12-2017

Extensions and Improvements to Random Forests for Classification

Anna Quach
Utah State University

EXTENSIONS AND IMPROVEMENTS TO RANDOM FORESTS FOR CLASSIFICATION

by

Anna Quach

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematical Sciences

Approved:

Adele Cutler, Ph.D. Major Professor	Christopher Corcoran, Ph.D. Committee Member
Richard Cutler, Ph.D. Committee Member	Jürgen Symanzik, Ph.D. Committee Member
Kyumin Lee, Ph.D. Committee Member	Mark R. McLellan, Ph.D. Vice President for Research Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2017

ABSTRACT

Extensions and Improvements to Random

Forests for Classification

by

Anna Quach, Doctorate of Philosophy

Utah State University, 2017

Major Professor: Dr. Adele Cutler
Department: Mathematics and Statistics

The motivation of my dissertation is to improve two weaknesses of Random Forests. One, the failure to detect genetic interactions in higher dimensions when the interacting genes both have weak main effects and two, the difficulty of interpretation in comparison to parametric methods such as logistic regression, linear discriminant analysis, and linear regression.

There are approximately 10 million single nucleotide polymorphisms (SNPs) in the whole genome of the human species. We focus on detecting pairwise SNP interactions in genome case-control studies. Analyzing SNP-SNP interactions is computationally and methodologically challenging because the search space for k-way interactions of d SNPs is $\binom{d}{k}$. We present an efficient filtering method and compare it to leading methods. We show that our new filtering method is computationally faster with good detection power.

One of the advantages Random Forests has over statistical methods is its capability of handling data sets when $d >> n$, where n is the number of observations. It is common to use Random Forests as a filtering technique to reduce the number of predictors. We determine the best parameter settings to optimize the detection of SNP interactions and improve the efficiency of Random Forests specifically for data from the Genome-Wide Association Studies.

Random Forests allows us to identify clusters, outliers, and important features for subgroups of observations through the visualization of the proximities. The old implementation of Random Forests uses a multidimensional scaling plot to visualize the symmetric proximities. We improve the interpretation of Random Forests through the proximities. The result of the new proximities are asymmetric and reproduce the predictions in Random Forests. The appropriate visualization of the new proximities requires an asymmetric model for interpretation. We propose a new visualization technique for asymmetric data and compare it to existing approaches.

(90 pages)

PUBLIC ABSTRACT

Extensions and Improvements to Random

Forests for Classification

Anna Quach

The motivation of my dissertation is to improve two weaknesses of Random Forests. One, the failure to detect genetic interactions between two single nucleotide polymorphisms (SNPs) in higher dimensions when the interacting SNPs both have weak main effects and two, the difficulty of interpretation in comparison to parametric methods such as logistic regression, linear discriminant analysis, and linear regression.

We focus on detecting pairwise SNP interactions in genome case-control studies. We determine the best parameter settings to optimize the detection of SNP interactions and improve the efficiency of Random Forests and present an efficient filtering method. The filtering method is compared to leading methods and is shown that it is computationally faster with good detection power.

Random Forests allows us to identify clusters, outliers, and important features for subgroups of observations through the visualization of the proximities. We improve the interpretation of Random Forests through the proximities. The result of the new proximities are asymmetric, and the appropriate visualization requires an asymmetric model for interpretation. We propose a new visualization technique for asymmetric data and compare it to existing approaches.

This work is dedicated to my parents.

ACKNOWLEDGMENTS

I could not thank my major professor, Dr. Adele Cutler enough for all the knowledge she passed on to me. I am indebted to her for her time in providing constant help and guidance. I would not be the Statistician I am today without her. She has been a great mentor and a great friend.

I couldn't be more grateful for the opportunity and valuable skills I've gained from working on Dr. Heidi Wengreen's Nutrition project, Dr. Guifang Fu's leaf project, and working with Dr. Jürgen Symanzik on a data visualization competition.

I would like to thank my committee members: Dr. Christopher Corcoran, Dr. Richard Cutler, Dr. Jürgen Symanzik, and Dr. Kyumin Lee for their support and great suggestions. Many thanks go to the Mathematics and Statistics Department at Utah State University, to all professors, students, and staff members. Last, I would like to thank my family for always being there for me, and my friends for all the memorable times and laughs we've shared over the years.

Anna Quach

CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Introduction

In this dissertation, we dive deep into some types of problems for which Random Forests are shown to do poorly and improve the interpretation, a perceived weakness of Random Forests compared to traditional models such as linear regression, linear discriminant analysis, and logistic regression. We propose a new filtering method to detect genetic interactions, optimize Random Forests for detecting genetic interactions, and extend the interpretation of Random Forests by introducing a new visualization method.

The remainder of the introduction presents the background and literature reviews for interaction detection in Genome-Wide Association (GWA) studies, and for Random Forests.

1.2 Genome-Wide Association Study Literature Review and Background

GWA studies are a powerful tool used to identify genetic variants that are associated with diseases. Variation at a single base pair is a Single Nucleotide Polymorphism (SNP). It is estimated that there are approximately 10 million SNPs on the whole genome of the human species (National Library of Medicine, 2017). Due to increasing evidence that individual SNPs only explain a portion of the genetic causes of a disease, researchers have started analyzing SNPs at two loci (the specific location of a SNP on a chromosome) or more. k-SNP (k$\geq$ 2) interactions are also known as epistasis or epistatic interactions. Analyzing gene-gene or SNP-SNP statistical interactions is a challenging task because the search space for k-way interactions of d SNPs is $\binom{d}{k}$. Therefore there is a need for fast implementations to detect epistatic interactions.

1.2.1 Problem Statement

GWA studies have allowed researchers to investigate the association of diseases (phenotype data) and SNPs (genotype data containing genetic information of an individual). The genotype data that we will be observing are bi-allelic SNPs with a major allele denoted with a capital letter, A, and a minor allele represented with a lower case letter, a. There are three possible genotypes: AA, Aa (or aA), and aa. AA, Aa, and aa are typically coded as values 0, 1, 2, respectively and cases (the disease group) are usually coded as 1's and controls as 0's.

Determining which interacting SNPs are causal of a disease is a $d >> N$ problem, where d is the number of SNPs, and N is the number of observations. It's approximated that there are about 10 million SNPs in the whole genome. Due to a phenomenon called linkage disequilibrium, the nonrandom association of alleles at two or more loci (Slatkin, 2008), we don't have to look at every single SNP. Statistically speaking, linkage disequilibrium is the correlation that occurs between two SNPs. However, the number of single SNPs remains quite large and obtaining enough cases and controls to be part of a study is a challenge.

The number of possible 2-way interactions of d SNPs is $\frac{d\times(d-1)}{2}$, which creates a computational challenge in detecting real causal SNP-SNP interactions. Many existing methods are unable to handle the enormous number of possible combinations of SNPs. It would take years for some of the methods to complete.

Quite a few methods carry out a statistical test to determine the significance of an epistatic interaction. Handling an immense number of possible interacting SNPs creates the problem of controlling the number of false positives while retaining high detection power. Therefore developing an efficient and effective method to detect epistatic interactions is desirable.

1.2.2 Definition of Interaction

The exact definition of an interaction is debatable. Most researchers are familiar with an interaction in a model-based sense, such as, in linear regression or logistic regression. Fisher (1918) defined an interaction as two or more variables that deviate from additive effects. For example, the logistic regression model has the following form:

