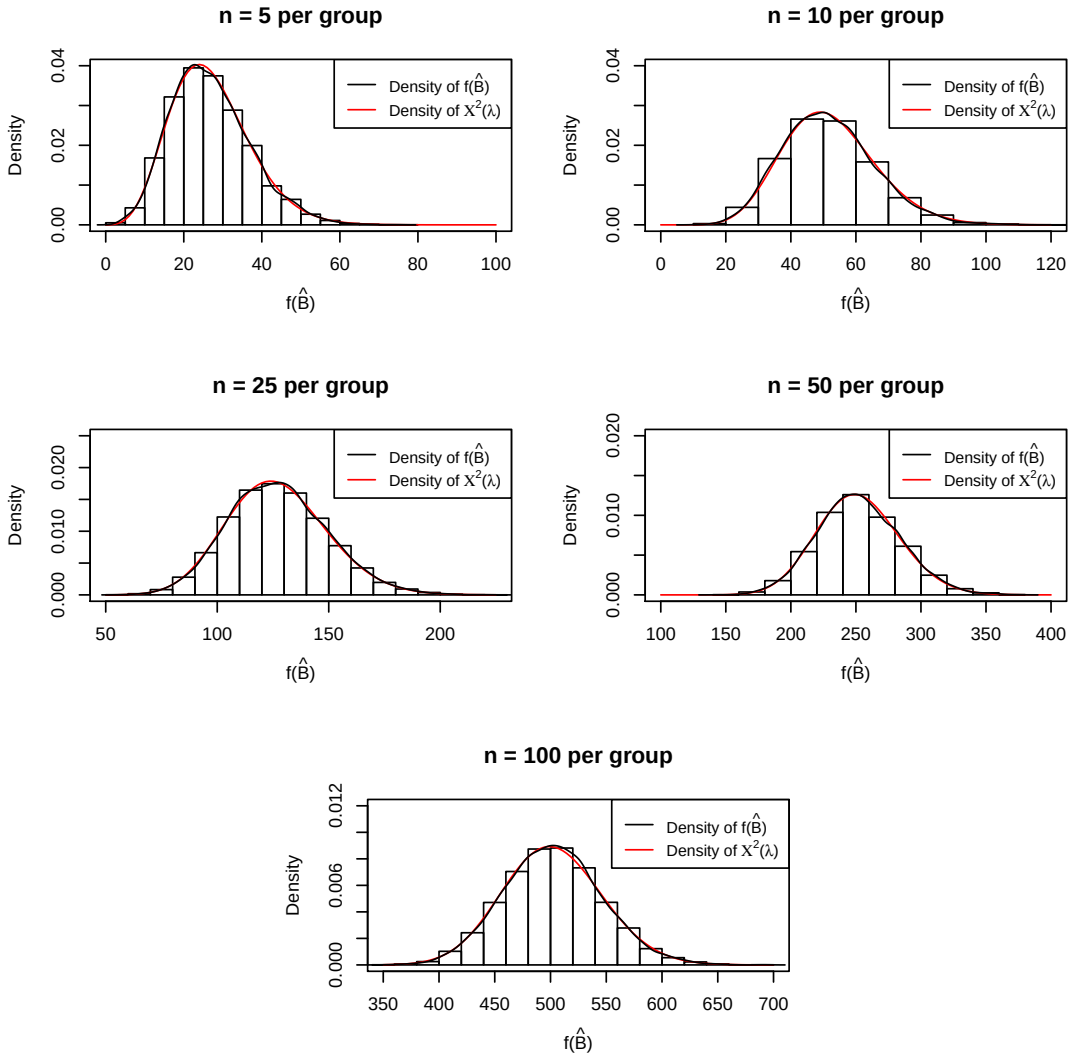


Figure 4.1: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from distributions with means far apart and the same covariance matrix with a non-zero term for covariance between the two predictors (simulation setting 1). In each figure, the red line represents the distribution of the target distribution of ${\chi_2}^2(\lambda = \boldsymbol{\mu}^{*T}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*)$ and the black line represents the kernel density estimate of the empirical distribution of the simulated quantity $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$.

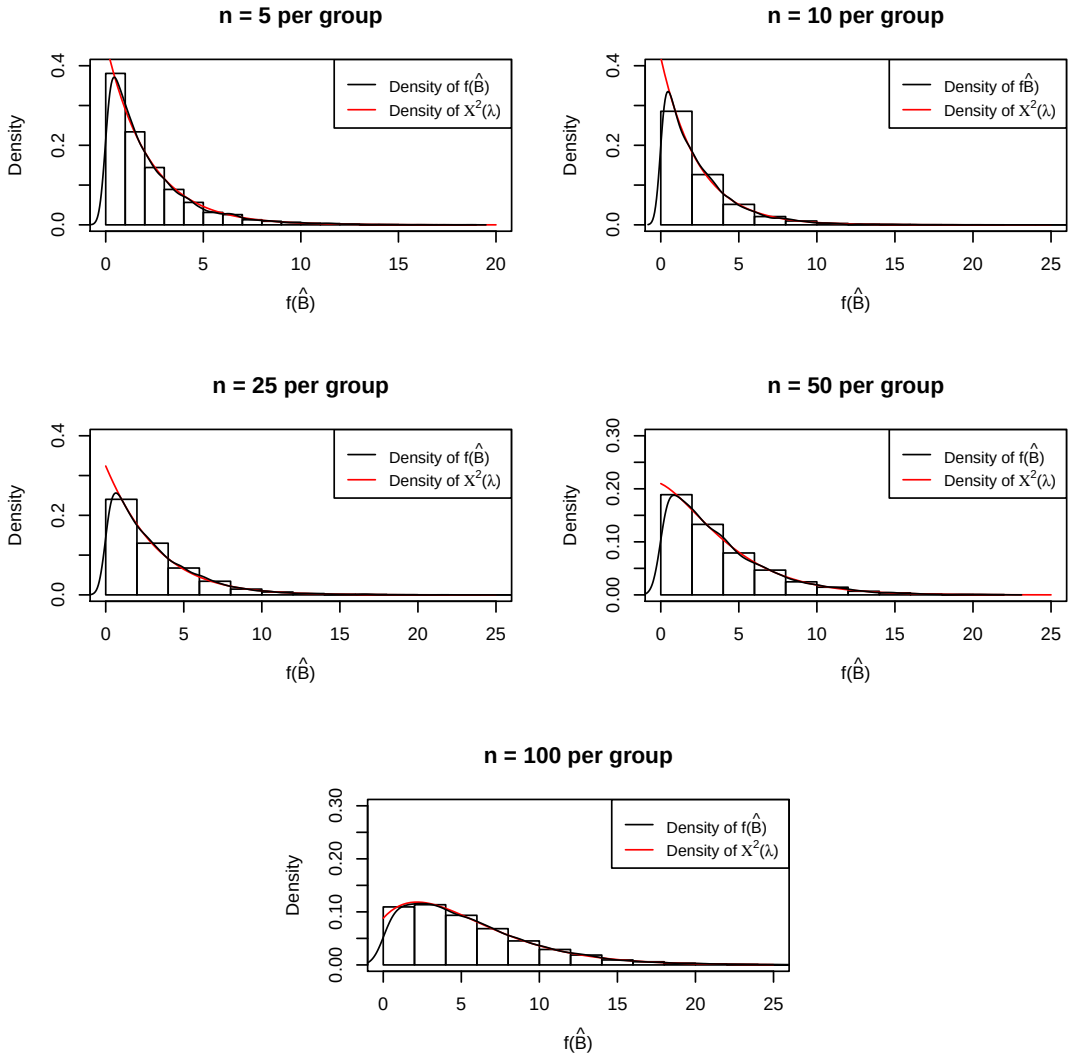


Figure 4.2: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from distributions with similar means and the same covariance matrix with a non-zero term for covariance between the two predictors (simulation setting 2). In each figure, the red line represents the distribution of the target distribution of ${\chi_2}^2(\lambda = \boldsymbol{\mu}^{*T}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*)$ and the black line represents the kernel density estimate of the empirical distribution of the simulated quantity $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$.

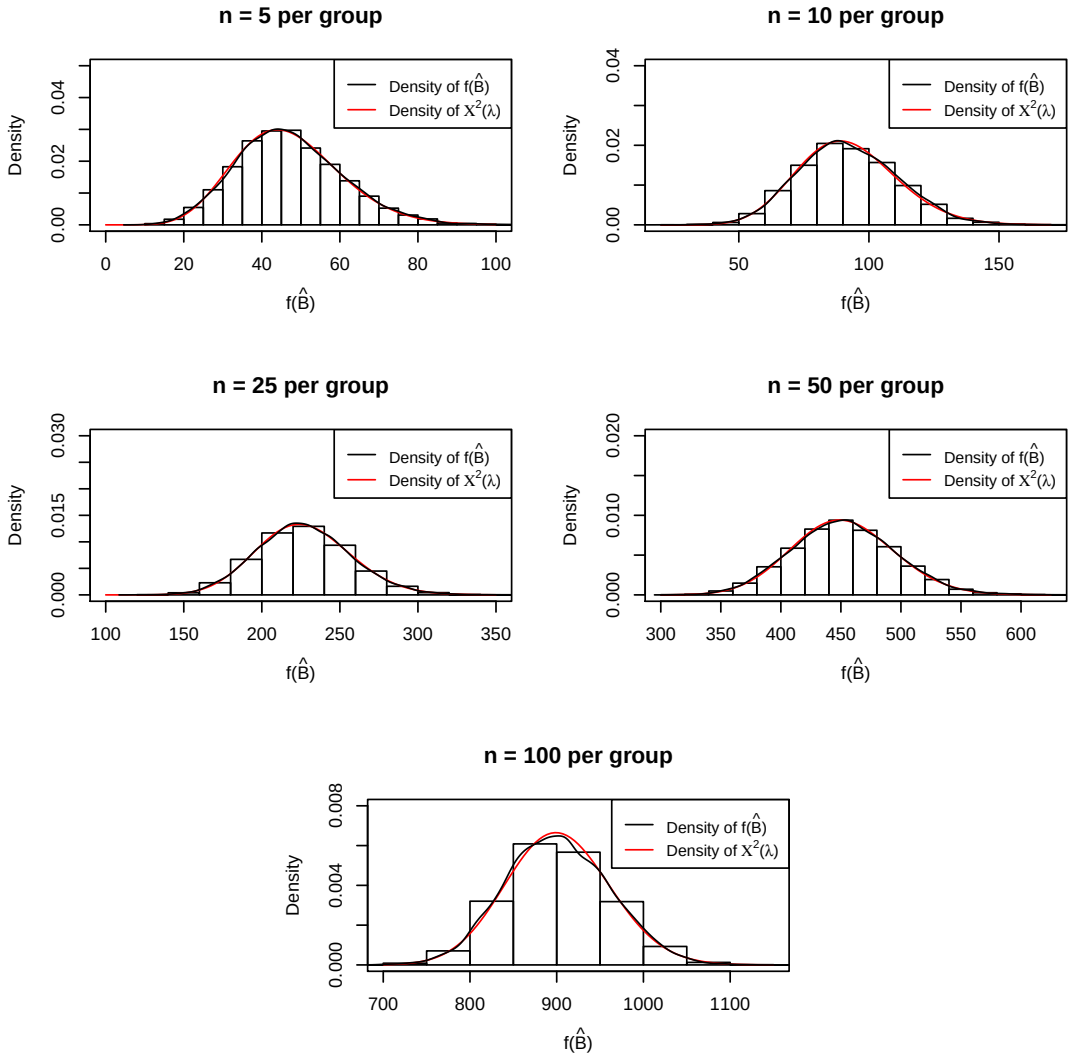


Figure 4.3: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from distributions with means far apart and the same covariance matrix with a zero term for covariance between the two predictors (simulation setting 3). In each figure, the red line represents the distribution of the target distribution of ${\chi_2}^2(\lambda = \boldsymbol{\mu}^{*T}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*)$ and the black line represents the kernel density estimate of the empirical distribution of the simulated quantity $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$.

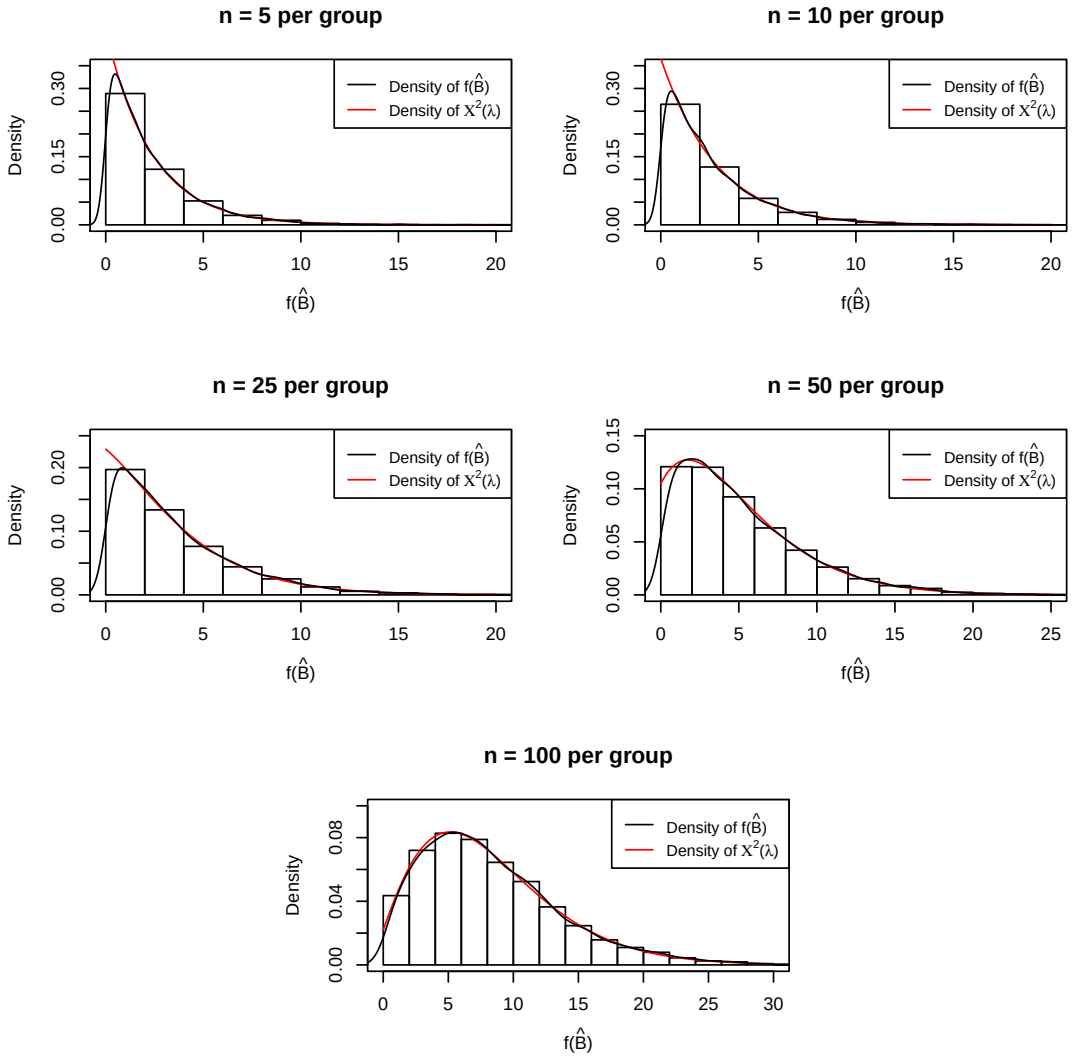


Figure 4.4: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from distributions with similar means and the same covariance matrix with a zero term for covariance between the two predictors (simulation setting 4). In each figure, the red line represents the distribution of the target distribution of ${\chi_2}^2(\lambda = \boldsymbol{\mu}^{*T}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*)$ and the black line represents the kernel density estimate of the empirical distribution of the simulated quantity $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$.

It is clear from these figures that as the sample sizes increase, the simulated distribution approaches the target distribution very quickly for all parameter settings. Quantile-quantile (q-q) plots are used for determining if data come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Here, the first data set is the simulated data and the second data set is 10,000 random sample from the ${\chi_2}^2(\lambda = {\boldsymbol{\mu}^*}^T {\boldsymbol{\Sigma}^*}^{-1} \boldsymbol{\mu}^*)$ distribution. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. Figures 4.5 - 4.8 show the q-q plots for each sample size under each of the four parameter settings. In simulations 2 and 4 that consider differences in distributions with close means, it seems that there is more departure from the reference line. However, overall the q-q plots show a linear relationship and do not provide evidence for differences in the distribution of $\frac{8n_1n_2}{n_1+n_2}\hat{B}$ and the target of ${\chi_2}^2(\lambda)$.

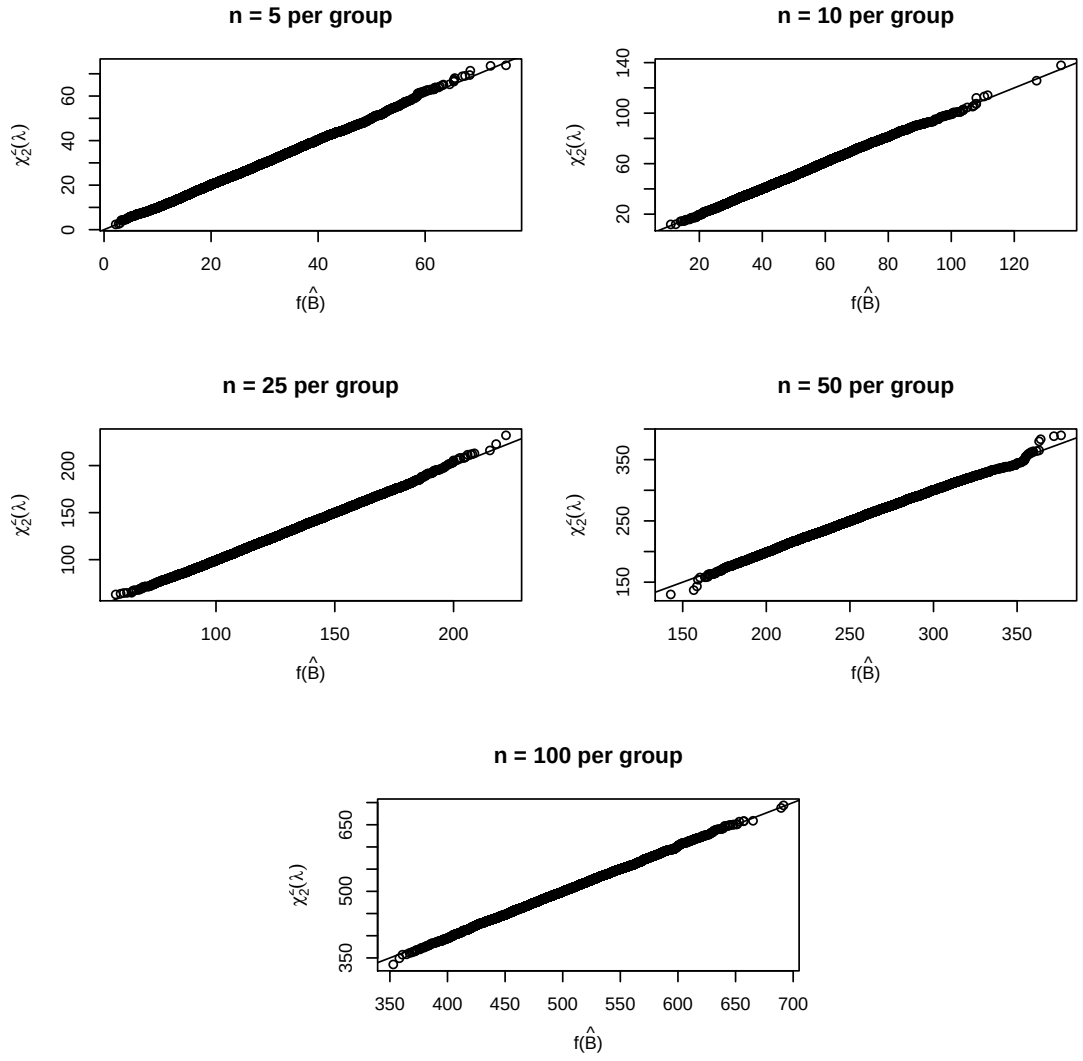


Figure 4.5: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from distributions with means far apart and the same covariance matrix with a non-zero term for covariance between the two predictors (simulation setting 1). In each figure, the quantiles of the simulated $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$ are plotted against the quantiles of the target distribution of ${\chi_2}^2(\lambda = \boldsymbol{\mu}^{*T}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*)$.

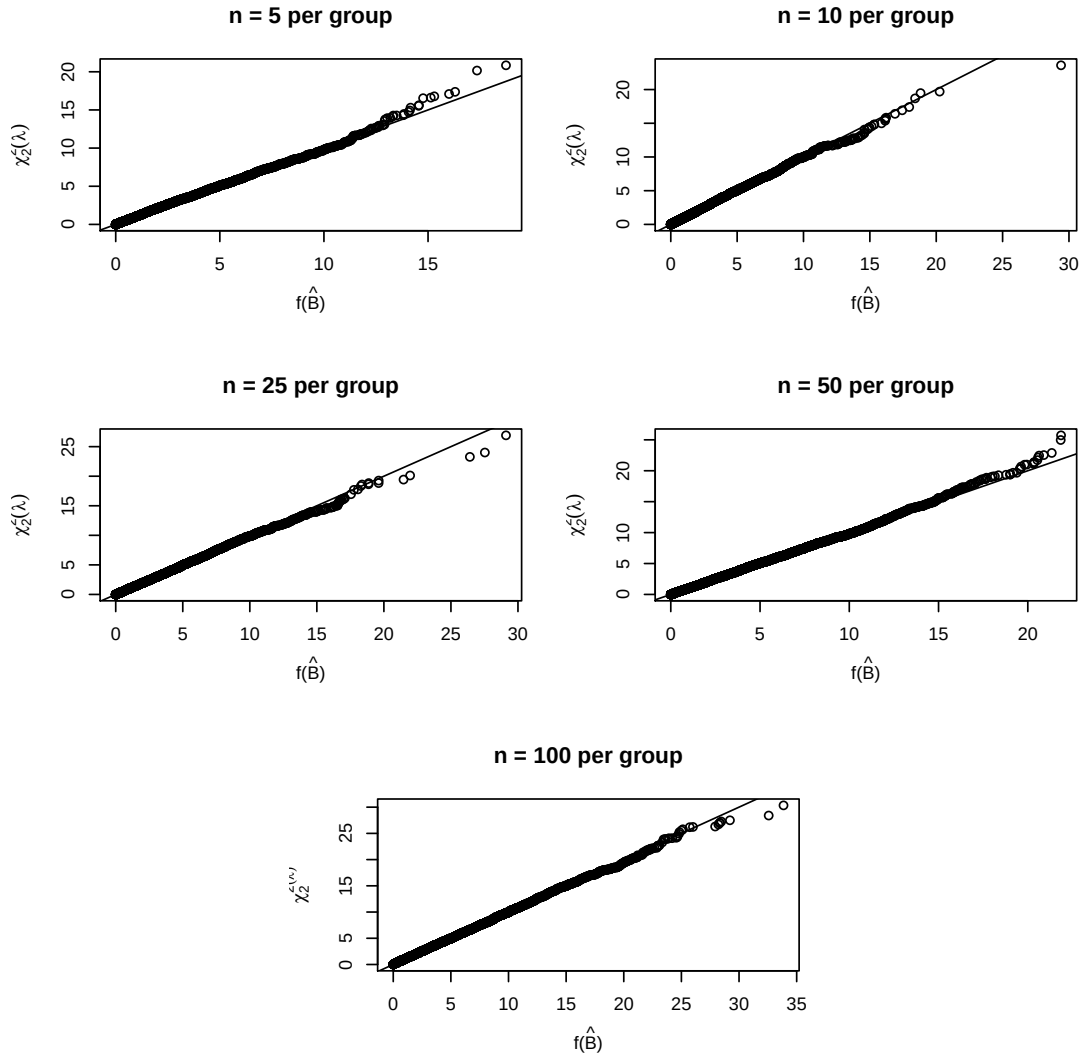


Figure 4.6: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from distributions with similar means and the same covariance matrix with a non-zero term for covariance between the two predictors (simulation setting 2). In each figure, the quantiles of the simulated $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$ are plotted against the quantiles of the target distribution of ${\chi_2}^2(\lambda = \boldsymbol{\mu}^{*T}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*)$.

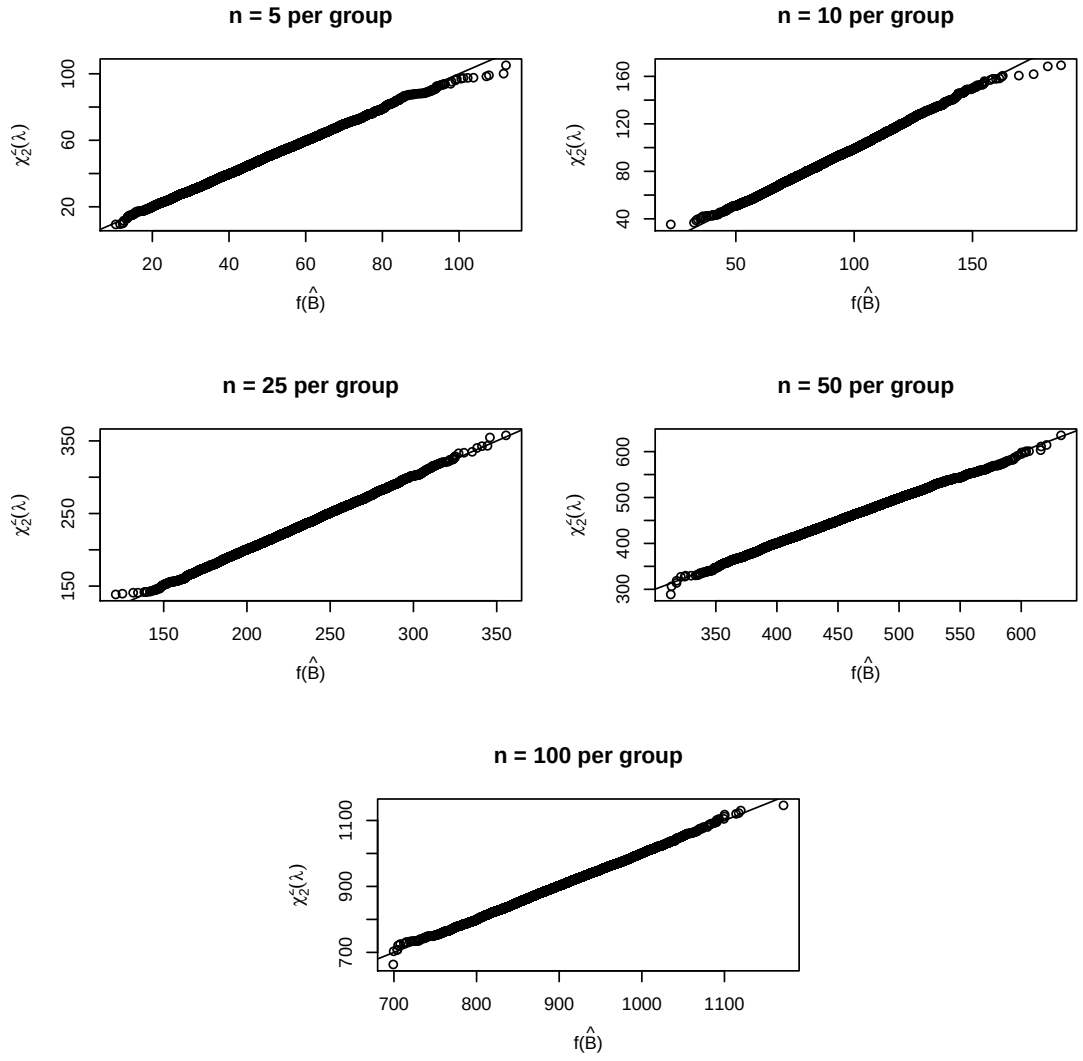


Figure 4.7: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from distributions with means far apart and the same covariance matrix with a zero term for covariance between the two predictors (simulation setting 3). In each figure, the quantiles of the simulated $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$ are plotted against the quantiles of the target distribution of ${\chi_2}^2(\lambda = {\boldsymbol{\mu}^*}^T {\boldsymbol{\Sigma}^*}^{-1} \boldsymbol{\mu}^*)$.

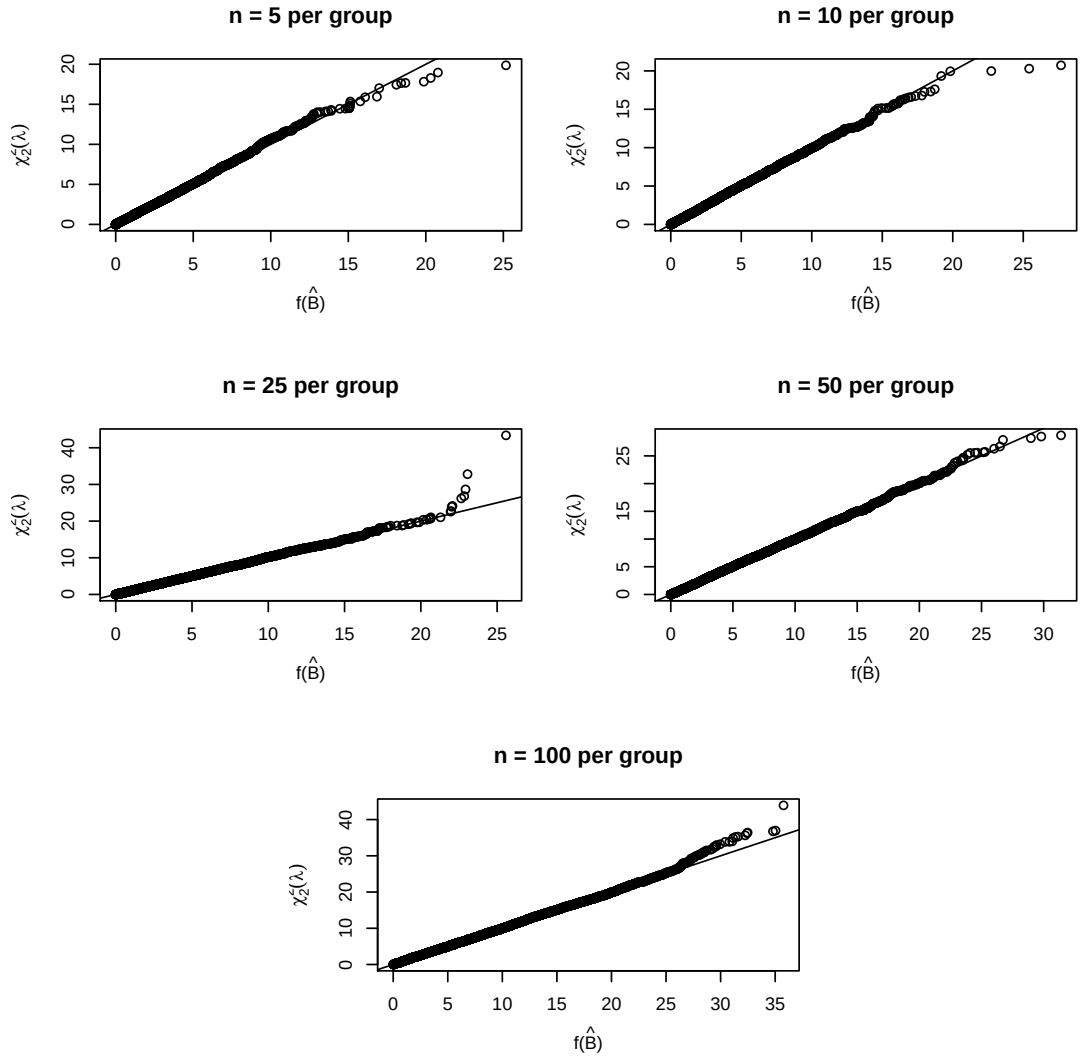


Figure 4.8: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from distributions with similar means and the same covariance matrix with a zero term for covariance between the two predictors (simulation setting 4). In each figure, the quantiles of the simulated $\frac{8n_1n_2}{(n_1+n_2)}\hat{B}$ are plotted against the quantiles of the target distribution of ${\chi_2}^2(\lambda = {\boldsymbol{\mu}^*}^T{\boldsymbol{\Sigma}^*}^{-1}\boldsymbol{\mu}^*)$.

It is clear from both the histograms and q-q plots that the simulated data follows the target distribution and that under these assumptions about $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$ $f(\hat{B}) \sim {\chi_2}^2(\lambda)$, where $\lambda = {\boldsymbol{\mu}^*}^T {\mathbf{\Sigma}^*}^{-1} \boldsymbol{\mu}^*$. Although this result is interesting, it is not particularly useful in the case where B-distance is used to identify interactions among predictors. In this situation, the covariances are not expected to be known nor equal, just independent. In the following sections, these assumptions will be relaxed.

4.1.3 Distribution of the Sample Bhattacharyya Distance Assuming Unknown Equal Covariances

Now consider relaxing the condition of known covariances, but assume that $\mathbf{\Sigma_1} = \mathbf{\Sigma_2}$ and assume that the groups have equal sample sizes, i.e., $n_1 = n_2 = n$. Now consider the first term of $\hat{B}$,

$$\frac{1}{8}(\bar{\boldsymbol{X}}_\mathbf{1} - \bar{\boldsymbol{X}}_\mathbf{2})^T \hat{\mathbf{\Sigma}}^{-1} (\bar{\boldsymbol{X}}_\mathbf{1} - \bar{\boldsymbol{X}}_\mathbf{2}).$$

Since the maximum likelihood estimates are being used and it is assumed that $\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = \mathbf{\Sigma}$,

$$\bar{\boldsymbol{X}}_\boldsymbol{i} \sim \mathcal{N}_p(\boldsymbol{\mu_i}, \frac{1}{n}\mathbf{\Sigma})$$

for $i = 1, 2$. Letting $\mathbf{X}^* = \bar{\boldsymbol{X}}_\mathbf{1} - \bar{\boldsymbol{X}}_\mathbf{2}$,

$$\mathbf{X}^* \sim \mathcal{N}_p\left(\boldsymbol{\mu}^*, \mathbf{\Sigma}^*\right).$$

where $\boldsymbol{\mu}^* = \boldsymbol{\mu_1} - \boldsymbol{\mu_2}$ and $\mathbf{\Sigma}^* = \frac{2}{n}\mathbf{\Sigma}$.

It is also known that $\hat{\mathbf{\Sigma}}_\mathbf{1}$ and $\hat{\mathbf{\Sigma}}_\mathbf{2}$ follow Wishart distributions,

$$\hat{\mathbf{\Sigma}}_\boldsymbol{i} \sim \mathrm{W}_p\left(n-1, \frac{1}{n-1}\mathbf{\Sigma}\right).$$

From this,

$$\hat{\mathbf{\Sigma}} = \frac{1}{2}(\hat{\mathbf{\Sigma}}_\mathbf{1} + \hat{\mathbf{\Sigma}}_\mathbf{2}) \sim \mathrm{W}_p\left(2(n-1), \frac{1}{2(n-1)}\mathbf{\Sigma}\right).$$

From Theorem 18.19 of [Arnold, 1981],

$$\frac{(2n-1-p)n}{2p(2n-2)}(\bar{\boldsymbol{X}}_\mathbf{1} - \bar{\boldsymbol{X}}_\mathbf{2})^T \hat{\mathbf{\Sigma}}^{-1} (\bar{\boldsymbol{X}}_\mathbf{1} - \bar{\boldsymbol{X}}_\mathbf{2}) \sim F_{p,2n-1-p}(\delta),$$

where

$$\delta = \frac{n}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T$$

is the non-centrality parameter.

Here $p = 2$ and so

$$\frac{n(2n-3)}{8(n-1)}(\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}})^T \hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}}) \sim F_{2,2n-3}(\delta),$$

where

$$\delta = \frac{n}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T$$

Let

$$Y = \frac{n(2n-3)}{8(n-1)}(\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}})^T \hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}})$$

and

$$Z = \frac{1}{8}(\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}})^T \hat{\boldsymbol{\Sigma}}^{-1}(\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}}),$$

noticing that Z is the first term of $\hat{B}$. Then

$$Y = \frac{n(2n-3)}{(n-1)}Z.$$

That is,

$$\frac{n(2n-3)}{(n-1)}Z \sim F_{2,2n-3}(\delta),$$

where

$$\delta = \frac{n}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T$$

and thus the distribution of the first term of $\hat{B}$ is now known in the case that $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2}$ and $n_1 = n_2$. Note that the expected value of the first term here is

$$\mathrm{E}[Z] = \frac{(n-1)(2+\delta)}{2n(2n-5)},$$

where

$$\delta = \frac{n}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T.$$

Hence, the first term of $\hat{B}$ is clearly a biased estimator of the first term of B.

Now consider the second term of the sample B-distance. Finding the distribution of this term

proves to be very difficult due to analytically intractable covariance parameters. I will show this here. Remember that it is assumed that $\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = \mathbf{\Sigma}$. So the second term of $\hat{B}$ is

$$\frac{1}{2}\ln\left(\frac{\det\hat{\mathbf{\Sigma}}}{\sqrt{\det\hat{\mathbf{\Sigma}}_1\det\hat{\mathbf{\Sigma}}_2}}\right),$$

which can be rewritten using properties of logarithms as

$$\frac{1}{2}\ln(\det(\hat{\mathbf{\Sigma}})) - \frac{1}{4}\ln(\det(\hat{\mathbf{\Sigma}}_1)) - \frac{1}{4}\ln(\det(\hat{\mathbf{\Sigma}}_2)). \tag{4.1}$$

Note the following corollary from Cai et al. [Cai et al., 2015] where n is the sample size, $\boldsymbol{S}$ is a sample covariance matrix, and p is the dimension of $\boldsymbol{S}$.

Corollary: If p is fixed, then the log determinant of $\boldsymbol{S}$ satisfies

$$\frac{\log\det\boldsymbol{S} - p(p+1)/(2(n-1)) - \log\det\mathbf{\Sigma}}{\sqrt{2p/(n-1)}} \xrightarrow{L} \mathcal{N}(0,1),$$

as $n \to \infty$. In the case of $\hat{\mathbf{\Sigma}}$ there are $p = 2$ dimensions and it is known that $\hat{\mathbf{\Sigma}}$ is a sample covariance matrix [Arnold, 1981]. Thus,

$$\frac{\sqrt{n-1}}{2}(\log\det\hat{\Sigma} - \frac{3}{n-1} - \log\det\Sigma) \xrightarrow{L} \mathcal{N}(0,1),$$

as $n \to \infty$ and so

$$\sqrt{n-1}\left(\frac{1}{2}\log\det\hat{\Sigma} - \frac{3}{2(n-1)} - \frac{1}{2}\log\det\Sigma\right) \xrightarrow{L} \mathcal{N}(0,1),$$

as $n \to \infty$.

In the case of $\hat{\mathbf{\Sigma}_1}$ and $\hat{\mathbf{\Sigma}_2}$, there are $p = 2$ dimensions and from the Corollary above,

$$\sqrt{n-1}\left(\frac{1}{4}\log\det\hat{\Sigma}_1 - \frac{3}{4(n-1)} - \frac{1}{4}\log\det\Sigma_1\right) \xrightarrow{L} \mathcal{N}\left(0,\frac{1}{4}\right),$$

as $n \to \infty$ and in the analogous case,

$$\sqrt{n-1}\left(\frac{1}{4}\log\det\hat{\Sigma}_2 - \frac{3}{4(n-1)} - \frac{1}{4}\log\det\Sigma_2\right) \xrightarrow{L} \mathcal{N}\left(0,\frac{1}{4}\right),$$

as $n \to \infty$.

It is known that the sum of three normally distributed random variables is also normally distributed and that the mean is equal to the sum of the three means. Thus, 4.1 follows a normal distribution and its expected value is approximately

$$\begin{aligned}&\frac{3}{2(n-1)}+\frac{1}{2}\log(\det(\boldsymbol{\Sigma}))-\frac{3}{4(n-1)}-\frac{1}{4}\log(\det(\boldsymbol{\Sigma}_1))-\frac{3}{4(n-1)}-\frac{1}{4}\log(\det(\boldsymbol{\Sigma}_2))\\&=\frac{1}{2}\log(\det(\boldsymbol{\Sigma}))-\frac{1}{4}\log(\det(\boldsymbol{\Sigma}_1))-\frac{1}{4}\log(\det(\boldsymbol{\Sigma}_2))\end{aligned}$$

The mean of the sum of these three terms is known, but the variance is also needed to derive the full distribution. Next, I would like to calculate the variance of the sum of these three pieces. Note that if the terms are independent, then the variance terms from the three parts can be summed. However, in the event that the variables are not independent, the variances are not additive due to the correlations between variables. It has been assumed that $\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{2}}$ are independent. However, since

$$\hat{\boldsymbol{\Sigma}}=\frac{\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}+\hat{\boldsymbol{\Sigma}}_{\mathbf{2}}}{2},$$

it is obvious that $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}$ are not independent and that $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{2}}$ are also not independent. Thus, since the term $\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}|-\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|-\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{2}}|$ is being considered,

$$\begin{aligned}\operatorname{Var}(\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}|-\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|-\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{2}}|) &= \operatorname{Var}(\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}|)+\operatorname{Var}(\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|)+\operatorname{Var}(\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{2}}|)\\&- 2\operatorname{Cov}(\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}|,\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|)-2\operatorname{Cov}(\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}|,\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{2}}|)\\&- 2\operatorname{Cov}(\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|,\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{2}}|).\end{aligned}$$

The variance terms are known, but the covariance terms need to be calculated. First consider

$$\begin{aligned}\operatorname{Cov}\left(\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}|,\,\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|\right)&=\operatorname{E}\left[\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}|\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|\right]-\operatorname{E}\left[\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}|\right]\operatorname{E}\left[\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|\right]\\&=\operatorname{E}\left[\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}|\frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|\right]-\left(\frac{3}{2(n-1)}+\frac{1}{2}\log|\boldsymbol{\Sigma}|\right)\left(\frac{3}{4(n-1)}+\frac{1}{4}\log|\boldsymbol{\Sigma}_{\mathbf{1}}|\right)\\&=\frac{1}{8}\operatorname{E}\left[\log|\hat{\boldsymbol{\Sigma}}|\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|\right]-\frac{9}{8(n-1)}-\frac{3}{8(n-1)}\log|\boldsymbol{\Sigma}_{\mathbf{1}}|\\&-\frac{3}{8(n-1)}\log|\boldsymbol{\Sigma}|-\frac{1}{8}\log|\boldsymbol{\Sigma}|\,\log\,|\boldsymbol{\Sigma}_{\mathbf{1}}|\end{aligned}$$

Since $\log|\hat{\boldsymbol{\Sigma}}|$ and $\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|$ are not independent, $\operatorname{E}\left[\log|\hat{\boldsymbol{\Sigma}}|\log|\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}|\right]$ cannot be written as

$\mathrm{E}\left[\log|\hat{\mathbf{\Sigma}}|\right]\mathrm{E}\left[\log|\hat{\mathbf{\Sigma}}_{\mathbf{1}}|\right]$. Also, since the joint distribution of $\log|\hat{\Sigma}|$ and $\log|\hat{\Sigma_1}|$ is unknown, this term cannot be analytically calculated. Notice that when considering $\mathrm{Cov}(\log|\hat{\mathbf{\Sigma}}|, \log|\hat{\mathbf{\Sigma}}_{\mathbf{2}}|)$, the same issues will arise. Thus, this covariance term is analytically intractable and the distribution of the second term of the $\hat{B}$ cannot be derived under in the case where $\mathbf{\Sigma_1} = \mathbf{\Sigma_2}$ and $n_1 = n_2$. However, the asymptotic mean is known for both terms of $\hat{B}$ and thus, $\mathrm{E}[\hat{B}]$ can be calculated. It it is clear, though, that $\hat{B}$ is biased. Thus, inference about the bias of $\hat{B}$ can be made and is considered in the following section.

4.1.4 Confidence Intervals for Bhattacharyya Distance

Even under assumptions of $\mathbf{\Sigma_1} = \mathbf{\Sigma_2}$, the distribution of $\hat{B}$ cannot be derived. In order to make inference about $\hat{B}$, I will first consider interval estimates to help estimate B-distance. Since the distribution of $\hat{B}$ is not known, percentile intervals can be used. Percentile intervals are a common method when working with estimators with unknown sampling distributions [Wilcox, 2011]. The distribution of $\hat{B}$ is unknown and clearly not normal since B-distance can never be less than zero. In fact, the sample B-distance will equal zero with probability zero. Thus, the non-normal nature of $\hat{B}$ needs to be accounted for when creating these intervals. Further, $\hat{B}$ is a biased estimator of B and so this will be accounted for when creating percentile intervals.

Since the distribution of $\hat{B}$ is unknown, percentile intervals for the expected value of $\hat{B}$ are created via bootstrapping. In simulation studies, the expected value $\mathrm{E}[\hat{B}]$ are calculated through Monte Carlo integration in order to estimate coverage probabilities.

Simulations.— Simulations are conducted to examine the accuracy of percentile intervals created via bootstrapping. The accuracy of these intervals is measured by coverage probability, i.e., how often the calculated intervals contain $\mathrm{E}[\hat{B}]$ as calculated by Monte Carlo integration. $\mathrm{E}[\hat{B}]$ is calculated as the mean of $\hat{B}$ from 10,000 simulations for each of 6 sample sizes ranging from 5 to 200. Then percentile intervals are created from 1,000 data sets. In each data set, 1,000 bootstrap samples are taken within groups with replacement from the original data and $\hat{B}$ is calculated for each bootstrap sample. The lower and upper limits of the interval are taken to be the 2.5^{th} and 97.5^{th} percentiles of the bootstrap samples of $\hat{B}$, respectively. This produces a 95% confidence interval for $\hat{B}$. The coverage probability is then calculated as the percentage of the 1,000 intervals that cover $\mathrm{E}[\hat{B}]$, and then compared to the desired probability of 95%.

Table 4.2: Simulation parameters for evaluating the coverage probabilities of percentile intervals for $\hat{B}$ are shown here. Column 1 denotes the number of the simulation setting. Columns 2 and 3 contain the means for the two distributions, while Columns 4 and 5 display the covariances of the two distributions. The different sample sizes that are used in the simulations are found in Column 6.

Simulation	$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\Sigma_1}$	$\boldsymbol{\Sigma_2}$	n
1	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
2	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
3	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
4	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5 10 25 50 100 200

Data are generated under four parameter settings incorporating different values for means and covariances. For each combination of mean and covariance choices, sample sizes from $n = 5$ per group up to $n = 200$ per group are used. The first setting generates data from the same distributions for the response groups. The remaining three combinations of parameter choices are chosen to evaluate the accuracy of this method under various conditions of differences in distributions. These conditions include when groups have the same means but different covariances, when groups have the same covariances but different means, and lastly when the groups have both different means and different covariances. All of these combinations need to be considered in order to determine what circumstances the percentile intervals have good coverage probability. The parameter settings for the simulations can be found in Table 4.2.

Simulation Results.— Simulations are conducted to evaluate the coverage probability of percentile intervals under various conditions. Table 4.3 displays these coverage probabilities.

Table 4.3: Simulation results for evaluating the coverage probabilities of percentile intervals for $\hat{B}$ are shown here. The coverage probabilities (Column 7) clearly increase as the sample size (Column 6) increases. Columns 1 and 2 contain the means for the two distributions, while Columns 4 and 5 display the covariances of the two distributions.

$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\Sigma_1}$	$\boldsymbol{\Sigma_2}$	n	Coverage Probability
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0.776
				10	0.88
				25	0.926
				50	0.94
				100	0.958
				200	0.95
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0.767
				10	0.895
				25	0.926
				50	0.941
				100	0.953
				200	0.953
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0.799
				10	0.843
				25	0.915
				50	0.929
				100	0.922
				200	0.934
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0.771
				10	0.869
				25	0.913
				50	0.930
				100	0.936
				200	0.933

It is clear that as the sample size increases, so does the coverage probability of the interval. In the first two simulation settings, 95% coverage is reached with sample sizes of $n = 100$ and $n = 200$. For the last two simulation settings, the intervals reach a little over 93% coverage, which is very close to 95%. Thus, 50 seems to be an acceptable sample size to achieve appropriate coverage probability. Therefore, it seems that this method is an appropriate way to create confidence intervals for $\hat{B}$. Now that there exists a method for creating confidence intervals, it would be useful to have a method to perform hypothesis testing as well.

4.1.5 Hypothesis Testing

When performing hypothesis testing, the goal is to determine if the response groups come from the same distribution or from different distributions. If it can be determined that the response groups come from two different distributions, this may be evidence of an interaction between the two variables of interest or of main effects of at least one of the variables. Thus, here the null and alternative hypotheses being tested are

H_0: The response groups come from the same distribution

H_1: The response groups come from different distributions.

In Section 4.1.2, the distribution of $\hat{B}$ was not able to be derived without making strict and unrealistic assumptions about $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$. However, it may be possible to derive the null distribution of a function of $\hat{B}$ that can be used for hypothesis testing. Since I am interested in using B-distance to identify interaction effects, it would be useful to have this distribution in order to perform hypothesis testing.

It is known that Hellinger's Distance is a transformation of B-distance that is often used for variable selection. Some information is known about the asymptotic distribution of the Hellinger distance and therefore the relationship between Hellinger distance and B-distance can be used to derive the distribution of a function of $\hat{B}$. Recall that if π_1 and π_2 are two populations and $f(x, \theta_i)$ is the fixed density of the random vector X in π_i. then the Hellinger distance between the two populations is

$$\Delta = \int \{f^{\frac{1}{2}}(x, \theta_1) - f^{\frac{1}{2}}(x, \theta_2)\}^2 d\lambda(x). \tag{4.2}$$

Let

$$\theta_1 = (\mu_{11}, \mu_{12}, \sigma_{1,11}, \sigma_{1,12}, \sigma_{1,22})$$

and

$$\theta_1 = (\mu_{21}, \mu_{22}, \sigma_{2,11}, \sigma_{2,12}, \sigma_{2,22}).$$

Under the null hypothesis $H_0 : \theta_1 = \theta_2$, the following is true about the sample Hellinger Distance

$$\frac{4n_1n_2}{n_1+n_2}\hat{\Delta} \rightarrow \chi^2_t \tag{4.3}$$

in distribution as $n_1 \rightarrow +\infty$, $n_2 \rightarrow +\infty$ and $\frac{n_1}{n_1+n_2} \rightarrow u \in]0,1]$ [Alba-Fernández et al., 2005]. That is, letting

$$X_n = \frac{4n_1n_2}{n_1+n_2}\hat{\Delta},$$

then asymptotically,

$$X_n \xrightarrow{d} \chi^2_t$$

where $\xrightarrow{d}$ denotes convergence in distribution.

Next, consider Slutsky's Theorem.

Theorem: Let X_n and Y_n be sequences of scalar/vector/matrix random elements. If X_n converges in distribution to a random element X and Y_n converges in probability to a constant c, then

$$X_n + Y_n \xrightarrow{d} X + c; \tag{4.4}$$

$$X_nY_n \xrightarrow{d} cX; \tag{4.5}$$

$$X_n/Y_n \xrightarrow{d} X/c, \text{ provided that c is invertible,} \tag{4.6}$$

where $\xrightarrow{d}$ denotes convergence in distribution.

Now let

$$f(\hat{B}) = \frac{4n_1n_2}{n_1+n_2}\left(1 - e^{-\hat{B}}\right) \tag{4.7}$$

where $\hat{B}$ is the sample B-distance. Then, from the definition of the two distances, the following relationship is true:

$$f(\hat{B}) = \frac{X_n}{2}.$$

Using Slutsky's Theorem, let $Y_n = 2$ for all n and consider (4.6). Then,

$$\frac{X_n}{Y_n} \xrightarrow{d} \frac{\chi_t^2}{2},$$

under the conditions associated with (4.3).

Note that a Gamma (k, θ) random variable with $k = \frac{\nu}{2}$ and $\theta = 2$, is a chi-squared random variable with ν degrees of freedom. And by properties of the gamma distribution, if

$$X \sim \text{Gamma}(k, \theta),$$

then for any $c > 0$,

$$cX \sim \text{Gamma}(k, c\,\theta),$$

(proof using moment generating functions).

Since $c = \frac{1}{2}$ here,

$$\frac{X_n}{Y_n} \xrightarrow{d} \Gamma(\frac{t}{2}, 1).$$

And since $f(\hat{B}) = \frac{X_n}{Y_n}$,

$$f(\hat{B}) \xrightarrow{d} \Gamma\left(\frac{t}{2}, 1\right)$$

where t is the number of parameters being estimated and n_1 and n_2 are the sample sizes for groups 1 and 2, respectively. In the case where all parameters are assumed unknown and the null hypothesis is assumed, only 5 parameters need to be estimated. That is, μ_{11}, μ_{22}, σ_{11}, σ_{22}, and $\sigma_{12} = \sigma_{21}$ need to be estimated, where

$$\boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \begin{pmatrix} \mu_{11} \\ \mu_{22} \end{pmatrix}$$

and

$$\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

With $t = 5$, the result is that

$$f(\hat{B}) \xrightarrow{d} \Gamma(2.5, 1). \tag{4.8}$$

Table 4.4: Simulation parameters for comparing the distribution of $\hat{B}$ to the $\Gamma(2.5, 1)$ distribution are shown here. Column 1 denotes the number of the simulation setting. Column 2 contains the mean for the two distributions, while Column 3 displays the covariance of the two distributions. The different sample sizes that are used in the simulations are found in Column 4.

Simulation	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	n
1	$\begin{pmatrix} 3 \\ 3 \end{pmatrix}$	$\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$	5 10 25 50 100
2	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$	5 10 25 50 100
3	$\begin{pmatrix} 3 \\ 3 \end{pmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	5 10 25 50 100
4	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	5 10 25 50 100

Simulations.— Simulations are conducted to examine the asymptotic convergence of $f(\hat{B})$ under the null distribution to a $\Gamma(2.5, 1)$ distribution as the sample sizes of the two groups, n_1 and n_2, increase. The two groups are simulated from the same distribution for four combinations of means and covariances. The parameter settings include zero and nonzero means, as well as covariances that include both a zero and a nonzero covariance term between the two predictors. Thus, the null distribution of $\hat{B}$ can be examined under various combinations of parameters to display what effect these differences have on the convergence of the distribution, if any. The simulation settings can be seen in Table 4.4.

Ten thousand data sets are simulated under these conditions. For each data set, $f(\hat{B})$ is calculated and then plotted in a histogram. The sample distribution of $f(\hat{B})$ that arises from these simulations can then be compared to the target distribution of $\Gamma(2.5, 1)$.

Simulation Results.— Simulations are conducted to examine the asymptotic convergence of $f(\hat{B})$ to a $\Gamma(2.5, 1)$ distribution as the sample sizes of the two groups, n_1 and n_2, increase. Figures 4.9 - 4.12 show results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group. In each figure, the blue line represents the density of the target distribution and

the red line represents the kernel density estimate of the empirical distribution of the simulated $f(\hat{B})$ values. It is clear from these figures that as the sample sizes increase, the simulated distribution approaches the target distribution fairly quickly. However, it does not match well for small sample sizes less than 25 per group. This is true across all parameter settings.

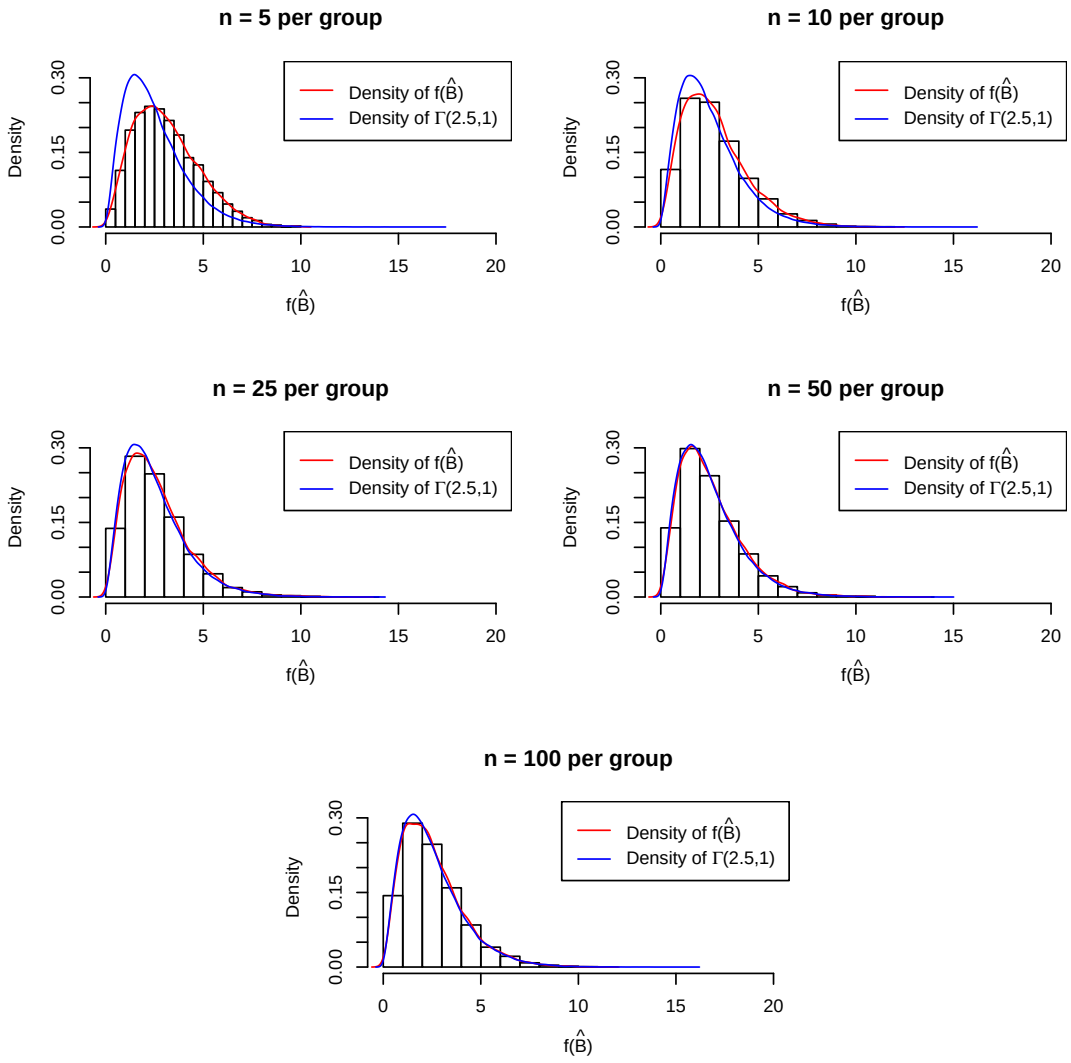


Figure 4.9: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from a distribution with non-zero mean and a covariance matrix with a non-zero term for covariance between the two predictors (simulation setting 1). In each figure, the blue line represents the distribution of the target distribution of $\Gamma(2.5, 1)$ and the red line represents the kernel density estimate of the empirical distribution of the simulated quantity $f(\hat{B})$.

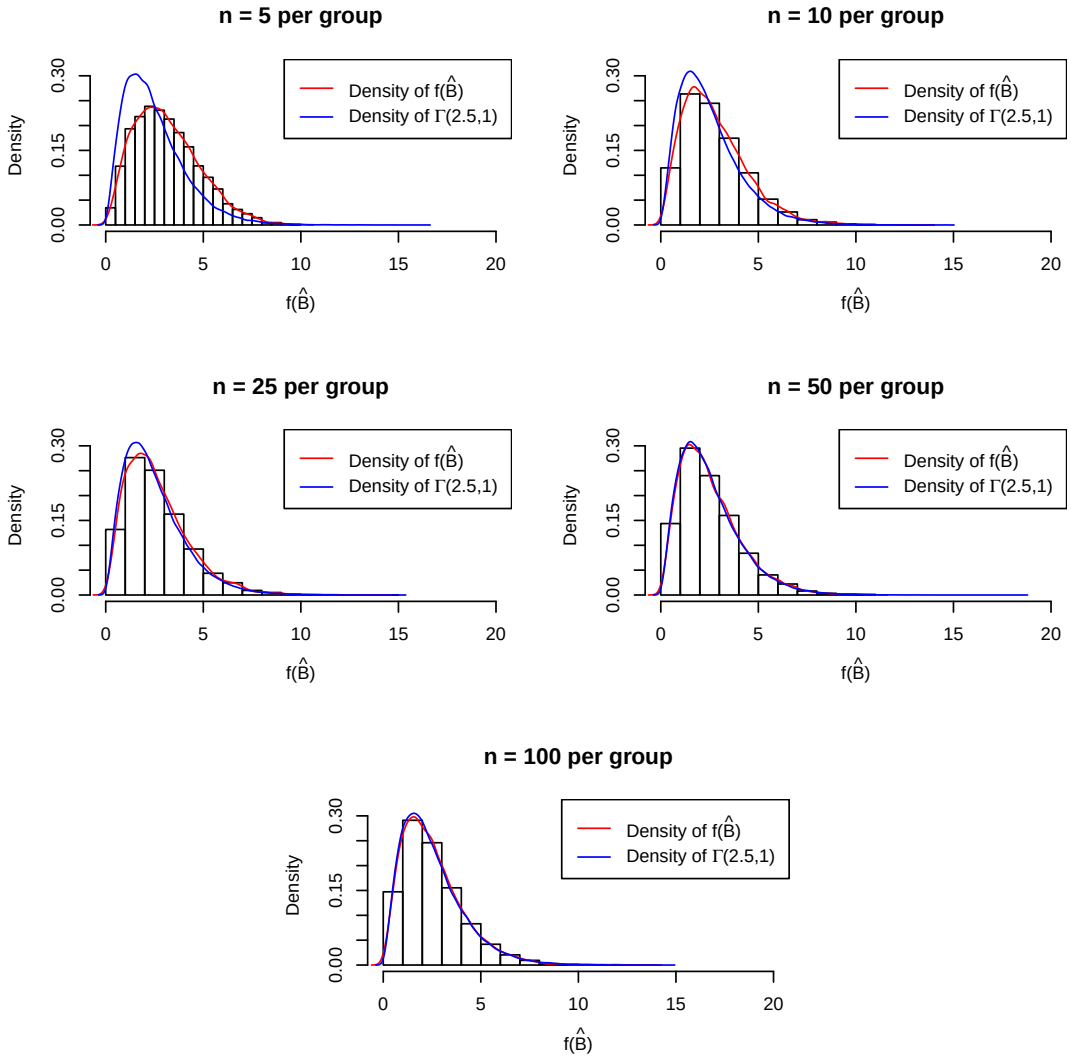


Figure 4.10: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from a distribution with zero mean and a covariance matrix with a non-zero term for covariance between the two predictors (simulation setting 2). In each figure, the blue line represents the distribution of the target distribution of $\Gamma(2.5, 1)$ and the red line represents the kernel density estimate of the empirical distribution of the simulated quantity $f(\hat{B})$.

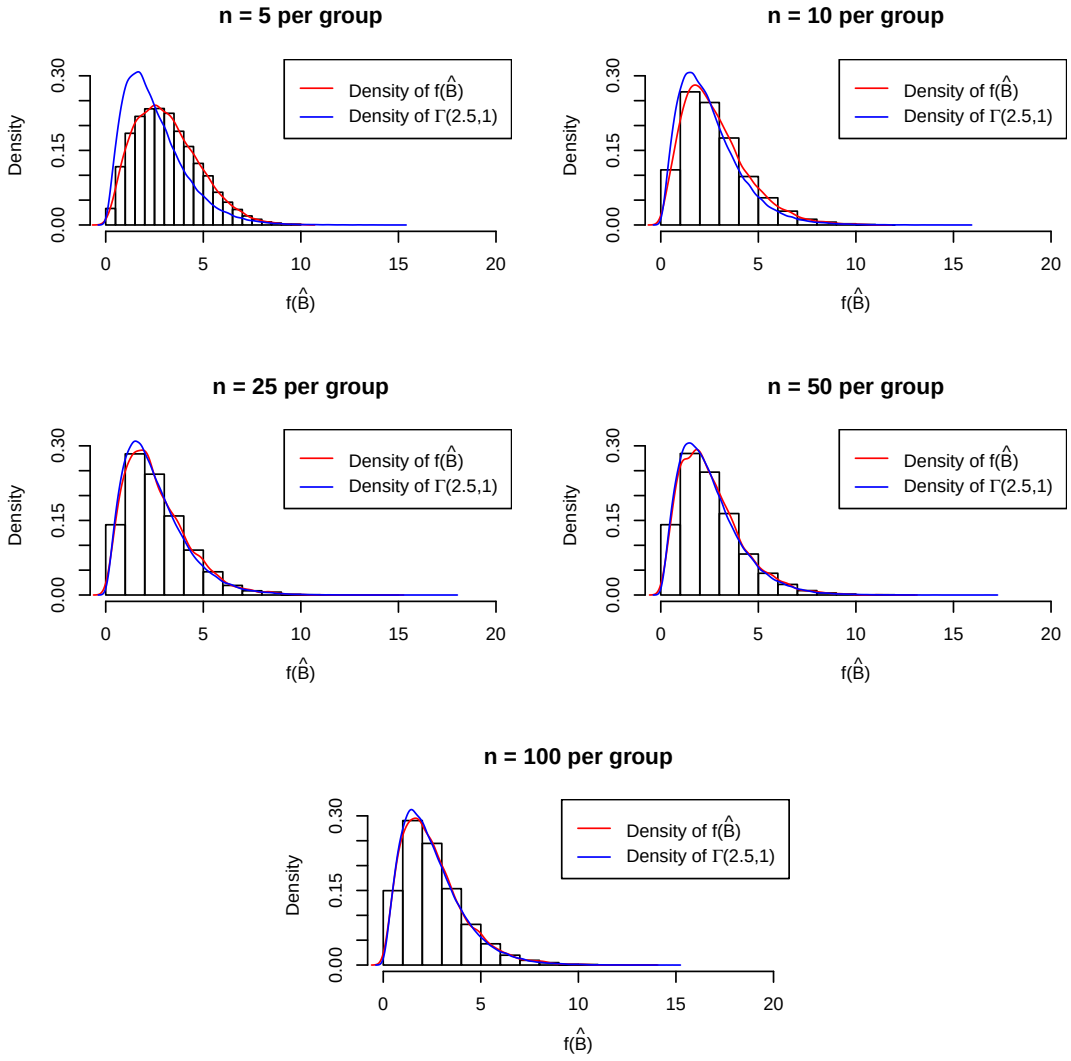


Figure 4.11: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from a distribution with non-zero mean and a covariance matrix with a zero term for covariance between the two predictors (simulation setting 3). In each figure, the blue line represents the distribution of the target distribution of $\Gamma(2.5, 1)$ and the red line represents the kernel density estimate of the empirical distribution of the simulated quantity $f(\hat{B})$.

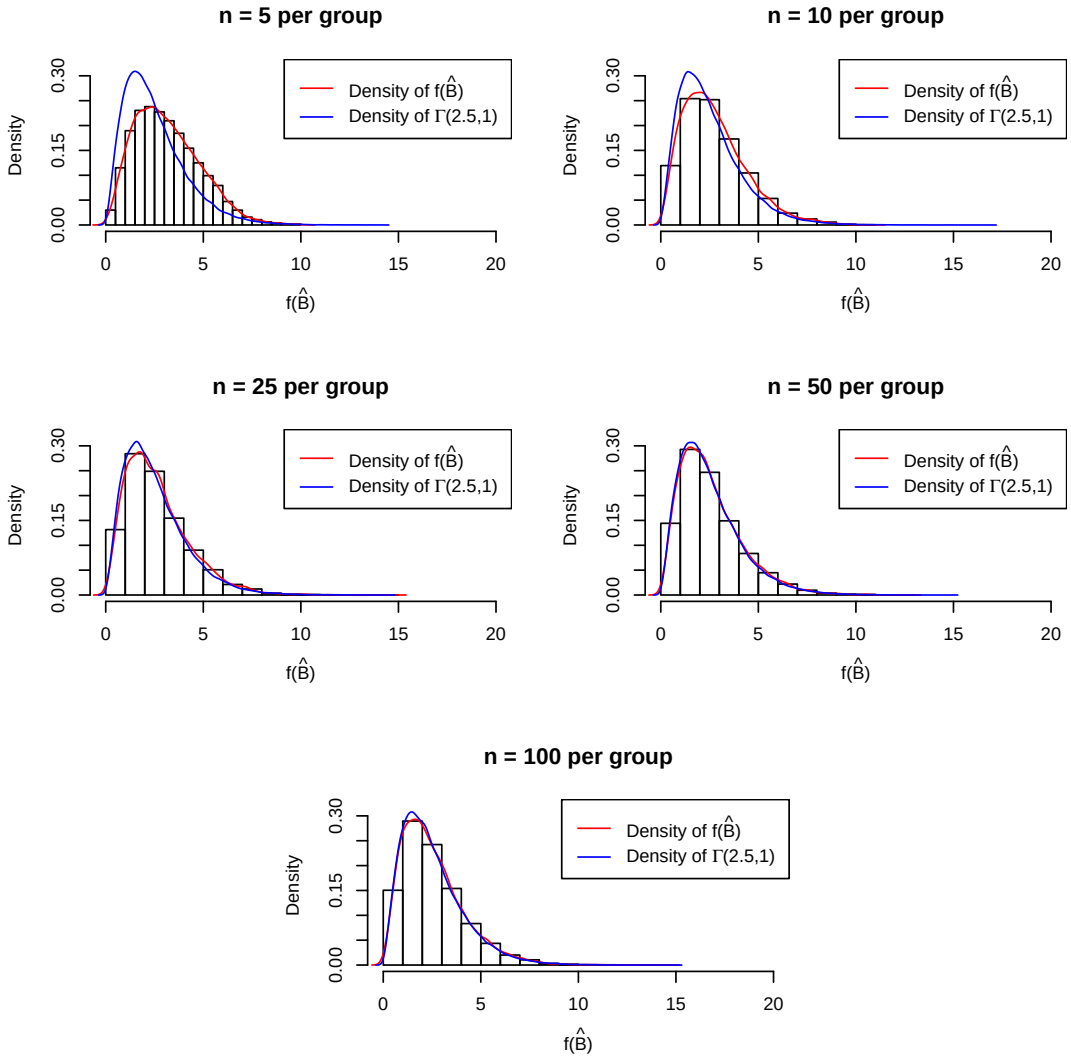


Figure 4.12: Results from 10,000 simulated data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group are plotted here. Data are simulated from a distribution with zero mean and a covariance matrix with a zero term for covariance between the two predictors (simulation setting 4). In each figure, the blue line represents the distribution of the target distribution of $\Gamma(2.5, 1)$ and the red line represents the kernel density estimate of the empirical distribution of the simulated quantity $f(\hat{B})$.

Now that the asymptotic null distribution is known, it can be used to test the null hypothesis $H_0 : \theta_1 = \theta_2$, that is, that the two groups have the same distribution. Thus, the type I error rate needs to be checked at each sample size in these four parameter settings. This comparison can been seen in Figure 4.13. The plots here show the simulated type I error rates against the true type I error rates for each of the four simulation settings. The black crosses, red triangles, yellow squares, green circles, and blue diamonds represent the type I error rates of 10,000 simulations with $n = 5$, 10, 25, 50, and 100 respectively. Simulated type I error rates are calculated as the percentage of simulated data that are greater than or equal to the value from quantile function of the specific gamma distribution that is being compared to for each of the desired probabilities. For example, if the desired probability is chosen to be $\alpha = 0.05$, then first the value at which the probability of the random variable is less than or equal to the given $1 - \alpha$ is found for the specific gamma distribution of interest $\Gamma(2.5, 1)$. In R, this can be done by typing in qgamma(0.95,shape=2.5,scale=1) and it is seen that this value is about 5.54. Thus, in order to test that the distribution of the sample B-distance under the null hypothesis matches the one defined, the percent of the sample B-distances that are at least as big as 5.54 need to be calculated to see if it is close to the desired α.

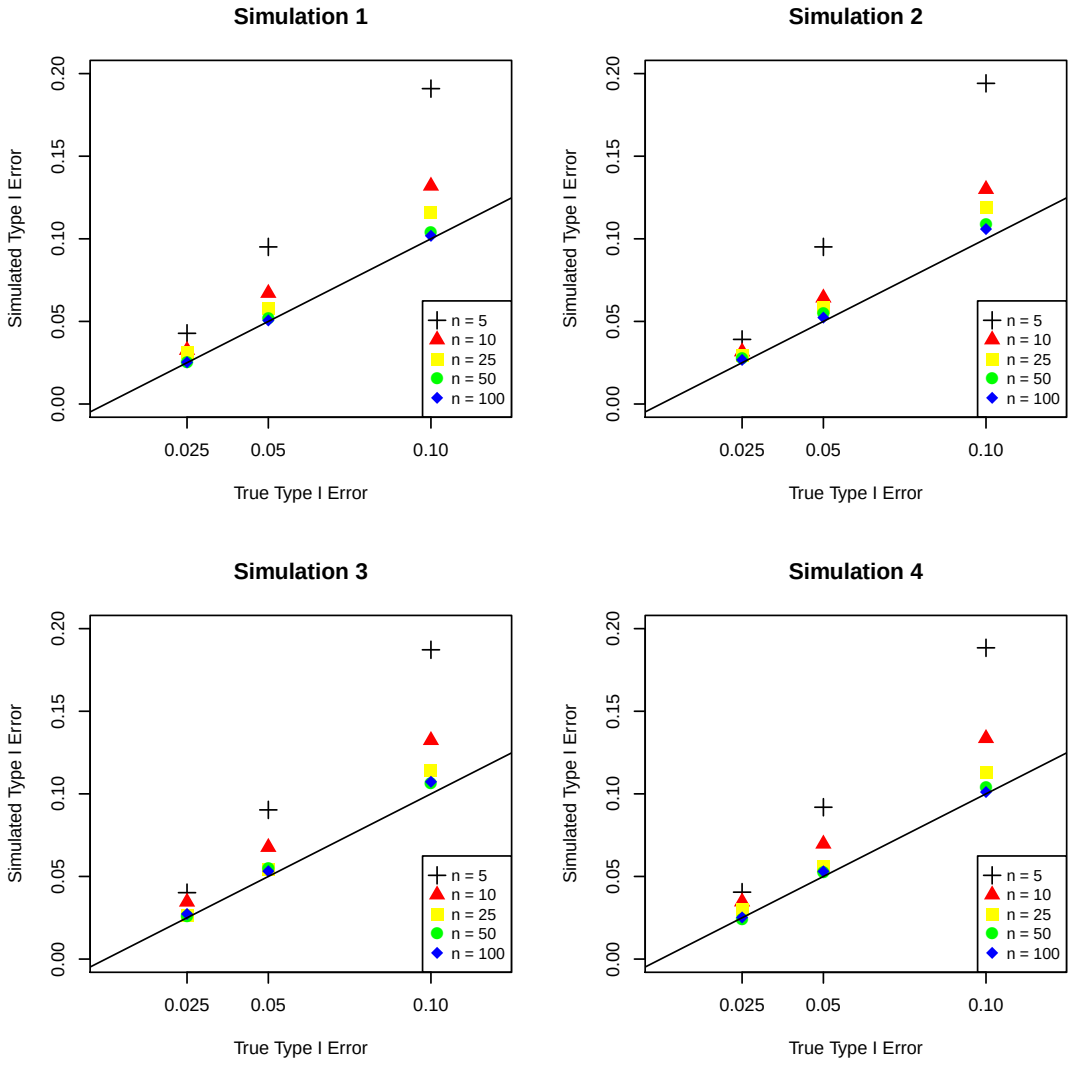


Figure 4.13: Simulated type I error rates calculated from 10,000 data sets are plotted against the true type I error rates of the asymptotic null distribution of $\Gamma(2.5, 1)$ for each of four parameter settings. The black crosses, red triangles, yellow squares, green circles, and blue diamonds represent the type I error rates of 10,000 simulations with $n = 5$, 10, 25, 50, and 100 respectively.

A table of these results can be found in the Appendix, A.2. It is clear from both the plot and table that the type I error rate is higher than desired at the small sample sizes of 5 and 10 per group. However, the type I error rate is controlled well at the 0.025 and 0.05 levels for the larger sample sizes of 25, 50, and 100 per group. Thus, in the case of small sample sizes, this asymptotic result cannot be used, because the type I error rates are too high. If the asymptotic results are used for the small sample sizes of interest, there is a risk of concluding that a sample B-distance is statistically significant when it is not more often than is comfortable. However, now there exists a general idea about how large the group sample sizes need to be in order to use this asymptotic test. In the case of small sample sizes, other statistical tools, such as permutation testing, can be used to conduct hypothesis tests and p-values.

Comparison of Asymptotic Test with Testing done via LRT Statistic.— Now that a method exists to perform hypothesis testing for adequate sample sizes based on an asymptotic distribution, it would be useful to compare this method to a more traditional hypothesis test. This can be done by performing a likelihood ratio test (LRT). The LRT statistic is calculated by comparing the full model including an interaction term to the reduced model with just an intercept term. These models are evaluated with logistic regression and then compared using an LRT statistic. Simulation studies are conducted to compare both the type I error rates and power of hypothesis testing based on the asymptotic distribution of $f(\hat{B})$ and testing by means of the LRT.

Simulations.— Simulations are conducted to examine the effectiveness of both hypothesis testing based on the asymptotic distribution of $f(\hat{B})$ and testing via the LRT statistic as the sample sizes of the two groups, n_1 and n_2, increase. Effectiveness is evaluated by measuring both the type I error rates and power of the two methods. Data are generated under the six parameter settings in Table 4.5. The first parameter setting will be used to evaluate the type I error rates of the testing methods since data from each group are simulated from the same distribution. The next five settings are designed to test for power since the groups are generated from different distributions. The second setting will test for differences in distributions with the same means, but with different covariances that contain a non-zero covariance term between predictors with opposite signs. The third setting tests for differences distributions with different means, but the same covariances that contain a non-zero term for correlated predictors. The fourth setting is designed to test for differences in distributions with both different means and different covariances, where the covariances contain non-zero covariance terms with different signs. The fifth setting tests for differences in distributions with

the same means, but different covariances, where one distribution has a non-zero covariance term and one has a zero covariance term between predictors. The last parameter setting is aimed to test for differences in distributions with different means and different covariances where one distribution has a non-zero covariance term and one has a zero covariance term between predictors. By calculating power via simulations under each of these combinations of parameters, the effectiveness of these two methods can be evaluated. This will provide an understanding about when one method may be preferred over another and how well the methods do when testing for various types differences in distributions. For each combination of mean and covariance choices, sample sizes from $n = 5$ per group up to $n = 200$ per group are used.

Table 4.5: Simulation parameters to compare the performance of hypothesis testing based on the asymptotic null distribution of $f(\hat{B})$ and the LRT method are shown here. Column 1 denotes the number of the simulation setting. Columns 2 and 3 contain the means of the two distributions, while Columns 4 and 5 display the covariances of the two distributions. The different sample sizes that are used in the simulations are found in Column 6.

Parameter Setting	$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\Sigma_1}$	$\boldsymbol{\Sigma_2}$	n
1	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
2	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
3	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
4	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
5	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5 10 25 50 100 200
6	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5 10 25 50 100 200

For each of six sample sizes under each parameter setting, 10,000 data sets are simulated. For each data set, $f(\hat{B})$ is calculated and the type I error rate or power is calculated as the percentage of these values that is as least as big as the cutoff value from the $\Gamma(2.5, 1)$ distribution. Type I error and power of testing via the LRT statistics is calculated as the percentage of p-values less than 0.05 across 10,000 simulated data sets for each parameter setting.

Simulation Results.— Simulations are conducted to evaluate the effectiveness of both hypothesis testing based on the asymptotic distribution of $f(\hat{B})$ and testing via the LRT statistic. However, the LRT statistic and therefore resulting p-value are not able to be calculated in the case of perfect separation. Table 4.6 displays type I error rate calculations for both permutation testing and testing via the LRT statistic. Notice that * denotes p-values calculated from less than the total 10,000 data sets since perfect separation occurred. Both methods control the type I error rate well at the 0.05 level for sample sizes of 25 or more per group. However, the type I error rates are a little high for sample sizes of 5 and 10 per group. Also, the asymptotic test works for all data sets, while testing via the LRT statistic fails due to perfect separation in 21.6% of the data sets where n is 5 per group and in a little less than 1% of data sets where n is 10 per group.

Table 4.6: This table contains simulated type I error rates calculated from 10,000 data sets for testing based on the asymptotic null distribution of $f(\hat{B})$ and the LRT method for each of six parameter settings. Columns 1 and 2 display the mean and covariance of the distribution that both groups were simulated from. The sample sizes of the simulations are in Column 3 and Columns 4 and 5 contain the simulated type I error rates for testing based on the asymptotic null distribution of $f(\hat{B})$ and the LRT method, respectively.

$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	n	$f(\hat{B})$	LRT
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0.0964	0.0504*
		10	0.0692	0.0995*
		25	0.0562	0.0712
		50	0.0516	0.0560
		100	0.0542	0.0562
		200	0.0537	0.0532

The simulated power for both of these methods can be found in Table 4.7. Under every parameter setting, hypothesis testing based on the asymptotic distribution of $f(\hat{B})$ does as well as, or better than, the LRT method. The method based on $f(\hat{B})$ achieves at least 91% power under all settings with a sample size of 25 or greater and does so for even smaller sample sizes in a few of the cases. It is clear that the method based on $f(\hat{B})$ does extremely well in the cases where means are different, regardless of whether covariances are the same or different, but is not as good when the means of the two distributions are the same. Notice that * denotes power that was calculated from

less than the total 10,000 samples when perfect separation occurs and p-values are not able to be calculated. Perfect separation occurs under every parameter setting and at every sample size and actually occurred in 100% of simulations of size 200 per group under the setting of different means and different covariances. The percent of simulations that failed due to perfect separation can be seen in Figure 4.14. The black crosses, red X's, green squares, blue triangles, yellow diamonds, and orange dots represent the percent of 10,000 simulations with perfect separation for the LRT method for parameter settings 1, 2, 3, 4, 5, and 6 respectively. The solid line at 0 represents the percent of perfect separation for permutation testing based on $f(\hat{B})$. These values can also be found in the accompanying Table 4.8. This, compared with the higher power of the method based on $f(\hat{B})$, provides evidence that testing based on $f(\hat{B})$ is the preferred method and is adequate for sample sizes of 25 or more per group.

Table 4.7: This table contains simulated power calculated from 10,000 data sets for testing based on the asymptotic null distribution of $f(\hat{B})$ and the LRT method for each of five parameter settings. Columns 1 and 2 display the means of the two distributions, while Columns 3 and 4 display the covariances of the distributions. The sample sizes of the simulations are in Column 5 and Columns 6 and 7 contain the simulated power for testing based on the asymptotic null distribution of $f(\hat{B})$ and the LRT method, respectively. Cases where power was not calculated from all 10,000 data sets due to perfect separation are denoted by *.

$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\Sigma_1}$	$\boldsymbol{\Sigma_2}$	n	Power of $f(\hat{B})$	Power of LRT
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$	5	0.5509	0.1820*
				10	0.9573	0.9216*
				25	1	1*
				50	1	1*
				100	1	1*
				200	1	1*
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0.9327	0.6601*
				10	1	1*
				25	1	1*
				50	1	1*
				100	1	1*
				200	1	1*
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$	5	1	1*
				10	1	1*
				25	1	1*
				50	1	1*
				100	1	1*
				200	1	NA*
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0.1983	0.0667*
				10	0.4261	0.3189*
				25	0.9123	0.6335*
				50	0.9983	0.9098*
				100	1	0.9972*
				200	1	1
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0.988	0.7521*
				10	1	1*
				25	1	1*
				50	1	1*
				100	1	1*
				200	1	1*

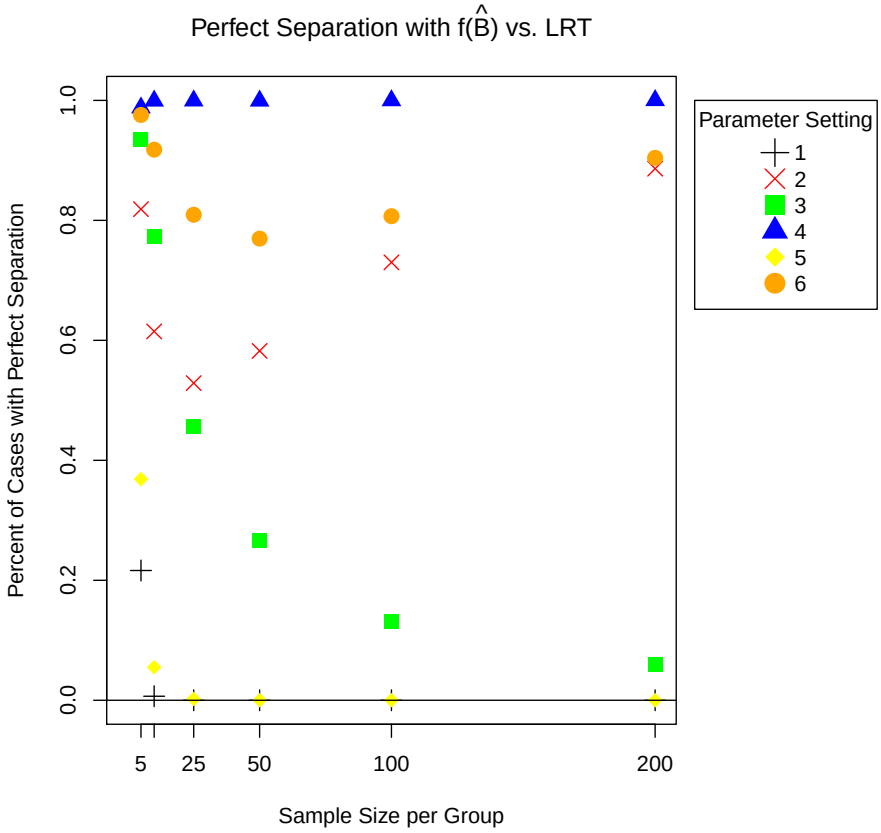


Figure 4.14: Perfect separation rates calculated from 10,000 data sets are plotted for each of the 6 sample sizes under each of six parameter settings. The black crosses, red X's, green squares, blue triangles, yellow diamonds, and orange dots represent the percent of 10,000 simulations with perfect separation for the LRT method for parameter settings 1, 2, 3, 4, 5, and 6 respectively. The solid line at 0 represents the percent of perfect separation for testing based on the asymptotic null distribution of $f(\hat{B})$.

Table 4.8: This table contains the percent of perfect separation from 10,000 simualted data sets for testing based on the asymptotic null distribution of $f(\hat{B})$ and the LRT method for each of six parameter settings. Columns 1 and 2 display the means of the two distributions, while Columns 3 and 4 display the covariances of the distributions. The sample sizes of the simulations are in Column 5 and Columns 6 and 7 contain the percent of 10,000 data sets that did not produce p-values due to perfect separation.

$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\Sigma_1}$	$\boldsymbol{\Sigma_2}$	n	Percent Perfect Separation of $f(\hat{B})$	Percent Perfect Separation of LRT
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0	0.2163
				10	0	0.0067
				25	0	0
				50	0	0
				100	0	0
				200	0	0
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$	5	0	0.8187
				10	0	0.6149
				25	0	0.5287
				50	0	0.5824
				100	0	0.7300
				200	0	0.8862
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0	0.9344
				10	0	0.7733
				25	0	0.4559
				50	0	0.2662
				100	0	0.1316
				200	0	0.0588
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$	5	0	0.9877
				10	0	0.9992
				25	0	0.9994
				50	0	0.9992
				100	0	0.9998
				200	0	1
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0	0.3687
				10	0	0.0550
				25	0	0.0022
				50	0	0.0001
				100	0	0.0001
				200	0	0
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0	0.9758
				10	0	0.9179
				25	0	0.8094
				50	0	0.7696
				100	0	0.8070
				200	0	0.9043

It is clear that hypothesis testing based on the asymptotic distribution of $f(\hat{B})$ is appropriate for large sample sizes. However, in the case of small sample sizes, i.e., n less than 25 samples per groups, although the type I error rates seem to be fine, the power is not high enough under all parameter settings. Thus, hypothesis testing can be done based on permutation testing methods.

Permutation Testing.— Permutation testing is a method that can be used to construct sampling null distributions, and thus empirically compute p-values. Like bootstrapping, a permutation test constructs, rather than assumes, a sampling distribution by resampling the observed data. Specifically, the observed data is shuffled or permuted by assigning different outcome values to each observation from the outcomes that are actually observed. Unlike bootstrapping, these permutations are done without replacement. Since the previous asymptotic result does not apply in the case of small sample sizes, permutation testing can be done instead to compute p-values.

Simulation studies are done to examine both the type I error rate and power for permutation testing with the sample B-distance. These results are also compared to hypothesis testing by means of a likelihood ratio test (LRT). The LRT test statistic is calculated by comparing the full model including an interaction term to the reduced model with just an intercept term. These models are evaluated with logistic regression and then compared using a LRT statistic.

Simulations.— Simulations are conducted to examine the effectiveness of both permutation testing with B-distance and testing via the LRT statistics as the sample sizes of the two groups, n_1 and n_2, increase. Effectiveness is evaluated by measuring both the type I error rate and the power of the two methods. Data are generated under six parameter settings incorporating different values for means and covariances. For each combination of mean and covariance choices, sample sizes from $n = 5$ per group up to $n = 200$ per group are used. The first setting is used to calculate type I error rates and thus data for the two response groups are generated from the same distribution. The remaining five combinations of parameter choices are chosen to evaluate the power of the two testing methods under various conditions of differences in distributions. These are the same settings used to compare testing with the asymptotic distribution of $f(\hat{B})$ and the LRT method. These settings are chosen to evaluate power under various conditions and will provide an understanding about when one method may be preferred over another and how well the methods do when testing for various types differences in distributions. These parameter settings can be found in Table 4.9.

Table 4.9: Simulation parameters to compare the performance of permutation testing based on $\hat{B}$ and the LRT method are shown here. Column 1 denotes the number of the simulation setting. Columns 2 and 3 contain the means of the two distributions, while Columns 4 and 5 display the covariances of the two distributions. The different sample sizes that are used in the simulations are found in Column 6.

Parameter Setting	$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\Sigma_1}$	$\boldsymbol{\Sigma_2}$	n
1	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
2	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
3	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
4	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$	5 10 25 50 100 200
5	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5 10 25 50 100 200
6	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5 10 25 50 100 200

For each of six sample sizes, 1,000 data sets are simulated in this manner. For each data set, $\hat{B}$ is calculated for each of 1000 permutations of the simulated data and a p-value is calculated from the percentage of permuted data sets that result in a $\hat{B}$ at least as big as the one from the original data set. Type I error is calculated as the percentage of p-values less than 0.05 across 1,000 simulation data sets for each parameter setting.

Simulation Results.— Simulations are conducted to evaluate the effectiveness of both permutation testing with B-distance and hypothesis testing via the LRT statistics. However, the LRT statistic and therefore resulting p-value are not able to be calculated in the case of perfect separation. Table 4.10 displays type I error rate calculations for both permutation testing and testing via the LRT statistic. Notice that * denotes p-values calculated without all 1,000 data sets since perfect separation occurred. Both methods control the type I error rate well at the 0.05 level. However, permutation testing works for all data sets, while testing via the LRT statistic fails due to perfect separation in 20% of the data sets where n is 5 per group and in less than 1% of data sets where n is 10 per group. These percentages of perfect separation can be found in Table 4.12. Therefore, permutation testing using B-distance is preferred over LRT testing.

Table 4.10: This table contains simulated type I error rates calculated from 10,000 data sets for permutation testing and the LRT method for each of four parameter settings. Columns 1 and 2 display the mean and covariance of the distribution that both groups were simulated from. The sample sizes of the simulations are in Column 3 and Columns 4 and 5 contain the simulated type I error rates for permutation testing and the LRT method, respectively.

$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	n	Permutation Testing	LRT
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0.052	0.040*
		10	0.054	0.096*
		25	0.053	0.078
		50	0.044	0.067
		100	0.038	0.06
		200	0.057	0.059

Table 4.11 displays power calculations for both permutation testing and testing via the LRT statistic. Notice that * denotes p-values calculated without all 1,000 data sets since perfect separation occurred.

Table 4.11: This table contains simulated power calculated from 10,000 data sets for permutation testing based $\hat{B}$ and the LRT method for each of five parameter settings. Columns 1 and 2 display the means of the two distributions, while Columns 3 and 4 display the covariances of the distributions. The sample sizes of the simulations are in Column 5 and Columns 6 and 7 contain the simulated power for testing based on the asymptotic null distribution of $f(\hat{B})$ and the LRT method, respectively. Cases where power was not calculated from all 1,000 data sets due to perfect separation are denoted by *.

$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\Sigma_1}$	$\boldsymbol{\Sigma_2}$	n	Power of Permutation Testing	Power of LRT
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$	5	0.292	0.146*
				10	0.896	0.926*
				25	1	1*
				50	1	1*
				100	1	1*
				200	1	1*
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0.880	0.703*
				10	1	1*
				25	1	1*
				50	1	1*
				100	1	1*
				200	1	1*
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$	5	0.997	1*
				10	1	NA*
				25	1	NA*
				50	1	1*
				100	1	NA*
				200	1	NA*
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0.096	0.078*
				10	0.328	0.318*
				25	0.879	0.640*
				50	1	0.904*
				100	1	0.999
				200	1	1
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0.958	0.684*
				10	1	1*
				25	1	1*
				50	1	1*
				100	1	1*
				200	1	1*

Just as before when the power of the test based on the asyptotic null distribution of $f(\hat{B})$ is compared with the LRT method, the permutation testing based on $\hat{B}$ does as well as or better than the LRT method in every single case but one. The only time LRT performs better in these simulations is under the setting of distributions with the same means, but different covariances of opposite signs for the covariance term at the sample size of 10. Otherwise, permutation testing outperforms LRT, especially since, once again, p-values cannot be computed for many of the LRT simulations due to perfect separation. Power seems to be high across all parameter settings for sample sizes of 10 or more per group. However, both methods lack adequate power at sample sizes of 5 per group under both of the settings where means are the same, but covariances are different. This is more evidence that not only is B-distance useful for identifying combinations of predictors that are important, but it is useful for identifying main effects that logistic regression often cannot due to perfect separation.

The percentage of simulations where perfect separation occurs in plotted in 4.15. The black crosses, red X's, green squares, blue triangles, yellow diamonds, and orange dots represent the percent of 10,000 simulations with perfect separation for the LRT method for parameter settings 1, 2, 3, 4, 5, and 6 respectively. The solid line at 0 represents the percent of perfect separation for permutation testing based on $\hat{B}$. Table 4.12 also displays this information on the percentage of p-values that are not calculated due to perfect separation. Clearly this is a problem when attempting to use the LRT method for hypothesis testing and occurs in as many as 100% of cases for these simulations. Therefore, permutation testing with B-distance is very obviously the preferred method due to its ability to analyze all data sets, as well as its high power to detect true differences. This is evidence that permutation testing is an appropriate method for performing hypothesis testing for samples of size 10 or greater. Note that although power is not very high for permutation in sample sizes of 5 per group, the type I error rate is still controlled. Therefore, it would not be inappropriate to use permutation testing in the case of these small sizes, but the testing method is has lower power for detecting differences due to covariances when means are the same compared to differences in means. However, even in this case, permutation testing outperforms LRT.

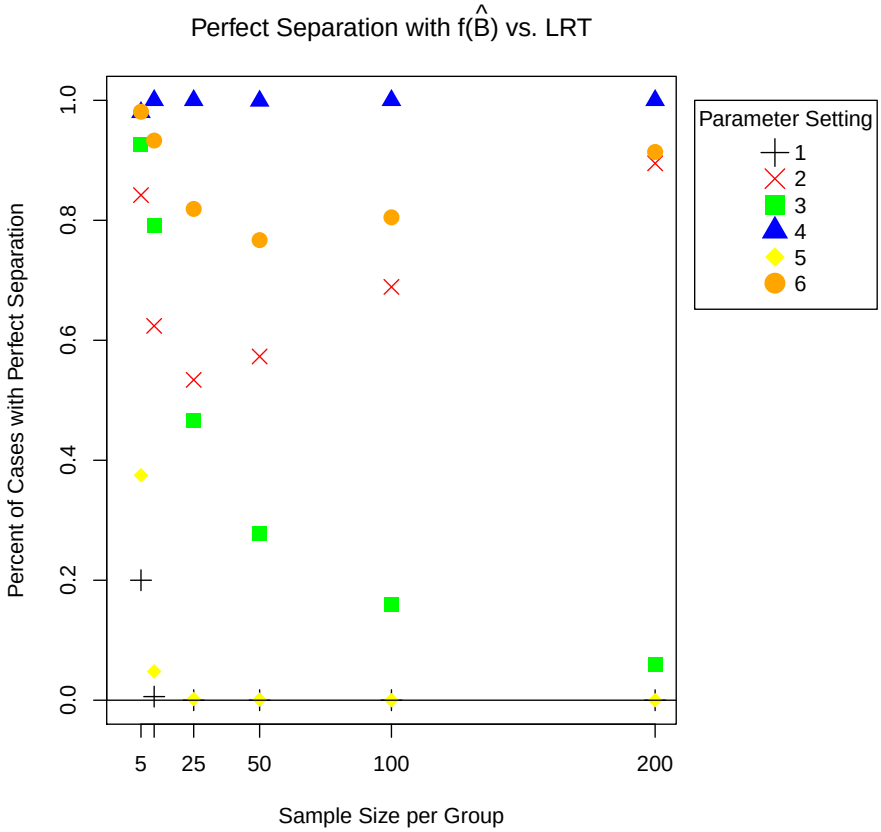


Figure 4.15: Perfect separation rates calculated from 10,000 data sets are plotted for each of the 6 sample sizes under each of six parameter settings. The black crosses, red X's, green squares, blue triangles, yellow diamonds, and orange dots represent the percent of 10,000 simulations with perfect separation for the LRT method for parameter settings 1, 2, 3, 4, 5, and 6 respectively. The solid line at 0 represents the percent of perfect separation for permutation testing based on $\hat{B}$.

Table 4.12: This table contains the percent of perfect separation from 10,000 simulated data sets for permutation testing based $\hat{B}$ and the LRT method for each of six parameter settings. Columns 1 and 2 display the means of the two distributions, while Columns 3 and 4 display the covariances of the distributions. The sample sizes of the simulations are in Column 5 and Columns 6 and 7 contain the percent of 10,000 data sets that did not produce p-values due to perfect separation.

$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\Sigma_1}$	$\boldsymbol{\Sigma_2}$	n	Percent Perfect Separation of Permutation Testing	Percent Perfect Separation of LRT
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0	0.200
				10	0	0.006
				25	0	0
				50	0	0
				100	0	0
				200	0	0
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0	0.842
				10	0	0.624
				25	0	0.534
				50	0	0.573
				100	0	0.689
				200	0	0.895
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0	0.926
				10	0	0.791
				25	0	0.466
				50	0	0.278
				100	0	0.159
				200	0	0.06
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0	0.980
				10	0	1
				25	0	1
				50	0	0.999
				100	0	1
				200	0	1
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0	0.375
				10	0	0.048
				25	0	0.002
				50	0	0.001
				100	0	0
				200	0	0
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0	0.981
				10	0	0.933
				25	0	0.819
				50	0	0.767
				100	0	0.805
				200	0	0.914

Now that some methods have been developed to perform hypothesis testing and create confidence intervals, inference can be made based on $\hat{B}$. Thus, a real data analysis is performed to demonstrate the validity and usefulness of these methods.

4.2 Bhattacharyya Distance Real Data Analysis

Data come from 8 salamanders, each with 20,035 normalized gene expression measurements. A total of 4 salamanders are in the control group and the other 4 salamanders are allocated to the treatment group. Here, the treatment is the application of a chemical that inhibits tail regeneration. The goal is to identify differentially expressed genes that are related to regeneration. Gene IDs were previously annotated based on BLAST searches.

One approach to identify potentially interesting genes between groups is to perform 20,035 t-tests, one for each gene individually. Taking this approach, there are 1,966 genes that are significant at the 0.01 alpha level. However, solely performing t-tests has several limitations. Ideally, I would like to use gene expressions to predict the outcome of regeneration and not vice versa. However, by using t-tests, the presence or absence of regeneration is not treated as the response, which is the main goal of the experiment. Secondly, t-tests can only provide results about how a single gene expression is related to the outcome of regeneration. More information can be gained by considering combinations of genes that predict regeneration. To do this, logistic regression can be performed. However, it is now the case of small sample sizes, and perfect separation occurs quite often just by chance with this limitation on sample sizes. In fact, perfect separation occurs in more than 70% out of the total 200,690,595 possible combinations of genes. When perfect separation occurs, an error message is produced like the one in 4.16.

```
Call:  glm(formula = group ~ axo00002.f_at * axo00012.r_at, family = binomial,
    data = dat_orig)

Coefficients:
                (Intercept)                 axo00002.f_at
                  6.314e+03                    -4.368e-01
              axo00012.r_at  axo00002.f_at:axo00012.r_at
                 -4.437e-01                     3.057e-05

Degrees of Freedom: 7 Total (i.e. Null);  4 Residual
Null Deviance:      11.09
Residual Deviance: 4.801e-10    AIC: 8
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Figure 4.16: When perfect separation occurs, logistic regression fails and statistical software programs, such as R, will produce error messages. This is the error message produced by R when trying to fit a logistic regression model to data with perfect separation.

Plotting the data that produced this error in Figure 4.17 displays that perfect separation exists between the subjects with regeneration from those without regeneration, represented here by a dashed line.

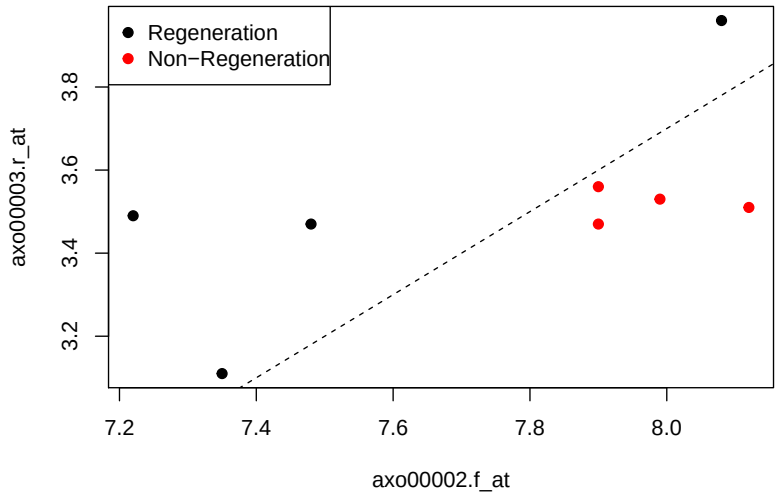


Figure 4.17: This plots shows the relationship between two predictors and the response groups. Black dots denote observations from the regeneration group and red dots denote observations from the non-regeneration group. In this case, the predictors can perfectly separate observations in the regeneration group from the non-regeneration group (by the dashed line) and logistic regression fails. It is clear, though, that this combination of variables could be useful in determining group membership.

In order to address this issue of perfect separation, B-distance is used to analyze this data. FSA is run with B-distance as the optimization criterion in search for two-way interactions. Based on the lower bound identified in my previous work, number of iterations of the algorithm is chosen to obtain the statistically optimal solution with at least 90% probability. That is, in order to obtain c, where c is the percentage of predictors that is used as the number of random starts, the following equation is solved:

$$1 - e^{-c2^2} = 0.90$$

$$\iff c = 0.5756463$$

Since there are $p = 20,035$ predictors, the number of random starts needed is $0.58(20035) =$

11, 533.07, or 11,534 random starts. Due to the large number of iterations of the algorithm, the analysis is paralellized to run on a computer cluster. Based on the method chosen to parallelize this analysis, there are actually 11,840 random starts conducted, which is greater than the necessary 11,534 iterations found from the lower bound formula.

In total, there are 950 feasible solutions identified by FSA. A subset of these feasible solutions can be seen in Table 4.13. A table of the full results can be found in the appendix in Tables A.1 - A.22. There are 791 solutions containing genes that are not identified by t-tests at the $\alpha = 0.01$ significance level. These feasible solutions provide an additional genes identified as some solutions contained one additional gene, while others contained two genes not previously identified through t-test analysis. FSA also identified the highest non-infinity B-distance value and in fact, it is the solution chosen most often. A plot of this feasible solution can be seen in Figure 4.18.

Table 4.13: FSA produced 950 feasible solutions and a subset of those are shown here, including the statistically optimal solution denoted in bold with $\hat{B} = 6627.62$ (Column 3). Columns 1 and 2 show the probes that are identified in each of the models. Columns 3 and 4 display the sample B-distance associated with each model and the number of times each solution was chose by FSA, respectively.

Variable 1	Variable 2	B-distance	Time Chosen by FSA
axo00315.f-at	**axo15507.f-at**	**6627.619917**	**2400**
axo09358.f-at	axo24943.f-at	4261.718149	3
axo04944.f-at	axo26845.f-at	3125.810796	215
axo15772.f-at	axo20526.f-at	2828.213655	4
axo17411.r-at	axo19789.f-at	2463.550339	5
axo02376.f-at	axo11058.f-at	2096.440067	66
axo02994.f-at	axo04573.f-at	2077.560286	3
axo18985.f-at	axo27182.f-at	1928.560211	26
axo15122.f-at	axo22646.r-at	1920.598369	4
axo16951.r-at	axo18028.f-at	1789.161704	4

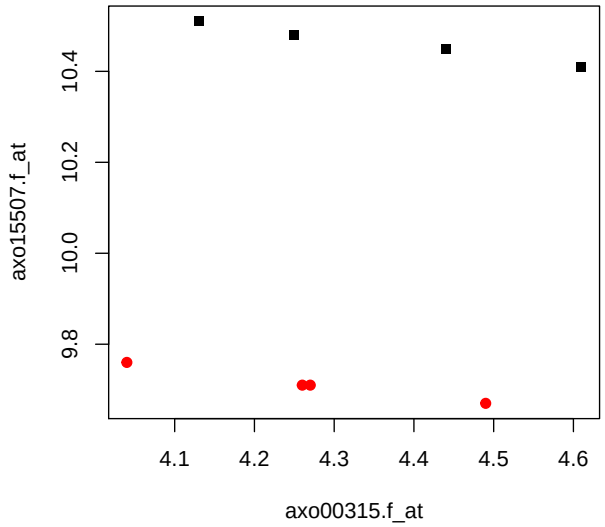


Figure 4.18: A plot of the combination of predictors that resulted in the largest B-distance is seen here. Black dots denote observations from the regeneration group and red dots denote observations from the non-regeneration group. Notice that the two response groups can be perfectly separated by these two predictors. This is a case where logistic regression would fail, but B-distance can be calculated.

After running FSA, exhaustive search is also conducted in order to learn more about B-distance and its ability to detect interactions. B-distance is calculated for all 200,690,595 possible combinations of predictors. Upon plotting a histogram of these distances, those that have a B-distance greater than 380 are examined further. This resulted in 340 out of the total 200,690,595 possible pairs of gene expressions being chosen. P-values are calculated via permutation testing for the 340 combinations of genes that do not have a B-distance of infinity. Permutation testing was performed by calculating B-distance for each of 35 combinations of ways to choose four subjects per group. There are actually 70 ways to choose four subjects per group from the total of eight subjects. However, because these are symmetric, i.e. having subjects 1, 2, 3, and 4 in the regeneration group will give the same B-distance as having these subjects in the non-regeneration group with the others being in the regeneration group, I only looked at the 35 ways in which these groups can differ. By shuffling the individuals in this way, I can look at what percentage of the time a B-distance as large as that in the original data is seen. This provides an idea of whether the distance seen is actually large or not. Since there are 35 distances calculated, one being the original distance, the smallest p-value that can be achieved is $\frac{1}{35}$ or 0.0286. Thus, a combination of genes is considered significant if the permuted p-value for that combination is 0.0286.

Out of the 340 gene combinations that permutation p-values are calculate for, all are considered significant, resulting in 340 combinations of genes. To consider whether a gene should be added to the list of potentially interesting genes, it is useful to compare the genes identified here to those genes identified by the univariate t-tests. Consider the histogram in Figure 4.19. This is a histogram of the t-test p-values for the genes that are significant as identified by permutation testing with B-distance values. It is clear that many of these genes have a t-test p-value less than 0.01. However, there are still many genes identified by permutation testing that are not identified by the t-test method.

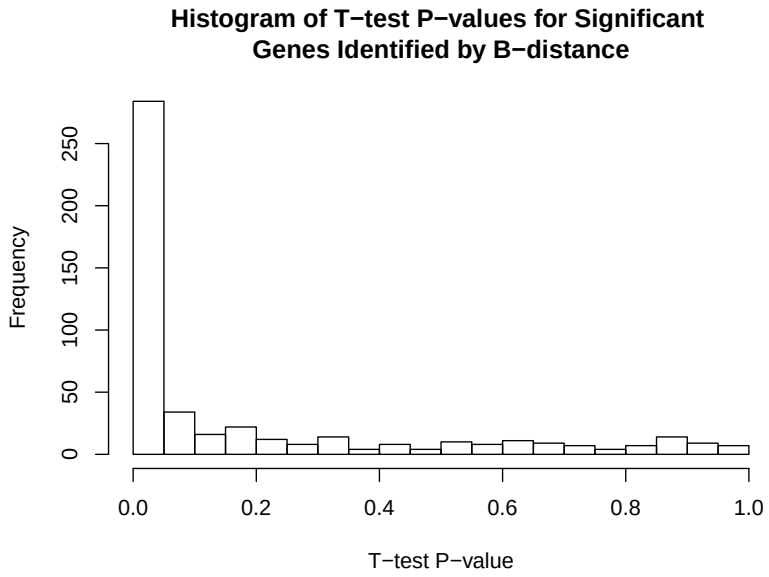


Figure 4.19: The p-values for univariate t-tests are plotted here for the values of B-distance greater than 380 that are significant based on permutation testing. Although a large amount of these p-values are less than 0.01, there are many p-values that are not significant based on t-tests, but that B-distance is able to identify as significant.

In fact, out of the 340 combinations that are significant by permutation testing, 4.12% or 14 of these have both t-test p-values greater than 0.01. These 14 combinations are considered the most interesting and should be on the differentially expressed gene list. Also note, that of these 14 interesting combinations of predictors, 13 are also identified by FSA with B-distance as the criterion function. Two of these interesting combinations of predictors can be seen in Figures 4.20 and 4.21. The remaining interesting combinations of predictors can be found in the Appendix in Figures A.1 - A.4.

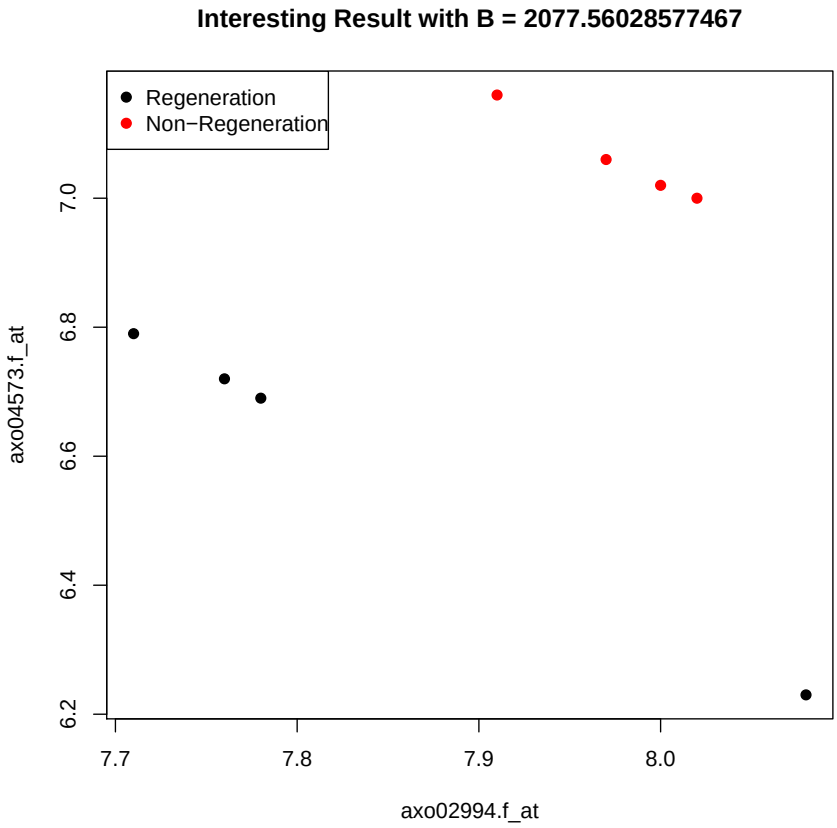


Figure 4.20: A plot of a combination of predictors that resulted in a significant permutation p-value from B-distance is seen here. Black dots denote observations from the regeneration group and red dots denote observations from the non-regeneration group. Notice that the two response groups can be perfectly separated by these two predictors. This is a case where logistic regression would fail, but B-distance can be calculated. Univariate t-tests would also fail to identify either of these predictors as significant at the 0.01 level, but clearly when combined, the two predictors provide valuable information about which observations correspond to regeneration or non-regeneration.

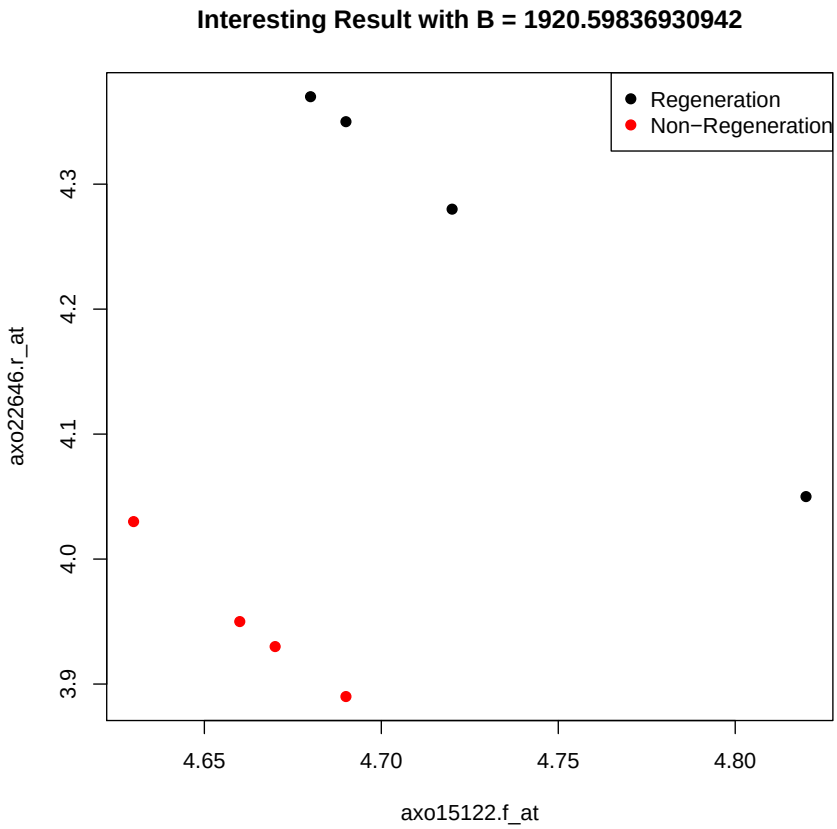


Figure 4.21: A plot of another combination of predictors that resulted in a significant permutation p-value from B-distance is seen here. Black dots denote observations from the regeneration group and red dots denote observations from the non-regeneration group. Notice that the two response groups can be perfectly separated by these two predictors. This is a case where logistic regression would fail, but B-distance can be calculated. Univariate t-tests would also fail to identify either of these predictors as significant at the 0.01 level, but clearly when combined, the two predictors provide valuable information about which observations correspond to regeneration or non-regeneration.

These results are very interesting and provide insight to the fact that B-distance can be used as a simple tool for exploratory data analysis. Logistic regression fails in every single one of the most interesting cases documented in these results. B-distance is able to be calculated in all of these cases and is also much faster than logistic regression. B-distance also seems to be better at identifying main effects than interaction effects. However, this is still useful in the case of perfect separation, since logistic regression fails and those main effects would not be discovered.

As a brief note, I will discuss the removal of combinations with a B-distance of infinity. This occurs for two reasons. The first is if there is no variability between the points in one or both groups, which is uninteresting. The second case is when the two predictors are perfectly correlated for one or both groups. This causes $|\hat{\mathbf{\Sigma}}_\mathbf{1}|$ or $|\hat{\mathbf{\Sigma}}_\mathbf{2}|$ to be equal to zero, and thus produces a B-distance of infinity.

Examples of both of these cases can be seen in Figures 4.22 and 4.23.

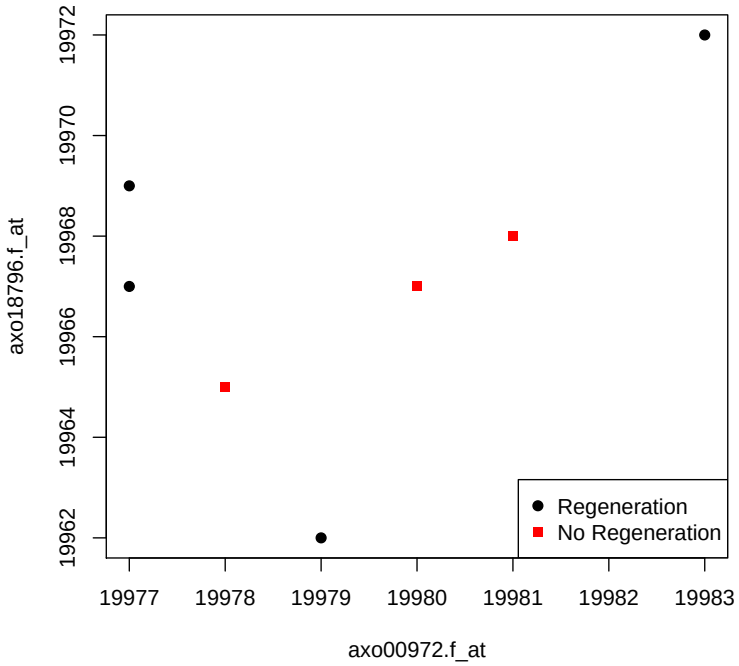


Figure 4.22: A combination of predictors that resulted in a sample B-distance of infinity is seen here. Black dots denote observations from the regeneration group and red dots denote observations from the non-regeneration group. The non-regeneration group has perfect linearity, which produces a B-distance of infinity. Thus, inference cannot be made in this case.

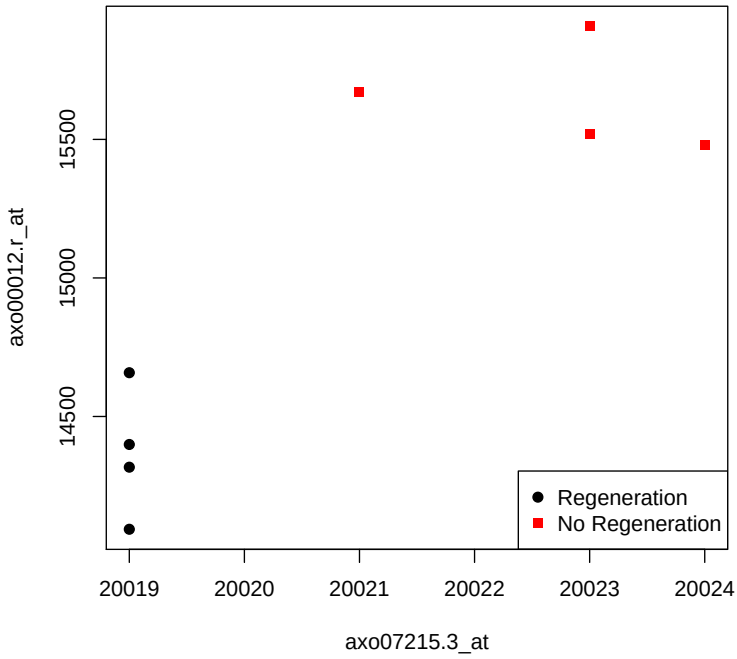


Figure 4.23: A combination of predictors that resulted in a sample B-distance of infinity is seen here. Black dots denote observations from the regeneration group and red dots denote observations from the non-regeneration group. The regeneration group has no variability in axo7215.3_at, which produces a B-distance of infinity. This case is not interesting and is removed from consideration.

This real data analysis of gene expression data from salamanders displays the validity and usefulness of using B-distance to make inference about combinations of predictors that are related to the response of regeneration. B-distance is used here to find interesting results that cannot be found through univariate t-tests or with logistic regression.

Chapter 5

Discussion and Future Directions

Many methods exist for modeling data with large p. However, these methods are often limited due to computational complexity, inability to model interactions effectively, and inflexibility of both input and output. FSA is a stochastic algorithm introduced to address some of these concerns. However, without statistical guidance for the number of iterations to run, users are required to arbitrarily choose this value. Thus, I have provided a lower bound on the probability of identifying the statistically optimal solution that can guide users to make an appropriate choice aimed at balancing the computational efficiency with the probability of attaining the statistically optimal solution. Thus, users are likely more likely to identify this statistically optimal solution, along with other potentially biologically meaningful solutions. Theoretical properties about B-distance have also been developed here in order to address the severe limitation of logistic regression when dealing with predictors that can perfectly separate the response groups. B-distance is proposed as an alternative to logistic regression due to its existence even when perfect separation occurs. However, little theory exists about B-distance and thus no information about making inference from $\hat{B}$ exists. These limitations are addressed and properties are derived in order to provide insight into making inferences about the difference in distribution of the sample groups with $\hat{B}$.

5.1 FSA

Although FSA addresses limitations of existing modeling techniques, little is known about its theoretical properties. To address one aspect of this limitation, I have provided a way to determine the number of iterations of FSA needed to obtain the statistically optimal solution of an m-way interaction model with a certain probability. For example, when considering a two-way interaction model, if one would like the probability of obtaining the statistically optimal solution to be at least 80%, then the number of random starts of FSA needs to be chosen to be 40% of the number of possible explanatory variables in the data set of interest. This lower bound on the probability of obtaining the statistically optimal solution can be easily implemented by data analysts running FSA and will be incorporated into the currently available Shiny application for FSA in the near future. Further, simulation study and real data analysis demonstrate the validity and usefulness of this lower bound.

The work here provides a foundation for further study of theoretical properties of FSA. For instance, the simulation study results show that the derived lower bound is conservative. Thus, in future work, tightening the lower bound would increase the computational efficiency of FSA. However, in this case, the conservative lower bound does provide statistical guidance to FSA users. In addition, little is known about how conservative this bound is in the presence of smaller effect sizes. Studying this will increase the impact of this work by providing more specific guidance to the data analyst. Knowing how to choose the number of times to run FSA will improve the computational usability of FSA by allowing the user to choose fewer random starts based on the desired likelihood of obtaining the statistically optimal solution while still being computationally feasible, and continue providing a valid alternative to exhaustive search and other model selection methods.

5.2 Bhattacharyya Distance

Data analysis for the cases of big p and small n are hindered due not only to computational concerns, but also issues that arise when predictors can perfectly separate the groups in binary response data, which happens often just by chance when dealing with small sample sizes. B-distance is proposed here to address these limitations. Although B-distance is used often in various feature selection and extraction methods, little is known about its theoretical properties outside of its relationship to the Bayes probability error of classification. This work provides information about the distribution of the sample B-distance under various assumptions.

Currently, there is limited knowledge about properties of $\hat{B}$. There exists no theory about the maximum likelihood estimate of the true B-distance between two distributions, B_0. Maximum likelihood estimation is a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters. Estimators derived by the method of maximum likelihood have some desirable properties. This estimator is referred to as the maximum likelihood estimator (MLE). Thus, if $\hat{B}$ is the MLE of the true B-distance, then these known properties about MLEs can be used.

For example, under some regularity conditions, it is known that for an MLE, $\hat{\theta}_n$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{p} \mathcal{N}(0, \frac{1}{I(\theta_0)})$$

where $\hat{\theta}_n$ is the consistent root of the likelihood equation and $I(\theta_0)$ is the Fisher information. Thus,

if $\hat{B}$ is the MLE of B_0, then $\hat{B}$ is expected to converge in distribution to a normal distribution. That is, it is expected that

$$\sqrt{n}(\hat{B}_n - B_0) \xrightarrow{p} \mathcal{N}(0, \frac{1}{I(B_0)})$$

However, currently the likelihood function has not been calculated. Thus, simulations are conducted to examine the distribution of $\sqrt{n}(\hat{B}_n - B_0)$.

Table 5.1: Simulation parameters to evaluate the convergence of $\sqrt{n}(\hat{B}_n - B_0)$ to a normal distribution are shown here. Columns 1 and 2 contain the mean and covariance the two distributions. The different sample sizes that are used in the simulations are found in Column 3.

$\boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \boldsymbol{\mu}$	$\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma}$	n
$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$	50 100 500 1000 2000

Simulations are conducted under the parameter settings in Table 5.1 to examine the asymptotic convergence of $\sqrt{n}(\hat{B}_n - B_0)$ to a normal distribution as the sample sizes of the two groups, n_1 and n_2, increase. For each data set, $\hat{B}$ is calculated and histograms are plotted. Preliminary plots in 5.1 show results from 10,000 simulated data sets at each of 5 sample sizes ranging from 50 per group up to 2,000 per group and indicate that $\hat{B}$ may not be the MLE. In Figure 5.1 where data for the two response groups are simulated from the same distribution, it does not look as though the simulated quantities of $\sqrt{n}(\hat{B}_n - B_0)$ converge to a normal distribution, even as the sample size increases. More work needs to be done in order to determine if this relationship holds or not.

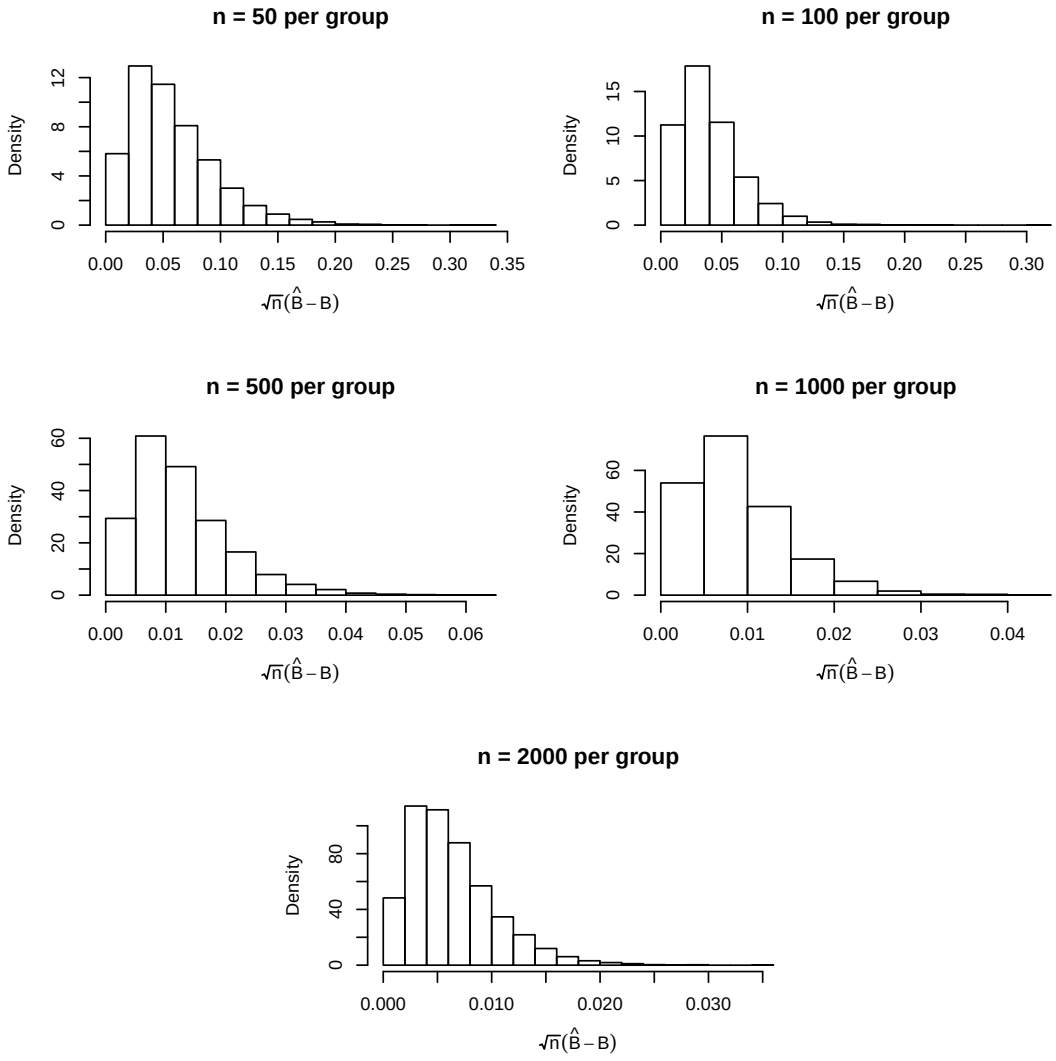


Figure 5.1: The quantity $\sqrt{n}(\hat{B_n} - B_0)$ is calculated and plotted for each of 10,000 data sets at each of 5 sample sizes ranging from 5 per group up to 100 per group. Data are simulated from a distribution with zero mean and a covariance matrix with a non-zero term for covariance between the two predictors. This is done to look for evidence of convergence of $\sqrt{n}(\hat{B_n} - B_0)$ to a normal distribution.

Prior to this work, no statistical methods exists on how to make inference from $\hat{B}$. Thus, I have derived the asymptotic null distribution that can be used for hypothesis testing with adequate sample sizes, i.e. with group sizes larger than approximately 25, and have also provided a way to conduct hypothesis testing via permutation testing when the group sample sizes are not large, i.e. less than 25 samples per group. Both of these hypothesis testing methods are compared to testing via an LRT statistic and show similar type I error rates and power. However, the LRT method fails in the case of perfect separation, while B-distance is not limited in this case. Percentile intervals are also used

in order to create confidence intervals for estimates of $\hat{B}$. Simulation study and real data analysis demonstrate the validity and usefulness of B-distance for identifying combinations of variables that are important, especially in the case of small sample sizes.

In order to utilize B-distance in the case of small sample sizes, researchers need an idea about what values of B-distance constitute a large value. Since much small sample theory has yet to be developed for the use of B-distance, it would be beneficial to determine a cutoff value that enables users to determine what values of sample B-distance are considered to be large. In order to do this, one could start by examining the percent of data sets containing perfect separation just due to chance, i.e. data sets simulated from the same distributions where perfect separation is not expected to occur. Then, a cutoff value could potentially be determined by examining all possible permutations of the outcomes of the data sets and the associated sample B-distance values. This would provide insight into what values of sample B-distance are likely to be seen when no true relationship exists between a combination of predictors and the group outcome in the case of small sample sizes. Thus, a cutoff value may be determined for identifying large B-distance values in this setting.

The work done in Chapter 4 provides insight into developing further theoretical properties about B-distance, as well as understanding its usefulness in real data analysis. In the future, it would be beneficial to develop a better understanding of how to interpret the relationship between predictors identified by B-distance with the response variables. For example, even though I chose B-distance in order to incorporate differences in shape and direction through the covariance terms, it still seems that B-distance is influenced much more by the term that incorporates the locations or means of the two response groups than the term that incorporates the covariances of the two sample groups. For example, consider Figures 5.2 and 5.3. Both represent combinations of variables that I would like to identify through the use of B-distance. Each group contains simulated data from 100 samples. However, B-distance will typically identify the differences displayed in Figure 5.3 over those displayed in Figure 5.2.

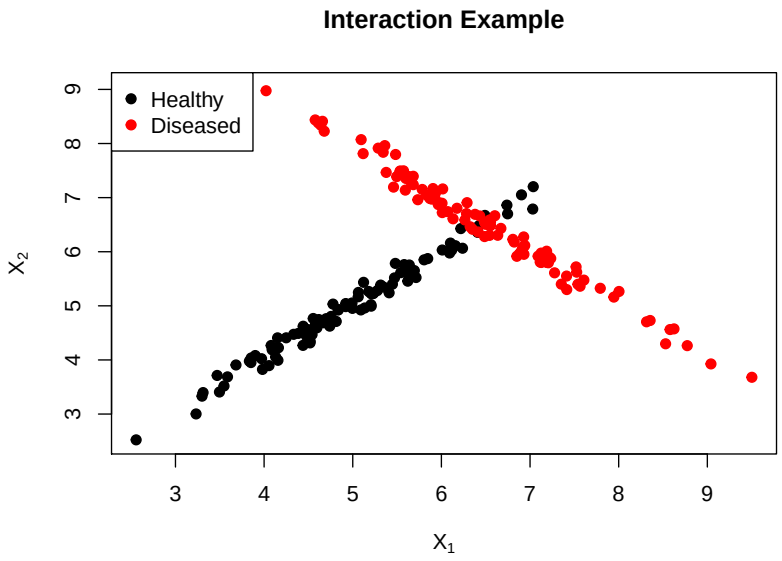


Figure 5.2: This plot represents a relationship B-distance aims to identify. Black dots denote observations from one response group and red dots denote observations from the other response group. Data are simulated from distributions with different means and covariances. Although B-distance is proposed to identify interaction effects like the one seen here, it does a better job of identifying the effects displayed in Figure 5.3.

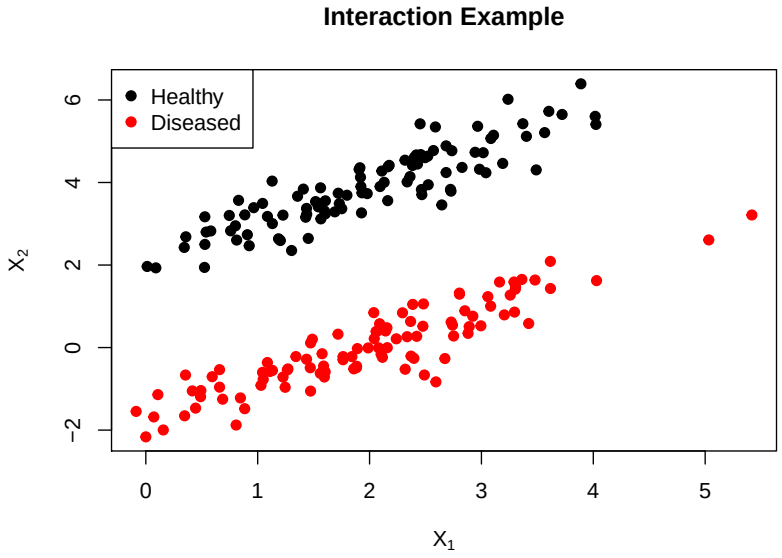


Figure 5.3: This plot represents a relationship B-distance aims to identify. Black dots denote observations from one response group and red dots denote observations from the other response group. Data are simulated from distributions with different means, but the same covariances. B-distance is proposed to identify effects like the one seen here and simulations show that B-distance is able to identify the effects displayed here.

Also, B-distance has only been considered thus far for identifying two-way interactions. It would be very interesting and potentially useful to examine the effectiveness of B-distance in identifying three-way interactions. In order to test for three-way interactions, the multivariate normal distributions of the response groups would increase in dimension to include a third predictor. By increasing the dimension of the distributions under consideration, the bias of $\hat{B}$ may also increase, especially in small samples. Further study is needed to explore additional theoretical properties of B-distance, but it is clear that it is helpful in exploratory data analysis.

5.3 Summary

The work done here provides insight into methods for modeling interaction effects in big data. FSA can be a useful tool for identifying these interactions, but without guidance on the number of iterations to run, users are uninformed about how this choice relates not only to the probability of identifying the statistically optimal solution, but also about how it relates to the computational efficiency of the algorithm. I have addressed this limitation by providing users with statistical guidance on how to appropriately choose the number of iterations of the algorithm to use in order to balance computational time with the probability of identifying the statistically optimal solution. Other issues that arise in big data come from the presence of perfect separation in data with a binary outcome. Thus, properties of B-distance and guidance on how to make inference about the sample B-distance have been developed. These methods for hypothesis testing and interval estimates are not only faster than traditional logistic regression methods, but also address the severe limitation of logistic regression in the case of perfect separation by providing an alternate method to analyze data.

Appendix

A.1 Tables

Table A.1: This table shows all 33 feasible solutions identified by FSA. Columns 1 and 2 show the SNPs that are identified, column 3 shows how many times each feasible solution was chosen by FSA, and column 4 shows the R^2 value associated with each model.

Variable 1	Variable 2	Times Chosen by FSA	R^2
mb104327194	mb91638370	42	0.0957401
mb13136127	mb31255782	898	0.1245719
mb28636979	**mb87344525**	**107**	**0.1256308**
mb111935889	mb43233761	25	0.1065257
mb62443411	mb99541026	23	0.1088855
mb112250554	mb96331482	56	0.1123864
mb14715054	mb19242337	66	0.1118795
mb112608319	mb96222909	144	0.113829
mb30780223	mb34260455	155	0.1142272
mb29078498	mb69000296	198	0.1229996
mb36825925	mb85205875	62	0.1146195
mb34871920	mb42993313	101	0.116866
mb14701689	mb37035520	95	0.1125078
mb110752558	mb15269553	25	0.09883659
mb107124592	mb57686005	26	0.1091593
mb107672282	mb56850293	22	0.1048774
mb14796254	mb19293075	21	0.1123462
mb69375025	mb93036659	15	0.1155624
mb104477633	mb91521906	18	0.09528844
mb61386637	mb99542699	20	0.1060636
mb29652874	mb74632475	11	0.09989297
mb72262602	mb92721283	13	0.111981
mb114381537	mb18255346	32	0.09999336
mb114239715	mb35229230	11	0.09914173
mb67397357	mb75739738	3	0.08240846
mb72256709	mb92533218	60	0.1166223
mb12221288	mb31300335	12	0.1047527
mb64403722	mb75460557	4	0.09193129
mb31343516	mb4585952	3	0.1055288
mb114382061	mb18266042	5	0.09632044
mb104707625	mb96032784	8	0.08665745
mb61425863	mb99527427	1	0.1050678
mb58406490	mb75674799	2	0.08783811

Table A.2: Type I error rates from 10,000 simulations for the comparison of the asymptotic null distribution of $f(\hat{B})$ to a $\Gamma(2.5, 1)$ are found here. Columns 1 and 2 display the means and covariances for the simulated data. Column 3 shows the sample sizes used for each simulation. Columns 4, 5, and 6 contain the simulated type I error rates at each of 0.025, 0.05, and 0.10 significance levels. This table corresponds the values plotted in Figure 4.13.

$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	n	α		
			0.025	0.05	0.10
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0.0432	0.097	0.191
		10	0.0334	0.0653	0.1354
		25	0.0278	0.0536	0.1066
		50	0.0262	0.0531	0.1075
		100	0.0256	0.051	0.1022
$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0.0455	0.0958	0.1953
		10	0.0358	0.0724	0.134
		25	0.0292	0.0574	0.1107
		50	0.0270	0.0520	0.1037
		100	0.0257	0.0516	0.1053
$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	5	0.0436	0.0969	0.1898
		10	0.0313	0.0631	0.1305
		25	0.0266	0.0544	0.1106
		50	0.0265	0.0529	0.1055
		100	0.0238	0.0489	0.103
$\begin{pmatrix}3\\3\end{pmatrix}$	$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$	5	0.0416	0.0955	0.1920
		10	0.0343	0.0677	0.1354
		25	0.0285	0.0577	0.1107
		50	0.0289	0.0551	0.1069
		100	0.0275	0.0546	0.1025

Table A.3: This table shows all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo00315.f-at	axo15507.f-at	6627.619917	2400
axo09358.f-at	axo24943.f-at	4261.718149	3
axo04944.f-at	axo26845.f-at	3125.810796	215
axo15772.f-at	axo20526.f-at	2828.213655	4
axo17411.r-at	axo19789.f-at	2463.550339	5
axo02376.f-at	axo11058.f-at	2096.440067	66
axo02994.f-at	axo04573.f-at	2077.560286	3
axo18985.f-at	axo27182.f-at	1928.560211	26
axo15122.f-at	axo22646.r-at	1920.598369	4
axo16951.r-at	axo18028.f-at	1789.161704	4
axo07342.r-at	axo12827.r-at	1696.683819	2
axo05747.f-at	axo12450.f-at	1589.491331	384
axo13479.f-at	axo16605.f-at	1582.921084	6
axo11927.f-at	axo16754.f-at	1492.07273	5
axo05548.f-at	axo19830.f-at	1430.534091	27
axo00347.f-at	axo15658.f-at	1397.856629	9
axo14778.f-at	axo31327.f-at	1218.22224	133
axo17616.f-at	axo22207.f-at	1154.60174	16
axo18225.f-at	axo28493.f-at	1149.687926	47
axo14055.r-at	axo14221.f-at	1127.776432	8
axo00845.f-at	axo01103.f-at	1081.040615	45
axo21651.r-at	axo25814.f-at	1065.595366	19
axo07174.f-at	axo26078.f-at	1045.552458	73
axo07245.r-at	axo24684.f-at	1004.032474	17
axo10317.f-at	axo15981.f-at	977.8579731	109
axo24285.f-at	axo27801.f-at	961.1020155	1671
axo12622.f-at	axo25858.f-at	919.618774	7
axo25315.f-at	axo25631.f-at	910.1378696	33
axo03076.f-at	axo11978.f-at	899.1791324	3
axo14168.f-at	axo25142.f-at	893.1647235	21
axo12874.r-at	axo25932.f-at	875.2573376	20
axo10070.f-at	axo28711.f-at	872.0978626	5
axo03422.f-at	axo15699.f-at	868.0965862	49
axo18130.r-at	axo19404.r-at	867.8999484	3
axo14458.r-at	axo27097.f-at	858.7742498	9
axo15017.f-at	axo28795.f-at	852.4527609	68
axo11386.r-at	axo22868.f-at	849.5245811	1
axo05327.f-at	axo17725.r-at	849.209176	21
axo08563.r-at	axo16297.f-at	839.3283611	10
axo05724.r-at	axo12970.f-at	837.1427941	8
axo03867.f-at	axo07474.f-at	817.1866988	40
axo19424.f-at	axo29151.f-at	816.8778713	83
axo03189.r-at	axo11632.r-at	810.8485353	462
axo00209.f-at	axo20011.f-at	792.540108	7
axo16406.f-at	axo31389.f-at	777.1519727	1

Table A.4: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo19787.f-at	axo28043.f-at	774.6741779	5
axo06550.f-at	axo12175.f-at	754.7713887	97
axo01084.r-at	axo15783.r-at	739.591376	14
axo19400.f-at	axo19703.f-at	722.5138158	37
axo08502.f-at	axo14301.f-at	721.1504504	2
axo14921.f-at	axo22382.f-at	721.0379747	18
axo09543.f-at	axo18167.f-at	709.8335618	1
axo11779.f-at	axo11796.r-at	707.7818454	112
axo06571.r-at	axo10212.f-at	703.9872711	7
axo12039.r-at	axo24165.f-at	701.0243758	2
axo04526.f-at	axo06372.r-at	689.7026802	15
axo04414.f-at	axo31366.f-at	686.1520629	1
axo13118.f-at	axo20326.f-at	681.3451949	9
axo10478.r-at	axo25223.f-at	676.4539191	48
axo00512.f-at	axo12095.r-at	642.9880953	17
axo09146.f-at	axo17762.f-at	631.0044436	331
axo03126.r-at	axo17986.f-at	630.7540894	54
axo24102.f-at	axo29428.f-at	617.1936015	1
axo09854.r-at	axo18520.f-at	614.8880811	35
axo18947.f-at	axo30833.f-at	613.4522388	3
axo01587.f-at	axo11192.r-at	609.8603677	4
axo08287.f-at	axo29623.f-at	607.6997445	15
axo16477.f-at	axo19170.r-at	601.4528869	22
axo03201.f-at	axo07901.f-at	597.6817113	2
axo19921.f-at	axo31548.f-at	595.714608	33
axo03591.f-at	axo13078.r-at	588.9752571	105
axo01561.r-at	axo16028.f-at	588.3373148	23
axo10647.f-at	axo23040.f-at	577.9345951	34
axo10179.f-at	axo30114.f-at	577.5116671	2
axo00378.r-at	axo15609.f-at	572.5783764	89
axo27655.f-at	axo28611.f-at	562.6642693	44
axo07768.f-at	axo24058.f-at	562.4299791	47
axo13934.f-at	axo31345.f-at	561.3297671	42
axo15834.r-at	axo28126.f-at	552.5326879	6
axo09621.r-at	axo29936.f-at	541.6465805	8
axo17270.f-at	axo30238.f-at	541.4483206	31
axo05140.r-at	axo05337.f-at	533.160244	13
axo11834.f-at	axo14306.f-at	526.7190708	6
axo12322.f-at	axo30175.f-at	524.5750415	4
axo07030.f-at	axo16275.f-at	519.2985346	4
axo08880.r-at	axo15086.r-at	516.9668911	5
axo11966.f-at	axo25007.f-at	514.5143244	5
axo18536.f-at	axo20620.f-s-at	506.7419369	5
axo25795.f-at	axo29519.f-at	506.4705942	6
axo02806.f-at	axo24512.f-at	505.6600658	5

Table A.5: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo09597.f-at	axo29189.f-at	504.2215151	24
axo22183.r-at	axo28428.f-at	503.7648535	9
axo03620.f-at	axo05776.f-at	502.6577369	3
axo20347.f-s-at	axo27714.f-at	501.5748537	1
axo03441.f-at	axo08206.f-at	501.3647212	3
axo10119.r-at	axo17622.f-at	500.7568582	16
axo04434.r-at	axo24112.r-at	499.3625905	23
axo08033.f-at	axo13300.r-at	495.7356767	9
axo13121.f-at	axo17687.f-at	493.213359	3
axo02002.f-at	axo26828.f-at	486.1288021	7
axo01448.r-at	axo02697.f-at	484.3558945	40
axo00362.f-at	axo11322.r-at	477.9861172	3
axo11258.f-at	axo12454.f-at	476.6697841	1
axo05810.f-at	axo14907.f-at	474.8137094	20
axo01726.f-at	axo26952.f-at	474.2220949	15
axo13234.f-at	axo16221.f-at	471.8504952	22
axo06234.f-at	axo06827.f-at	470.1781628	53
axo17887.f-at	axo26858.f-at	465.3478614	3
axo16574.f-at	axo29993.f-at	464.1742089	48
axo07872.f-at	axo16845.f-at	462.9983425	5
axo20981.f-s-at	axo31623.f-at	461.6073559	42
axo05691.f-at	axo06225.f-at	459.8635308	53
axo09564.f-at	axo24830.f-at	458.9158889	5
axo00992.f-at	axo09743.f-at	453.2636046	4
axo16022.f-at	axo23636.f-at	449.5592269	17
axo28124.f-at	axo31395.f-at	449.524055	13
axo19200.f-at	axo28984.f-at	448.4945544	4
axo08965.f-at	axo10459.f-at	444.7706528	20
axo04765.f-at	axo20251.f-at	444.5382484	2
axo17745.f-at	axo18541.f-at	443.2511347	6
axo06377.f-at	axo13778.f-at	441.8103762	2
axo16018.f-at	axo25819.f-at	441.2268357	3
axo01556.f-at	axo18943.f-at	440.9863323	9
axo00766.r-at	axo08099.f-at	440.52933	4
axo07750.f-at	axo28182.f-at	434.3966919	5
axo06131.f-at	axo07014.r-at	433.9045415	29
axo09037.r-at	axo29485.f-at	431.6477558	14
axo03630.f-at	axo11285.f-at	430.6659946	3
axo04508.f-at	axo09646.f-at	430.0839482	1
axo07571.f-at	axo24299.r-at	430.0749308	5
axo11276.r-at	axo22846.f-s-at	429.5770768	15
axo16490.f-at	axo24217.f-at	426.8002017	2
axo14292.r-at	axo29961.f-at	426.5821176	43
axo04752.r-at	axo12750.f-at	426.2452247	37
axo16595.r-at	axo27849.f-at	423.3556705	2

Table A.6: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo17596.r-at	axo27778.f-at	422.7398305	22
axo13873.r-at	axo26906.f-at	421.9478269	3
axo00009.r-at	axo11578.r-s-at	419.1596672	5
axo02902.r-at	axo12974.f-at	416.4060132	11
axo05797.f-at	axo15340.f-at	412.2339918	14
axo27946.f-at	axo28206.f-at	407.6139076	5
axo02748.f-at	axo19656.f-at	407.4595637	23
axo08873.r-at	axo16726.r-at	402.8523888	3
axo09445.f-at	axo23688.f-at	402.7279465	4
axo12500.r-at	axo30197.f-at	401.4139441	3
axo13780.f-at	axo28741.f-at	400.4379891	20
axo25124.f-at	axo30936.f-at	396.1117207	8
axo03341.f-at	axo08967.f-at	396.0279127	33
axo05649.f-at	axo27404.f-at	395.0452699	2
axo07054.f-at	axo11881.f-at	393.9437218	5
axo08346.f-at	axo10569.r-at	393.3031586	2
axo16466.f-at	axo19270.r-at	390.2353112	1
axo05440.r-at	axo25500.f-at	387.1801896	1
axo05906.r-at	axo11516.f-at	384.3517546	3
axo06348.f-at	axo24555.f-at	383.6899588	5
axo01009.f-at	axo09519.f-at	379.6655041	25
axo10732.f-at	axo28866.f-at	378.1340225	3
axo13050.f-at	axo16196.f-at	377.8915004	1
axo09014.f-at	axo13516.f-at	376.9469225	21
axo25641.f-at	axo27962.f-at	376.0639904	1
axo28085.f-at	axo30929.f-at	374.5823724	3
axo25832.f-at	axo28299.f-at	372.8997231	3
axo13307.r-at	axo14664.f-at	372.0888355	8
axo13418.f-at	axo27525.f-at	371.8841788	6
axo18387.r-at	axo25727.f-at	369.5271065	26
axo07979.r-at	axo13644.r-at	369.020611	32
axo04905.f-s-at	axo13686.f-at	367.8132292	19
axo05366.f-at	axo07528.r-at	367.3058721	5
axo24908.f-at	axo27250.f-at	365.063271	72
axo23123.f-at	axo29395.f-at	363.6140212	6
axo14085.r-at	axo24588.f-at	362.9672165	27
axo07952.f-at	axo20282.r-at	362.3465979	2
axo04647.f-at	axo16834.f-at	361.7309985	1
axo05737.r-at	axo24521.f-at	361.1794664	8
axo28657.f-at	axo31553.f-s-at	361.0065218	11
axo09580.f-at	axo16795.r-at	360.1454354	4
axo05904.f-at	axo24000.f-at	359.0626744	6
axo06357.f-at	axo08528.f-at	358.2696086	3
axo09094.f-at	axo19896.f-at	354.6797541	6
axo01423.r-at	axo15562.r-at	352.9477382	3

Table A.7: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo13347.f-at	axo16299.f-at	351.6846098	37
axo02294.f-at	axo03248.r-at	350.9918254	3
axo11006.f-at	axo12076.f-at	348.7161261	5
axo11593.f-at	axo18715.r-at	347.6022874	3
axo00427.f-at	axo18489.f-at	346.9491847	3
axo29666.f-at	axo30482.f-at	346.4992045	17
axo08059.r-at	axo13225.f-at	345.7715264	17
axo20972.f-at	axo26423.f-at	345.4975446	4
axo13080.f-at	axo24645.f-at	344.1151721	3
axo08395.r-at	axo12402.f-at	342.7840668	5
axo02625.f-at	axo03738.r-at	342.4766493	13
axo00202.r-at	axo25107.f-at	342.4500907	1
axo14784.f-at	axo21563.f-at	342.005855	9
axo05934.f-at	axo28788.f-at	340.942694	11
axo04320.r-at	axo26001.f-at	340.8343812	2
axo06330.r-at	axo22974.f-at	338.4261543	23
axo05596.r-at	axo09441.r-at	337.7083682	125
axo02936.r-at	axo10853.f-at	336.0311581	5
axo12098.f-at	axo16630.f-at	334.3792072	7
axo10708.f-at	axo31449.f-at	333.0581836	6
axo05742.r-at	axo06397.r-at	331.9598533	3
axo23866.f-at	axo27878.f-at	330.5738074	34
axo12118.r-at	axo20097.f-at	329.8410095	4
axo25635.f-at	axo30797.f-at	328.4345093	17
axo19011.f-at	axo19416.f-at	326.5017718	9
axo21610.r-at	axo24634.f-at	325.9171927	8
axo02731.f-at	axo26905.f-at	325.2983766	1
axo11431.r-at	axo12504.f-at	324.5712958	3
axo09417.f-at	axo22670.f-at	324.4906481	11
axo01565.f-at	axo18364.f-at	322.8319361	22
axo14440.f-at	axo19145.r-at	318.9199076	3
axo06635.f-at	axo07844.r-at	315.2144902	2
axo02211.f-at	axo25769.f-at	314.7098491	2
axo02997.f-at	axo10988.r-at	314.1473589	3
axo18604.r-at	axo19556.f-at	313.5518101	4
axo05624.r-at	axo11791.f-at	313.2862096	1
axo03310.r-at	axo16719.r-at	312.3737043	3
axo15361.r-at	axo28243.f-at	311.8847452	8
axo17861.f-at	axo19895.f-at	311.5431378	12
axo17507.f-at	axo30376.f-at	309.8522913	25
axo00832.f-at	axo16553.r-at	308.3756048	11
axo11157.f-at	axo28022.f-at	306.80073	2
axo05123.f-at	axo09374.f-at	306.751086	9
axo03643.r-at	axo17205.f-at	306.0764203	14
axo07854.r-at	axo20003.r-at	305.7728156	2

Table A.8: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo00481.f-at	axo09368.f-at	305.6494569	1
axo13106.f-at	axo28824.f-at	304.1040374	2
axo06014.r-at	axo31644.f-at	303.9995948	8
axo12438.f-at	axo24772.f-at	302.6918564	2
axo15268.f-at	axo28777.f-at	301.8826118	4
axo12594.f-at	axo21241.f-at	301.4799254	35
axo00438.r-at	axo09010.f-at	301.4365375	4
axo03321.f-at	axo29336.f-at	301.3157277	17
axo24595.f-at	axo27732.f-at	301.1896949	9
axo15519.r-at	axo18528.f-at	300.8692454	3
axo17451.f-at	axo26099.f-at	300.7580657	2
axo17581.f-at	axo20319.r-at	299.855087	2
axo01545.f-at	axo01575.r-at	299.0435448	2
axo15467.r-at	axo16829.f-at	298.7091552	27
axo00344.f-at	axo15121.f-at	298.428108	3
axo10968.f-at	axo13890.r-at	298.4045718	3
axo04631.f-at	axo16307.f-s-at	298.3638985	2
axo05058.f-at	axo18717.r-at	294.746986	4
axo00708.f-at	axo02690.f-s-at	294.4292531	3
axo01840.f-at	axo25573.f-at	293.6247125	12
axo01796.r-at	axo18197.f-at	293.2611479	7
axo08280.r-at	axo10787.f-at	293.2467848	3
axo08125.f-at	axo11361.f-at	293.01246	2
axo01986.f-at	axo31231.f-at	293.0027303	7
axo15733.f-at	axo17956.f-at	287.7408356	2
axo01247.f-at	axo12324.f-at	287.2244229	4
axo10155.f-at	axo23402.f-at	287.1105944	1
axo10677.f-at	axo15793.f-at	286.8111292	17
axo05739.r-s-at	axo24605.f-at	286.2387036	4
axo06236.r-at	axo11852.r-at	285.5610528	5
axo13810.f-at	axo13846.f-at	284.2859851	10
axo11000.f-at	axo23816.f-at	284.0734577	7
axo07877.r-at	axo27548.f-at	283.1828071	3
axo04533.f-at	axo10684.f-at	282.8408111	8
axo10590.f-at	axo12458.f-at	281.9537236	13
axo09986.f-at	axo24647.f-at	281.745512	25
axo14629.f-at	axo18438.r-at	280.686766	4
axo01432.f-at	axo14329.r-at	280.6241902	4
axo05333.r-at	axo24876.f-at	278.7564669	13
axo17915.f-at	axo25776.f-at	278.7563703	1
axo08664.r-at	axo14338.f-at	278.4296277	2
axo13471.f-at	axo30170.f-at	277.9686421	2
axo13524.r-at	axo29692.f-at	277.8063288	5
axo10535.f-at	axo29975.f-at	277.5849539	10
axo02933.r-at	axo04993.r-at	277.4426072	6

Table A.9: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo25389.f-at	axo30307.f-at	277.1853503	11
axo02362.f-at	axo07038.r-at	276.173961	3
axo05399.f-at	axo15155.f-at	275.9165402	2
axo00492.f-at	axo29042.f-at	275.8376552	7
axo00231.f-at	axo23230.f-at	275.7530591	13
axo25856.f-at	axo31309.f-at	275.2298472	8
axo03419.r-at	axo11560.f-at	274.5232809	9
axo04644.f-at	axo13874.f-at	273.9958675	3
axo02039.f-at	axo06663.f-at	273.3064385	66
axo15773.r-at	axo17084.f-at	273.0355539	3
axo04830.f-at	axo15683.f-at	272.7646626	5
axo20252.f-at	axo22665.f-at	270.6852353	4
axo14279.f-at	axo15022.r-at	268.9198099	24
axo03743.f-at	axo04985.f-at	268.8949079	20
axo01319.f-at	axo13267.f-at	268.5402168	9
axo12214.r-at	axo20594.f-at	267.5295591	6
axo02397.f-at	axo05252.f-at	267.1877033	1
axo05734.f-at	axo19903.f-at	266.6961598	5
axo07785.r-at	axo15508.f-at	266.6843882	1
axo05053.f-at	axo07917.r-at	264.5596733	2
axo03718.f-at	axo30373.f-at	263.4241713	3
axo19008.f-at	axo19077.f-at	262.849957	4
axo13035.f-at	axo18266.f-at	261.5764524	2
axo01896.f-at	axo19341.r-at	261.3003371	17
axo05854.f-at	axo16806.f-at	260.9924271	4
axo08187.f-at	axo14370.r-at	260.320172	3
axo13730.f-at	axo14167.r-at	260.1845937	1
axo02080.f-at	axo11765.f-at	260.0462481	8
axo00790.r-at	axo12369.f-at	259.2527783	5
axo13123.r-at	axo31194.f-s-at	258.6886416	3
axo18298.f-at	axo26230.f-at	258.4819759	1
axo04457.f-at	axo04798.f-at	258.4628803	5
axo02575.f-at	axo17239.f-at	257.7451881	3
axo06993.f-at	axo07319.f-at	256.9027292	18
axo06174.f-at	axo25173.f-at	256.6510729	1
axo08843.r-at	axo12960.f-at	256.4270569	4
axo08569.f-at	axo19594.f-at	256.4245865	2
axo09365.f-at	axo18602.f-at	256.1070298	4
axo10758.r-at	axo31466.f-s-at	255.9102435	8
axo07039.f-at	axo17080.f-at	255.2919769	2
axo12205.f-at	axo30518.f-at	255.2208806	4
axo12809.f-at	axo18055.f-at	255.0993919	3
axo05452.f-at	axo18194.f-at	252.9664445	5
axo02293.f-at	axo28826.f-at	252.7856621	15
axo17287.f-at	axo22159.r-at	252.4605227	4

Table A.10: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo01688.f-at	axo18609.r-at	252.1030933	3
axo08827.f-at	axo25398.f-at	251.5251559	7
axo04262.r-at	axo11319.f-at	251.3718938	4
axo01197.r-at	axo11411.f-at	250.7907864	19
axo08811.r-at	axo10249.f-at	249.3880224	19
axo14772.r-at	axo30672.f-at	249.2390175	2
axo05919.f-at	axo19751.f-at	249.1823956	2
axo19762.f-at	axo24331.r-at	248.2330458	5
axo02602.f-at	axo12802.r-at	247.085893	2
axo22243.f-at	axo25552.f-at	246.1943571	1
axo06447.f-at	axo28117.f-at	244.6719772	4
axo02987.r-at	axo24959.f-at	243.7679192	11
axo07201.f-at	axo24748.f-at	243.5025671	2
axo13628.f-at	axo27927.f-at	243.1775047	4
axo06623.r-at	axo10373.r-at	240.2317418	3
axo06652.r-at	axo15855.f-at	239.8025596	14
axo08411.f-s-at	axo26889.f-at	239.1366303	10
axo02356.f-at	axo04209.r-at	238.880843	9
axo04642.f-at	axo24147.f-at	238.0184101	2
axo01736.f-at	axo06128.f-at	237.8919889	4
axo11885.f-at	axo24427.f-at	237.7659017	5
axo01996.f-at	axo15538.f-at	237.6321373	7
axo00746.r-at	axo07499.f-at	237.3749744	5
axo08449.r-at	axo19900.f-at	237.0173769	4
axo04823.f-at	axo08546.f-at	236.5410462	2
axo19273.f-at	axo30637.f-at	236.5237364	2
axo10868.f-at	axo30009.f-at	236.4434355	15
axo23332.f-s-at	axo28255.f-at	236.3381017	4
axo08442.f-at	axo20463.f-s-at	235.6763857	2
axo13976.f-at	axo19959.f-at	235.1436766	12
axo09105.f-at	axo29128.f-at	234.8188972	3
axo06486.f-at	axo11024.f-at	234.7554918	16
axo00698.f-at	axo27279.f-at	234.0329164	2
axo25872.f-at	axo30725.f-at	233.9535256	4
axo05662.f-at	axo11020.r-at	233.2273438	3
axo08948.f-at	axo09359.f-at	232.7675679	5
axo07349.f-at	axo08861.f-at	231.5401602	7
axo07508.f-at	axo19537.f-at	230.8769904	1
axo02163.f-at	axo16996.f-at	230.0297975	1
axo10404.f-at	axo30268.f-at	229.79692	2
axo05330.r-at	axo25879.f-at	229.4920169	12
axo03021.f-at	axo05220.r-at	225.9045679	1
axo18233.r-at	axo24548.f-at	224.2855683	4
axo01673.r-at	axo12631.r-at	223.7480156	3
axo13916.r-at	axo21614.r-at	223.5949408	9

Table A.11: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo00012.r-at	axo19124.r-at	223.5914182	13
axo02631.r-at	axo27914.f-at	221.9165867	3
axo00033.f-at	axo15090.f-at	221.7342192	6
axo08526.f-at	axo27604.f-at	221.4212515	4
axo07228.r-at	axo23217.f-at	221.2262475	12
axo05976.f-at	axo12897.f-at	220.7294833	8
axo00797.r-at	axo05285.f-at	220.2307394	3
axo18767.r-at	axo25349.f-at	219.5498515	8
axo11759.f-at	axo12513.r-at	218.4865683	9
axo09043.r-at	axo13850.f-at	217.918214	3
axo11691.r-s-at	axo25615.f-at	217.3084934	1
axo06423.f-at	axo13219.f-at	216.8397968	2
axo07396.f-at	axo11378.r-at	216.4139687	2
axo14609.f-at	axo22563.r-at	215.7796207	5
axo13519.r-at	axo17115.f-at	215.7328108	9
axo12178.r-at	axo21492.f-at	215.6324168	3
axo02765.f-at	axo09727.f-at	215.4275653	6
axo28246.f-at	axo30841.f-at	215.4192423	12
axo10228.f-at	axo30144.f-at	215.2326643	1
axo10630.f-at	axo25035.f-at	215.1059215	4
axo06256.f-at	axo07090.f-at	215.062505	2
axo12797.f-at	axo31431.f-s-at	214.7702395	4
axo07467.f-at	axo08156.r-at	214.5120483	9
axo00752.f-at	axo31494.f-s-at	214.4729244	10
axo12951.f-at	axo18705.r-at	213.6568036	2
axo03989.f-at	axo05353.r-at	213.309962	7
axo03618.f-at	axo25753.f-at	211.9136523	10
axo10800.r-at	axo29425.f-at	211.7091427	16
axo12649.f-at	axo26418.f-at	211.5780169	3
axo02874.r-at	axo19515.r-at	211.3048865	5
axo28425.f-at	axo29302.f-at	211.264296	19
axo18458.r-at	axo30048.f-at	210.1092533	6
axo02581.f-at	axo03135.r-at	209.3935129	5
axo12126.r-at	axo31525.f-at	208.2971606	4
axo03212.f-at	axo10335.f-at	207.7078995	3
axo03110.f-at	axo12980.f-at	206.4597171	7
axo10646.f-at	axo17323.r-s-at	205.7351607	3
axo10437.f-at	axo22604.f-at	204.9159748	1
axo06188.f-at	axo14067.r-at	204.5129433	2
axo07093.f-at	axo09206.r-at	204.4231649	2
axo05777.f-at	axo14707.f-at	203.3971814	3
axo06967.r-at	axo10034.f-s-at	202.3761561	2
axo26985.f-at	axo27386.f-at	202.1482432	11
axo11901.f-at	axo24570.f-at	202.1262761	9
axo05430.f-at	axo05950.r-at	202.1061414	9

Table A.12: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo25815.f-at	axo28642.f-at	202.051211	5
axo04621.f-at	axo18685.r-at	201.9430451	1
axo06859.f-at	axo12412.r-at	201.093332	10
axo03103.f-at	axo30938.f-at	200.8221581	5
axo08576.f-at	axo22394.f-at	200.7873796	2
axo13604.f-at	axo28822.f-at	199.6037566	1
axo05433.f-at	axo09952.r-at	198.7936868	1
axo16235.f-at	axo25907.f-at	198.6261762	4
axo01820.f-at	axo12416.f-at	198.6241987	5
axo04880.f-at	axo19252.f-at	198.3492228	8
axo01572.r-at	axo09065.f-at	198.2167032	3
axo01392.f-at	axo02574.f-at	197.9647384	3
axo02176.f-at	axo08748.f-at	197.9008969	1
axo15276.f-at	axo19417.f-at	196.5992519	1
axo05713.f-at	axo10018.f-at	196.1047729	3
axo00104.r-at	axo01868.f-at	196.0992724	1
axo23568.f-at	axo31619.f-at	195.913077	4
axo06350.r-at	axo16649.r-at	195.6632371	3
axo02789.f-at	axo10604.f-at	194.5400702	3
axo03509.f-at	axo19414.r-at	194.4919537	4
axo10609.f-at	axo28626.f-at	194.4869466	2
axo08368.f-at	axo11600.f-at	193.6347474	9
axo01743.f-at	axo12337.r-at	193.2543707	7
axo14856.r-at	axo18042.f-at	193.2158682	6
axo10855.r-at	axo16111.r-at	193.1141477	4
axo01858.f-at	axo28029.f-at	193.043073	7
axo11387.f-at	axo17929.f-at	192.7458467	1
axo12351.f-at	axo18051.f-at	192.5804101	2
axo04169.f-at	axo16678.f-at	192.3047649	2
axo11931.f-at	axo15257.f-at	192.2913175	3
axo01878.f-at	axo06680.f-at	192.2286086	6
axo10875.f-at	axo11374.r-at	192.130475	6
axo20198.f-at	axo27006.f-at	192.101078	1
axo05997.f-at	axo17599.f-at	192.0332115	2
axo00504.f-at	axo00986.f-at	191.5595031	1
axo05505.r-at	axo14087.f-at	191.3922564	3
axo16468.r-at	axo25240.f-at	191.3281888	9
axo24604.f-at	axo31492.f-at	190.7525879	2
axo15647.f-at	axo18346.f-at	190.7372723	3
axo02116.f-at	axo14881.f-at	190.5791382	8
axo17626.f-at	axo28460.f-at	189.9464406	1
axo22073.f-at	axo28475.f-at	189.9382663	7
axo09607.f-at	axo31195.f-s-at	189.7476692	1
axo09935.f-at	axo17344.r-at	189.5437809	15
axo12380.f-at	axo15315.f-at	189.4817298	4

Table A.13: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo10195.f-at	axo27495.f-at	189.426621	3
axo01867.f-at	axo07083.r-at	188.8776395	2
axo06392.f-at	axo28276.f-at	187.3818734	2
axo08934.f-at	axo17506.f-at	187.1284934	3
axo09117.r-at	axo16436.r-at	187.0761329	7
axo04868.f-at	axo28274.f-at	186.7223254	13
axo13681.f-at	axo24136.f-at	186.0544926	3
axo09794.f-at	axo18328.r-at	185.9676132	1
axo07536.f-at	axo15292.r-at	185.6500664	3
axo00626.f-at	axo15559.f-at	185.3698725	9
axo15224.f-at	axo25415.f-at	185.1149363	8
axo05823.f-at	axo05940.f-at	185.0178811	8
axo03538.f-at	axo28864.f-at	184.3808561	6
axo03842.r-at	axo08609.f-at	183.9108874	2
axo06425.f-at	axo16008.f-at	183.0443242	1
axo16330.f-at	axo19509.f-at	182.7359445	5
axo08869.f-at	axo17145.r-at	182.5239301	3
axo11323.f-at	axo28008.f-at	182.5018005	3
axo06178.f-at	axo21124.f-at	182.4322562	1
axo07876.f-at	axo12350.f-at	181.0609099	3
axo11671.f-at	axo29096.f-at	180.898061	5
axo14806.f-at	axo30370.f-at	180.6609856	2
axo14390.f-at	axo15034.r-at	180.2755162	8
axo19140.f-at	axo26236.f-at	180.2390556	2
axo08701.f-at	axo19048.f-at	180.0264265	5
axo12002.f-at	axo23953.f-at	180.0060889	2
axo03492.f-at	axo16322.r-at	179.9489968	3
axo09390.f-at	axo18838.f-at	179.8377014	1
axo00213.f-at	axo16358.r-at	179.5295146	15
axo04719.r-at	axo31579.f-at	179.3378483	1
axo13731.f-at	axo18096.r-at	179.1133211	5
axo12249.r-at	axo30910.f-at	179.0552159	9
axo06771.r-at	axo17955.f-at	178.9118122	4
axo22480.f-at	axo30583.f-at	178.718927	2
axo08969.f-at	axo09053.f-at	176.532083	7
axo14383.r-at	axo15917.f-at	175.9370051	8
axo00768.f-at	axo15305.r-at	175.8501613	16
axo13724.r-at	axo24389.f-at	175.8239398	5
axo01600.r-at	axo15503.f-at	175.6872541	4
axo03445.r-at	axo08253.f-at	175.2512864	4
axo09859.f-at	axo30169.f-at	175.1469982	1
axo07897.f-at	axo16125.f-at	174.7704785	2
axo18157.f-at	axo27517.f-at	174.4418363	1
axo10419.r-at	axo14238.f-at	174.3993865	1
axo22605.f-at	axo30739.f-at	173.9513634	2

Table A.14: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo00715.r-at	axo07312.f-at	173.6505918	8
axo11226.r-at	axo15539.f-at	173.5524725	7
axo15635.f-at	axo22086.f-at	173.2733061	2
axo02796.f-at	axo13460.r-at	173.088495	3
axo27254.f-at	axo28665.f-at	172.5364789	2
axo07522.f-at	axo23819.r-at	171.7530039	3
axo00381.f-at	axo09466.r-at	170.5435047	9
axo13436.f-at	axo25150.f-at	170.5063067	6
axo02391.f-at	axo15449.f-at	170.432592	6
axo18442.f-at	axo30028.f-at	170.2806938	5
axo02480.r-at	axo24234.f-at	170.2469708	4
axo24483.f-at	axo29086.f-at	169.9648973	3
axo25296.f-at	axo28210.f-at	169.8500166	4
axo19657.f-at	axo28680.f-at	169.4636599	2
axo07690.r-at	axo16374.r-at	169.4042932	9
axo00158.f-at	axo02576.f-at	168.9016722	2
axo01946.r-at	axo29367.f-at	168.8728698	4
axo04194.r-at	axo15179.r-at	168.850299	2
axo09514.r-x-at	axo11777.f-at	168.5635563	11
axo06239.f-at	axo16335.r-at	167.9636927	3
axo03007.f-at	axo30483.f-at	167.865506	6
axo11479.r-at	axo21623.f-at	167.2871035	5
axo12832.f-at	axo19391.r-at	167.2523526	1
axo17302.f-at	axo23778.f-at	167.234679	1
axo03656.f-at	axo07863.r-at	166.9683359	4
axo08421.f-at	axo17412.r-at	166.8342996	7
axo11052.f-at	axo18908.r-at	166.6184085	5
axo17050.f-at	axo22242.r-at	166.601185	5
axo11096.f-at	axo15537.f-at	165.9718612	3
axo04824.r-at	axo07000.f-at	165.875221	3
axo05531.f-at	axo24978.f-at	165.450214	1
axo03703.f-at	axo09783.f-at	165.419717	12
axo17598.f-at	axo18928.r-at	165.4095169	2
axo02632.f-at	axo12252.f-at	164.7372479	5
axo04107.f-at	axo27114.f-at	164.2501454	12
axo12979.f-at	axo17124.f-at	164.2180523	9
axo05983.f-at	axo29772.f-at	164.1793222	6
axo04780.f-at	axo25970.f-at	164.1639075	4
axo08369.r-at	axo12299.r-at	164.1118875	2
axo09497.r-at	axo11797.f-at	164.0929587	2
axo19212.f-at	axo19694.r-at	164.078166	6
axo04523.r-at	axo22427.f-at	164.0008352	2
axo17249.f-at	axo18566.f-at	163.9597375	3
axo12701.f-at	axo31368.f-at	163.7913532	2
axo19428.r-at	axo21568.f-at	163.7873712	3

Table A.15: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo09570.f-at	axo25015.f-at	163.3768506	1
axo15166.f-at	axo18818.f-at	163.1563297	3
axo04267.r-at	axo14861.f-at	163.0352354	2
axo04195.r-at	axo19325.f-at	162.779267	2
axo06795.f-at	axo24815.f-at	162.3936452	2
axo01343.f-at	axo16424.f-at	162.1145937	1
axo00103.f-at	axo24039.f-at	161.8869841	2
axo19769.r-at	axo28466.f-at	161.8102113	5
axo09121.f-at	axo18547.f-at	161.6597184	3
axo04462.r-at	axo12190.f-at	161.2240285	4
axo07417.f-at	axo20436.f-s-at	160.8641751	1
axo02121.r-at	axo30581.f-at	160.7751871	1
axo03306.f-at	axo17567.f-at	160.7484624	4
axo04691.r-at	axo16012.f-at	160.7050485	14
axo10237.r-at	axo25940.f-at	160.575747	3
axo00783.r-at	axo01147.r-at	160.0618704	2
axo02130.f-at	axo22784.r-at	159.8498481	3
axo05294.f-at	axo23992.f-at	159.8256757	4
axo07948.f-at	axo19962.r-at	159.2227388	3
axo05954.r-at	axo23711.f-at	158.8026379	3
axo01025.f-at	axo27407.f-at	158.2850323	9
axo23137.f-at	axo27912.f-at	157.7755797	2
axo01763.f-at	axo08505.f-at	157.4929546	1
axo16141.r-at	axo17092.r-at	157.4894739	11
axo02413.f-at	axo13502.f-at	157.1803753	3
axo08978.f-at	axo18888.f-at	156.8807707	4
axo03510.r-at	axo25803.f-at	156.4892693	3
axo00036.f-at	axo06323.r-at	156.3500893	2
axo08351.f-at	axo27896.f-at	156.0130886	2
axo07606.f-at	axo12378.f-at	155.6842004	11
axo11663.f-at	axo29912.f-at	155.359731	1
axo07776.f-at	axo20082.f-at	155.3424213	2
axo00415.f-at	axo16147.f-at	155.3049241	2
axo17002.f-at	axo29588.f-at	155.1112863	2
axo04825.r-at	axo16168.r-at	155.0965903	3
axo10843.f-at	axo15511.r-at	154.9740931	2
axo02169.f-at	axo27459.f-at	154.321304	5
axo13832.f-at	axo16696.f-at	154.0266414	2
axo11722.f-at	axo27654.f-at	153.9811638	1
axo04943.r-at	axo06391.r-at	153.7510022	9
axo05569.f-at	axo21948.f-at	153.7393321	1
axo06921.f-at	axo27101.f-at	153.4921792	2
axo13701.f-at	axo16525.r-at	153.4912718	1
axo00302.f-at	axo10520.f-at	152.6967842	3
axo14783.r-at	axo30381.f-at	152.4066913	2

Table A.16: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo01328.r-at	axo11057.r-at	152.3412458	3
axo00056.r-at	axo13739.f-at	152.1830956	6
axo12943.r-at	axo17266.f-at	151.8639279	4
axo15766.f-at	axo15929.f-at	151.8527115	1
axo00703.r-at	axo24844.f-at	151.712809	1
axo11463.f-at	axo30224.f-at	151.5757099	3
axo12702.f-at	axo27091.f-at	150.7517702	2
axo00159.f-at	axo29106.f-at	150.591265	5
axo13919.f-at	axo15673.f-at	150.447842	3
axo07148.f-at	axo08508.f-at	150.1408172	3
axo06149.f-at	axo25817.f-at	149.8596243	2
axo02566.f-at	axo21688.f-at	149.8389697	5
axo11972.r-at	axo27104.f-at	149.7297497	2
axo01891.f-at	axo11315.f-at	149.555541	2
axo04445.f-at	axo28988.f-at	149.3998886	5
axo06893.f-at	axo22545.f-s-at	149.3629798	1
axo00270.f-at	axo13319.f-at	149.2435061	2
axo13446.f-at	axo18923.r-at	149.0200498	1
axo02283.r-at	axo09742.r-at	148.9084618	3
axo03242.r-at	axo07413.f-at	148.7762545	1
axo00968.f-at	axo18589.f-at	148.7542804	3
axo08361.f-at	axo25527.f-at	148.4353047	5
axo15200.r-at	axo28017.f-at	147.9991163	2
axo03006.f-at	axo24582.f-at	147.6917145	1
axo12271.f-at	axo12855.f-at	147.3738348	4
axo03262.f-at	axo25479.f-at	146.5417417	4
axo10735.f-at	axo16644.f-at	146.0833791	4
axo16740.f-at	axo25496.f-at	145.9792804	2
axo09099.f-at	axo12570.f-at	145.2387712	3
axo24666.f-at	axo29765.f-at	145.0632888	2
axo04978.f-at	axo11338.r-at	145.0281926	6
axo16936.r-at	axo17042.f-at	144.9239643	3
axo06269.f-at	axo08225.f-at	144.8381156	3
axo06394.f-at	axo07368.f-at	144.7447883	6
axo04208.f-at	axo07404.f-at	144.6410786	4
axo03031.f-at	axo10857.f-at	144.376147	2
axo09128.f-at	axo27543.f-at	143.9982993	3
axo00636.f-at	axo28303.f-at	143.849328	4
axo08542.r-at	axo13148.r-at	143.5668282	7
axo06371.f-at	axo10652.f-at	143.1626681	1
axo04973.r-at	axo07418.r-at	142.9395751	5
axo12200.f-at	axo19258.f-at	142.6840494	2
axo08506.f-at	axo18505.r-at	142.4621259	3
axo08821.f-at	axo20233.f-at	142.3445308	3
axo10999.f-at	axo17688.f-at	142.3385484	2

Table A.17: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo01586.r-at	axo07646.f-at	142.3318786	1
axo06772.f-at	axo18014.r-at	142.2484629	1
axo00346.r-at	axo12194.f-at	141.2248634	1
axo05485.f-at	axo20024.r-at	141.077256	3
axo02885.r-at	axo09864.f-at	140.9353917	3
axo08011.f-at	axo08240.f-at	140.8972824	2
axo05746.f-at	axo15015.r-at	140.6992449	3
axo19672.f-at	axo31667.f-at	140.27008	3
axo02577.r-at	axo25584.f-at	139.9460125	3
axo16536.f-at	axo18696.f-at	139.7167558	2
axo09162.f-at	axo16486.f-at	139.6182501	4
axo01503.f-at	axo09363.r-at	139.5971658	2
axo09018.f-at	axo30648.f-at	139.5196249	5
axo10825.f-at	axo16707.f-at	139.4798446	4
axo09923.r-at	axo27196.f-at	139.2118172	2
axo05352.f-at	axo12401.f-at	139.0851239	4
axo20167.r-at	axo21847.f-at	138.8758851	4
axo00031.f-at	axo07003.f-at	138.6838041	3
axo11787.f-at	axo19877.r-at	138.4744837	1
axo10795.f-at	axo29487.f-at	137.6660589	7
axo03569.f-at	axo11414.f-at	137.5130102	2
axo04246.r-at	axo15913.r-at	137.432455	1
axo02975.r-at	axo09510.r-at	137.2217266	1
axo16233.f-at	axo18789.f-at	137.1641198	3
axo17852.f-at	axo22896.f-at	135.909303	2
axo06694.r-at	axo20191.f-at	135.8166276	3
axo05407.r-at	axo30549.f-at	135.711958	4
axo13369.f-at	axo27507.f-at	135.5815989	2
axo00097.f-at	axo08045.f-at	135.1068763	4
axo02420.f-at	axo21880.r-at	134.4394781	4
axo16928.f-at	axo30467.f-at	134.4165772	3
axo06183.f-at	axo30485.f-at	134.1744613	3
axo17057.r-at	axo19863.r-at	133.8979031	2
axo09008.f-at	axo26894.f-at	133.7443636	3
axo07100.f-at	axo17895.f-at	133.6915207	1
axo12856.f-at	axo14174.f-at	133.3602901	2
axo05895.f-at	axo16417.f-at	133.0703533	2
axo17950.f-at	axo29842.f-at	132.943194	3
axo04151.r-at	axo24094.r-at	132.8247613	1
axo08882.f-at	axo18695.r-at	132.7525347	1
axo12398.r-at	axo16199.r-at	132.4771104	5
axo13192.f-at	axo27233.f-at	132.4531812	5
axo30596.f-at	axo30823.f-at	132.2975933	2
axo04680.r-at	axo30591.f-at	132.2717869	3
axo16256.f-at	axo24858.f-at	132.0249485	1
axo01436.f-at	axo11332.f-at	131.6967874	7

Table A.18: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo01613.f-at	axo07225.f-at	131.6384868	4
axo05419.f-at	axo24656.f-at	131.2706249	2
axo11122.f-at	axo15717.r-at	130.5535774	3
axo06857.f-at	axo13398.r-at	130.1619222	1
axo12981.f-at	axo22139.f-at	130.0765283	11
axo12218.f-at	axo19601.r-at	130.0012722	6
axo06032.f-at	axo24002.f-at	129.9739723	1
axo15003.f-at	axo24870.f-at	129.8120408	1
axo00028.f-at	axo10431.f-at	129.7981029	5
axo03091.f-at	axo16521.f-at	129.6634833	3
axo16472.f-at	axo18994.f-at	129.2740583	2
axo11630.f-at	axo16065.f-at	129.0739348	3
axo05665.f-at	axo08464.f-at	128.831189	2
axo09177.f-at	axo14675.f-at	128.8281081	2
axo22648.f-at	axo24561.f-at	128.2116112	4
axo11168.r-at	axo21646.f-at	127.9017387	3
axo02107.f-at	axo06565.r-at	127.4007425	1
axo01892.f-at	axo15632.f-at	127.3955043	4
axo18535.r-at	axo19913.r-at	126.5712914	8
axo02344.r-at	axo12672.f-at	126.5051095	1
axo07062.f-at	axo12625.f-at	126.4217759	4
axo21729.r-at	axo30886.f-at	126.4197517	3
axo15617.r-at	axo17146.r-at	126.2770104	2
axo07886.f-at	axo24613.f-at	126.0225488	1
axo07816.f-at	axo26827.f-at	125.6934742	1
axo13642.f-at	axo24251.f-at	125.6921965	5
axo02893.f-at	axo30071.f-at	124.9754202	1
axo11821.f-at	axo31454.f-at	124.612245	4
axo09387.f-at	axo19506.f-at	124.1422009	4
axo05425.r-at	axo29942.f-at	124.0420193	2
axo07925.f-at	axo16105.f-at	124.0323485	2
axo04376.f-at	axo29705.f-at	123.9961775	1
axo07377.r-at	axo14583.f-at	122.852885	1
axo04253.f-at	axo18217.r-at	122.784225	2
axo07268.r-at	axo19840.f-at	122.5344011	4
axo19570.f-at	axo25072.f-at	122.3480085	1
axo15581.f-at	axo19218.f-at	122.1363942	3
axo16701.r-at	axo26426.f-at	122.1155536	2
axo07945.r-at	axo25993.f-at	121.9837457	5
axo07108.f-at	axo22026.r-at	121.5358856	2
axo02696.f-at	axo07572.f-at	121.4999107	5
axo10199.f-at	axo25146.f-at	121.348254	4
axo02166.f-at	axo07445.f-at	121.118848	1
axo00330.f-at	axo31258.f-at	121.1004549	3
axo01724.r-at	axo09002.f-at	120.5108493	3
axo07898.f-at	axo19544.f-at	120.394673	1

Table A.19: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo25263.f-at	axo30847.f-at	120.1104708	3
axo04431.f-at	axo31399.f-at	119.6423184	4
axo16881.r-at	axo29304.f-at	119.5497182	9
axo01552.r-at	axo11563.f-at	119.3390904	1
axo03500.r-at	axo15525.r-at	118.7459604	2
axo01433.r-at	axo12787.f-at	117.5837699	6
axo10222.r-at	axo10945.f-at	117.2421801	3
axo02335.f-at	axo04917.f-at	117.1938591	2
axo21612.f-at	axo24218.f-at	116.8425777	2
axo14983.f-at	axo17667.f-at	116.5359775	3
axo09114.r-at	axo15905.f-at	116.4911174	1
axo06840.f-at	axo17047.f-at	115.8538478	1
axo04158.f-at	axo06864.r-at	114.7771695	2
axo06262.f-at	axo29905.f-at	114.554718	5
axo07107.r-at	axo18072.f-at	114.4622913	6
axo04164.f-at	axo15040.f-at	114.3691233	2
axo19997.r-at	axo30722.f-at	114.2381434	4
axo03290.f-at	axo13177.r-at	114.2039631	3
axo20884.f-s-at	axo27692.f-at	114.1569832	1
axo02390.r-at	axo13515.f-at	113.9501763	3
axo07858.f-at	axo18412.f-at	113.7325973	1
axo15504.f-at	axo25178.f-at	113.4748415	4
axo25497.f-at	axo28519.f-at	113.237962	3
axo04132.r-at	axo30799.f-at	112.9415521	2
axo13327.f-at	axo29672.f-at	112.314436	2
axo08510.f-at	axo25601.f-at	111.7818895	3
axo07799.f-at	axo14215.f-at	111.7032158	1
axo03456.f-at	axo12107.f-at	111.6494916	3
axo05030.r-s-at	axo15348.r-at	111.4931631	2
axo01437.f-at	axo10976.r-at	111.4255978	2
axo15988.f-at	axo17094.f-at	111.3490973	3
axo13535.f-at	axo25570.f-at	110.9740651	2
axo08584.r-at	axo24639.f-at	110.7127634	2
axo02811.f-at	axo04343.f-at	110.2616375	5
axo12136.r-at	axo22171.f-at	110.0814873	2
axo04498.f-at	axo17973.f-at	110.0453504	1
axo19963.f-at	axo30513.f-at	109.8078347	3
axo13825.r-at	axo16280.f-at	109.5566315	2
axo05096.f-at	axo08177.f-at	109.4689317	2
axo15904.r-at	axo27678.f-at	109.452145	2
axo14755.r-at	axo29172.f-at	109.4515053	1
axo01149.r-at	axo12281.f-at	109.406662	2
axo05469.r-at	axo08541.r-at	109.386928	2
axo10274.f-at	axo31394.f-s-at	109.3218702	3
axo04361.f-at	axo18036.f-at	109.1838179	3
axo19010.f-at	axo30229.f-at	109.0954607	1

Table A.20: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo06185.f-at	axo29769.f-at	109.0361469	2
axo06373.f-at	axo17518.f-at	109.0122699	3
axo08849.f-at	axo24472.r-s-at	108.6968361	2
axo11565.f-at	axo21895.r-at	108.6526531	1
axo08883.f-at	axo27665.f-at	108.3689807	1
axo15653.f-at	axo17025.r-at	108.2411447	1
axo13378.f-at	axo19597.f-at	108.1835439	2
axo04989.r-at	axo16623.r-at	107.8154924	1
axo18314.f-at	axo27119.f-at	107.5325783	1
axo03827.r-at	axo21958.r-at	107.2953228	3
axo09709.f-at	axo29927.f-at	107.0376979	5
axo08592.r-at	axo13945.f-at	106.9758139	1
axo05931.f-at	axo18117.r-at	106.691469	2
axo10143.f-at	axo25924.f-at	106.379234	1
axo14001.f-at	axo15377.f-at	106.2194842	1
axo10504.f-at	axo22489.f-at	105.8077798	3
axo25307.f-at	axo31646.f-at	105.4943572	3
axo10202.f-at	axo24995.f-at	105.0379762	1
axo13212.f-at	axo27679.f-at	104.8525948	1
axo14948.f-at	axo30094.f-at	104.4128979	1
axo13145.f-at	axo19527.f-at	104.2262509	6
axo17776.f-at	axo26966.f-at	103.8976029	1
axo15731.f-at	axo28980.f-at	103.8022292	5
axo10142.f-at	axo28816.f-at	103.7024387	3
axo02444.f-at	axo10539.f-at	103.6605922	3
axo14909.f-at	axo25020.f-at	103.4981433	3
axo10014.f-at	axo27401.f-at	103.3023646	2
axo20626.f-at	axo23288.r-at	103.1125237	1
axo03444.f-at	axo08984.f-at	102.8650076	4
axo04281.r-at	axo20834.f-at	102.7837274	3
axo02129.f-at	axo03054.f-at	102.5230608	1
axo04448.f-at	axo19887.f-at	102.4455907	4
axo20132.r-s-at	axo20953.f-s-at	102.3131313	3
axo04593.f-at	axo24311.r-at	102.2703743	5
axo14000.f-at	axo23668.r-at	102.1319253	1
axo16023.f-at	axo25759.f-at	101.6650936	1
axo12142.r-at	axo13902.r-at	101.2817618	1
axo03045.r-at	axo24849.f-at	101.236119	2
axo11622.f-at	axo21686.r-at	101.2080522	2
axo00972.f-at	axo04216.f-at	100.4603306	1
axo00399.r-at	axo13925.f-at	99.62461168	3
axo07400.f-at	axo08034.f-at	98.82551802	2
axo01966.f-at	axo15818.r-at	98.51652777	1
axo07032.f-at	axo28465.f-at	98.48149182	1
axo27108.f-at	axo27798.f-at	98.0131651	2
axo15468.f-at	axo27983.f-at	97.50714763	1

Table A.21: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo07069.r-at	axo19805.f-at	97.41828753	3
axo08017.f-at	axo25277.f-at	97.17923174	1
axo12690.r-at	axo30475.f-at	96.8666934	2
axo13034.f-at	axo25503.f-at	96.74368195	1
axo09995.f-at	axo19620.f-at	96.74096386	2
axo05076.f-at	axo19479.r-at	96.12591586	7
axo18263.f-at	axo20173.f-at	96.11777995	1
axo02114.f-at	axo16103.r-at	96.07745844	3
axo08030.f-at	axo17981.f-at	96.02669394	4
axo08549.f-at	axo31209.f-at	95.66616582	2
axo08871.f-at	axo28223.f-at	95.59907654	4
axo08423.f-at	axo10904.r-at	95.59214178	4
axo10452.f-at	axo29748.f-at	95.45855498	1
axo15163.f-at	axo25476.f-at	95.09678564	1
axo19284.f-at	axo29361.f-at	94.83665064	1
axo17347.r-at	axo27632.f-at	94.36378141	2
axo06027.f-x-at	axo09574.r-at	94.00099222	1
axo09003.f-at	axo18389.f-at	93.98061127	2
axo07796.f-at	axo08536.f-at	93.66484946	3
axo01534.f-at	axo08961.f-at	93.62968896	3
axo01441.f-at	axo02569.f-at	93.3066652	2
axo24220.r-at	axo28002.f-at	93.2635988	1
axo03117.f-at	axo25147.f-at	92.82980668	3
axo17337.r-at	axo17562.r-at	92.7046391	3
axo16596.f-at	axo27355.f-at	92.69061186	2
axo07620.f-at	axo28774.f-at	92.09946675	2
axo05543.f-at	axo18680.f-at	92.09177078	1
axo07851.f-at	axo18771.f-at	91.90281898	1
axo12455.r-at	axo13430.r-at	91.61577594	2
axo11598.f-at	axo31734.f-at	91.26869839	1
axo00090.r-at	axo23663.f-at	91.06737519	2
axo13343.f-at	axo16372.r-s-at	90.81686952	3
axo11041.f-at	axo13362.f-at	90.61648342	1
axo05905.f-at	axo29759.f-at	89.74426157	1
axo02996.r-at	axo28476.f-at	89.66720727	1
axo08459.f-at	axo20503.f-at	89.51436936	2
axo12885.f-at	axo12949.f-at	88.98089763	2
axo07923.f-at	axo30556.f-at	88.88653921	1
axo07324.f-at	axo13081.f-at	88.25730485	3
axo12918.f-at	axo14462.f-at	87.47141514	3
axo07326.r-at	axo14488.f-at	87.12812029	1
axo28851.f-at	axo30951.f-at	86.77923919	2
axo19369.r-at	axo27337.f-at	86.24890652	2
axo14587.f-at	axo19911.f-at	86.22672964	1
axo03314.f-at	axo04143.r-at	85.90263994	5
axo03123.f-at	axo16970.f-at	85.90135488	2

Table A.22: This table continues to show all 905 feasible solutions identified by FSA. Columns 1 and 2 show the probes that are identified, column 3 shows shows the $\hat{B}$ associated with each model, and column 4 how many times each feasible solution was chosen by FSA.

Variable 1	Variable 2	$\hat{B}$	Times Chosen by FSA
axo26036.f-at	axo29867.f-at	85.44844019	3
axo03745.f-at	axo15105.f-at	85.25672777	1
axo11754.f-at	axo23944.r-at	85.02504338	1
axo08295.f-at	axo29747.f-at	84.77364146	1
axo12772.f-at	axo23742.f-at	84.54500867	1
axo09075.f-at	axo19683.f-at	83.60864799	1
axo09502.f-at	axo29817.f-at	82.84489861	1
axo27289.f-at	axo28392.f-at	82.8332089	2
axo25606.f-at	axo30494.f-at	82.82974948	1
axo09692.f-at	axo24068.r-at	82.72023341	1
axo05355.f-at	axo16558.f-at	82.69510279	1
axo12847.f-at	axo14302.f-at	82.06501169	2
axo08891.f-at	axo18297.f-at	81.79876807	1
axo17431.r-at	axo29391.f-at	81.33041481	1
axo14848.f-at	axo16735.f-at	81.18223756	2
axo15826.f-at	axo26035.f-at	80.9893372	2
axo05434.r-at	axo08514.f-at	80.91770973	1
axo16399.f-at	axo31415.f-s-at	80.89970901	2
axo12794.r-at	axo25295.f-at	80.88948992	2
axo18714.f-at	axo18829.r-at	80.75414651	2
axo05185.f-at	axo28558.f-at	80.6664705	2
axo02837.f-at	axo20761.f-at	80.4671663	1
axo04774.f-at	axo05457.r-at	80.29728568	1
axo15056.r-at	axo24329.f-at	80.28899549	2
axo10577.f-at	axo13812.r-at	80.09843212	2
axo09056.f-at	axo09125.f-at	79.95596502	2
axo07087.f-at	axo19449.f-at	79.16115355	1
axo01331.f-at	axo30404.f-at	78.36089286	3
axo12463.f-at	axo31559.f-at	77.89786953	1
axo04155.r-at	axo22042.f-at	76.54125612	1
axo01998.f-s-at	axo14872.f-at	75.86602209	1
axo05109.r-at	axo12157.f-at	74.61523726	1
axo18600.r-at	axo30421.f-at	74.15059386	1
axo10886.f-at	axo20333.r-at	73.50469254	1
axo04273.f-at	axo24644.f-at	72.86632492	1
axo15099.f-at	axo30612.f-at	71.66238706	2
axo15621.f-at	axo16264.f-at	71.05323531	1
axo01367.r-at	axo26838.f-at	70.31903075	1
axo10215.r-at	axo13719.f-at	70.16317652	1
axo13437.r-at	axo14009.r-at	66.05492158	1
axo27220.f-at	axo31652.f-at	65.29004494	1
axo03780.f-x-at	axo15221.f-at	64.56296277	2
axo16952.f-at	axo24714.f-at	63.72813804	1
axo09446.r-at	axo20694.f-at	59.72850411	1
axo06677.f-at	axo19827.r-at	57.92164383	1

A.2 Figures

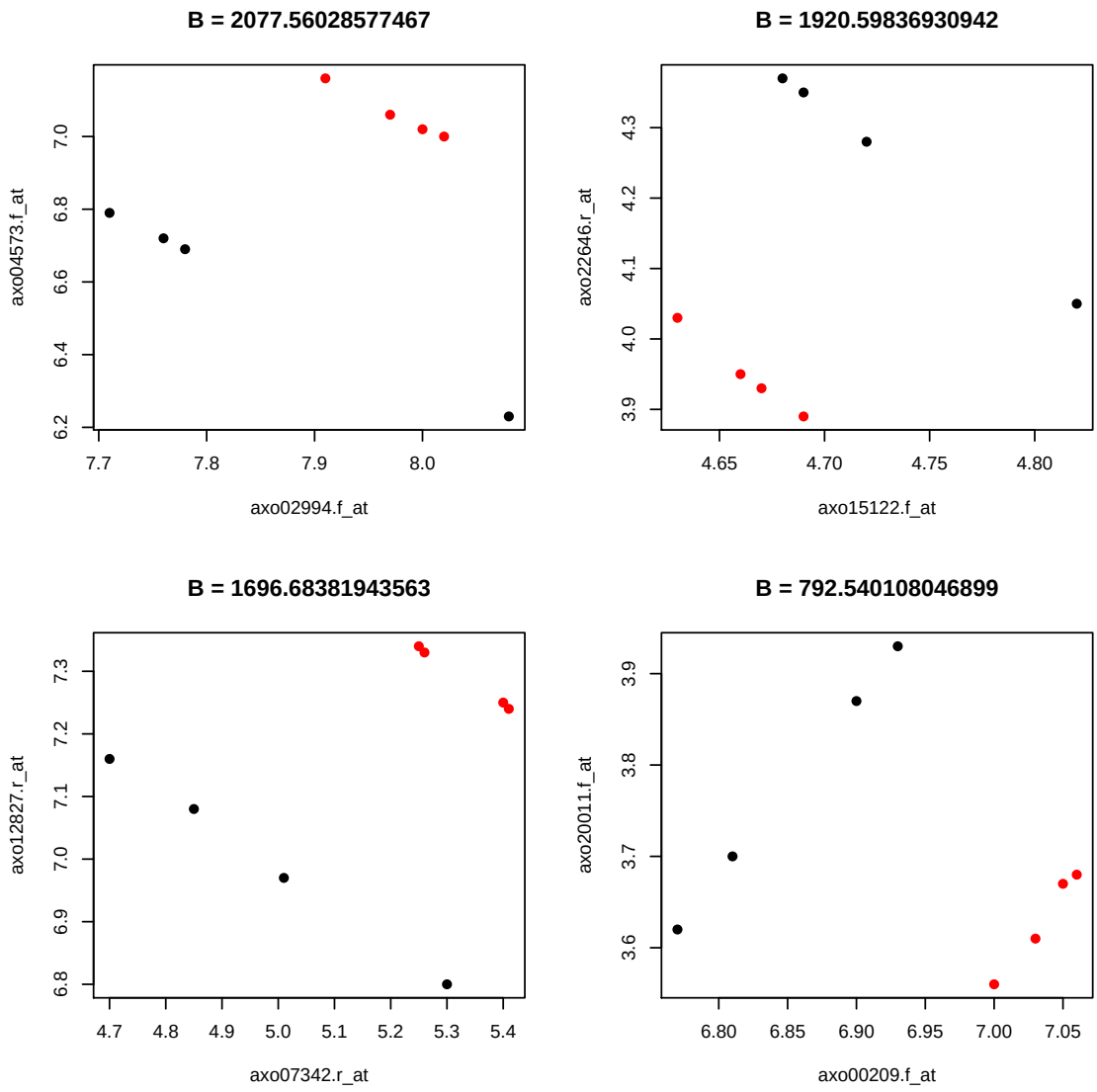


Figure A.1: These plots are of some of the most interesting combinations of predictors that resulted in a significant permutation p-value from B-distance is seen here. Black dots denote observations from the regeneration group and red dots denote observations from the non-regeneration group. Notice that the two response groups can be perfectly separated by these two predictors. This is a case where logistic regression would fail, but B-distance can be calculated. Univariate t-tests would also fail to identify either of these predictors as significant at the 0.01 level, but clearly when combined, the two predictors provide valuable information about which observations correspond to regeneration or non-regeneration

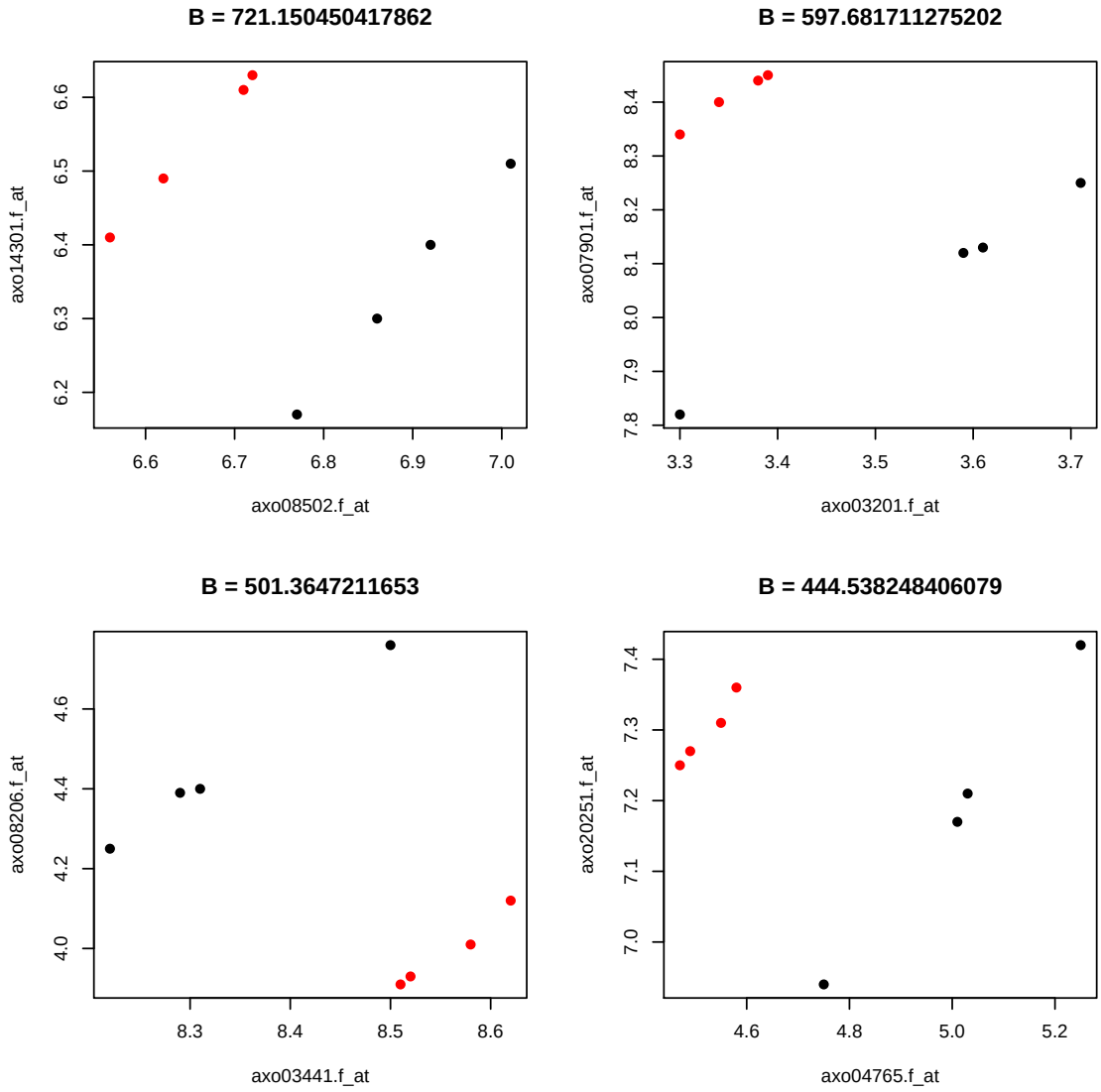


Figure A.2: These plots are of some of the most interesting combinations of predictors that resulted in a significant permutation p-value from B-distance is seen here. Black dots denote observations from the regeneration group and red dots denote observations from the non-regeneration group. Notice that the two response groups can be perfectly separated by these two predictors. This is a case where logistic regression would fail, but B-distance can be calculated. Univariate t-tests would also fail to identify either of these predictors as significant at the 0.01 level, but clearly when combined, the two predictors provide valuable information about which observations correspond to regeneration or non-regeneration

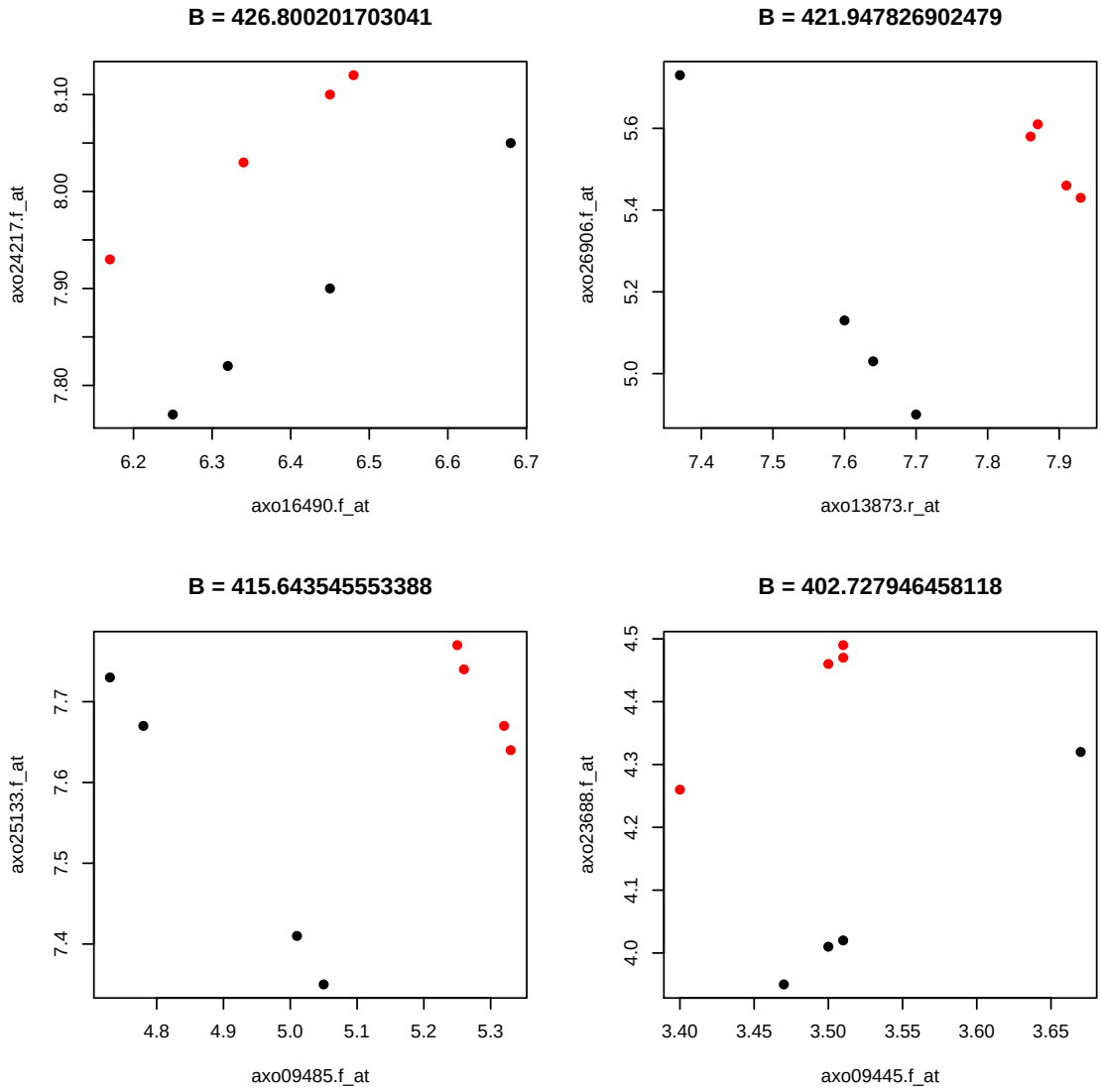
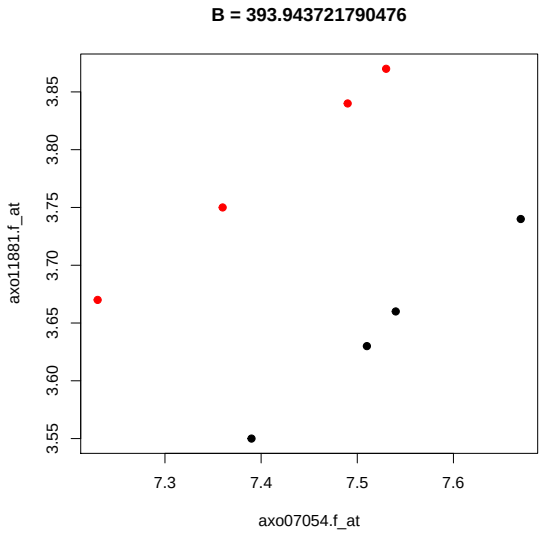


Figure A.3: These plots are of some of the most interesting combinations of predictors that resulted in a significant permutation p-value from B-distance is seen here. Black dots denote observations from the regeneration group and red dots denote observations from the non-regeneration group. Notice that the two response groups can be perfectly separated by these two predictors. This is a case where logistic regression would fail, but B-distance can be calculated. Univariate t-tests would also fail to identify either of these predictors as significant at the 0.01 level, but clearly when combined, the two predictors provide valuable information about which observations correspond to regeneration or non-regeneration



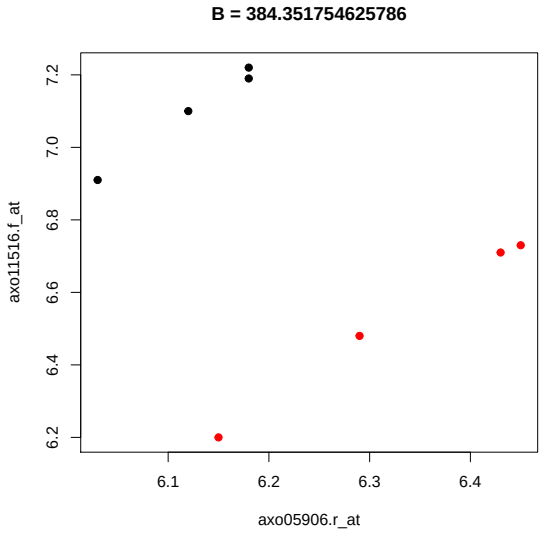


Figure A.4: These plots are of some of the most interesting combinations of predictors that resulted in a significant permutation p-value from B-distance is seen here. Black dots denote observations from the regeneration group and red dots denote observations from the non-regeneration group. Notice that the two response groups can be perfectly separated by these two predictors. This is a case where logistic regression would fail, but B-distance can be calculated. Univariate t-tests would also fail to identify either of these predictors as significant at the 0.01 level, but clearly when combined, the two predictors provide valuable information about which observations correspond to regeneration or non-regeneration

Bibliography

V Alba-Fernández, J Muñoz-García, and María Dolores Jiménez-Gamero. Bootstrap estimation of the distribution of matusita distance in the mixed case. *Statistics & probability letters*, 73(3): 277–285, 2005.

Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

Steven F. Arnold. *The theory of linear models and multivariate analysis*. Wiley series in probability and mathematical statistics. Wiley, New York, 1981. ISBN 0471050652.

Rudolf Beran. Minimum hellinger distance estimates for parametric models. *The Annals of Statistics*, pages 445–463, 1977.

Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distribution. *Bull. Calcutta Math. Soc*, 1943.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

T Tony Cai, Tengyuan Liang, and Harrison H Zhou. Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions. *Journal of Multivariate Analysis*, 137:161–172, 2015.

Aparna Chattopadhyay, Asis Kumar Chattopadhyay, and B-Rao Chandrika. Bhattacharyyas distance measure as a precursor of genetic distance measures. *Journal of biosciences*, 29(2):135, 2004.

Euisun Choi and Chulhee Lee. Feature extraction based on the bhattacharyya distance. *Pattern Recognition*, 36(8):1703–1709, 2003.

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.

I. Csiszar. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964. URL `http://ci.nii.ac.jp/naid/10006737982/en/`.

Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.

David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4), 2009.

Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 2013.

Paul J Gemperline. Computation of the range of feasible solutions in self-modeling curve resolution algorithms. *Analytical chemistry*, 71(23):5398–5404, 1999.

François Goudail, Philippe Réfrégier, and Guillaume Delyon. Bhattacharyya distance as a contrast parameter for statistical processing of noisy optical images. *JOSA A*, 21(7):1231–1240, 2004.

Benjamin Goudey, Mani Abedini, John L Hopper, Michael Inouye, Enes Makalic, Daniel F Schmidt, John Wagner, Zeyu Zhou, Justin Zobel, and Matthias Reumann. High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in genome wide association studies. *Health Information Science and Systems*, 3(1):1, 2015.

Xuan Guorong, Chai Peiqi, and Wu Minhui. Bhattacharyya distance feature selection. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, volume 2, pages 195–199. IEEE, 1996.

Douglas M Hawkins. A feasible solution algorithm for the minimum volume ellipsoid estimator in multivariate data. *Computational Statistics*, 8:95–95, 1993a.

Douglas M Hawkins. The feasible set algorithm for least median of squares regression. *Computational Statistics & Data Analysis*, 16(1):81–101, 1993b.

Douglas M Hawkins. The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics & Data Analysis*, 17(2):197–210, 1994a.

Douglas M Hawkins. The feasible solution algorithm for least trimmed squares regression. *Computational statistics & data analysis*, 17(2):185–196, 1994b.

Douglas M Hawkins and David J Olive. Improved feasible solution algorithms for high breakdown estimation. *Computational statistics & data analysis*, 30(1):1–11, 1999.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Thomas Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60, 1967.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Joshua Lambert. *rFSA: rFSA implements a Feasible Solution Algorithm (FSA) to optimal models of a specific form that include mth order interactions.*, 2015. R package version 1.0.

Thomas Lumley and Alan Miller. Leaps: regression subset selection. *R package version*, 2, 2004.

Li Ma, Andrew G Clark, and Alon Keinan. Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet*, 9(2):e1003321, 2013.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Prasanta Chandra Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India, 1936*, pages 49–55, 1936.

Brian Mak and Etienne Barnard. Phone clustering using the bhattacharyya distance. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2005–2008. IEEE, 1996.

Alan J Miller. Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, pages 389–425, 1984.

Jason H Moore and Scott M Williams. Epistasis and its implications for personal genetics. *The American Journal of Human Genetics*, 85(3):309–320, 2009.

Tetsuzo Morimoto. Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, 1963.

John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

Mikhail S Nikulin. Hellinger distance. *Encyclopedia of mathematics*, 151, 2001.

Dale Poirier. Jeffreys' prior for logit models. *Journal of Econometrics*, 63(2):327–339, 1994.

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set

for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

Carlisle Rainey. Dealing with separation in logistic regression models. *Political Analysis*, 24(3): 339–355, 2016.

Nalini Ravishanker and Dipak K Dey. *A first course in linear model theory*. CRC Press, 2001.

S Ray. On a theoretical property of the bhattacharyya coefficient as a feature evaluation criterion. *Pattern recognition letters*, 9(5):315–319, 1989.

Constantino Carlos Reyes-Aldasoro and Abhir Bhalerao. The bhattacharyya space for feature selection and its application to texture segmentation. *Pattern Recognition*, 39(5):812–826, 2006.

Peter J Rousseeuw and Andreas Christmann. Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*, 43(3):315–332, 2003.

Daniel F Schwarz, Inke R König, and Andreas Ziegler. On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics*, 26(14):1752–1758, 2010.

Fred C Schweppe. On the bhattacharyya distance and the divergence between gaussian processes. *Information and Control*, 11(4):373–395, 1967.

Zhiguang Su, Naoki Ishimori, Yaoyu Chen, Edward H Leiter, Gary A Churchill, Beverly Paigen, and Ioannis M Stylianou. Four additional mouse crosses improve the lipid qtl landscape and identify lipg as a qtl gene. *Journal of lipid research*, 50(10):2083–2094, 2009.

Xue W Tian and Joon S Lim. Bhattacharyya distance for identifying differentially expressed genes in colon gene experiments. In *Information Science and Applications (ICISA), 2013 International Conference on*, pages 1–2. IEEE, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Xiang Wan, Can Yang, Qiang Yang, Hong Xue, Xiaodan Fan, Nelson LS Tang, and Weichuan Yu. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87(3):325–340, 2010.

Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2011.

Naomi Wray and P Visscher. Estimating trait heritability. *Nature Education*, 1(1):29, 2008.

Chang Huai You, Kong Aik Lee, and Haizhou Li. An svm kernel with gmm-supervector based on the bhattacharyya distance for speaker recognition. *IEEE Signal processing letters*, 16(1):49–52, 2009.

Ling Sing Yung, Can Yang, Xiang Wan, and Weichuan Yu. Gboost: a gpu-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics*, 27(9):1309–1310, 2011.

Weidong Zhang, Ron Korstanje, Jill Thaisz, Frank Staedtler, Nicole Harttman, Lingfei Xu, Minjie Feng, Liane Yanas, Hyuna Yang, William Valdar, et al. Genome-wide association mapping of quantitative traits in outbred mice. *G3: Genes— Genomes— Genetics*, 2(2):167–174, 2012.

Christopher Zorn. A solution to separation in binary response models. *Political Analysis*, 13(2): 157–170, 2005.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

Vita

- Sarah A. Janse, Lexington, Kentucky
- Education
  - M.S. in Statistics: University of Kentucky, Lexington, KY 2014.
  - B.S. in Mathematics: Roanoke College, Salem, VA 2012.
- Professional Experience
  - Graduate Research Assistant, Department of Statistics, University of Kentucky, 2014-2017.
  - Graduate Teaching Assistant, Department of Statistics, University of Kentucky, 2012-2014.
- Scholastic Honors
  - Boyd Hershberger Travel Award, provided by National Science Foundation to attend SRCOS, 2016.
- Publications
  - Barbara Jones, Joey Clark, Jeffrey Bewley, Kristen McQuerry, **Sarah Janse**. Controlling Digital Dermatitis: Copper Sodium Hypochlorite Versus Copper Sulfate Footbath, *Journal of Dairy and Veterinary Sciences*, 3(1):555602, 2107.
  - Peter T. Nelson, Wang-Xia Wang, **Sarah A. Janse**, and Katherine L. Thompson. 2017. MicroRNA expression patterns in human anterior cingulate and motor cortex: a study of dementia with Lewy bodies and controls*Submitted to Brain Research.*
  - **Sarah A. Janse** and Katherine L. Thompson. 2017. A probabilistic bound on the number of iterations of stochastic algorithms, *Submitted to Probability and Statistics Letters.*
  - Mark A Williams, John G Strang, Ricardo T Bessin, Derek Law, Delia Scott, Neil Wilson, **Sarah Witt**, and Douglas D Archbold. An assessment of organic apple production in kentucky. *HortTechnology*, 25(2):154-161, 2015.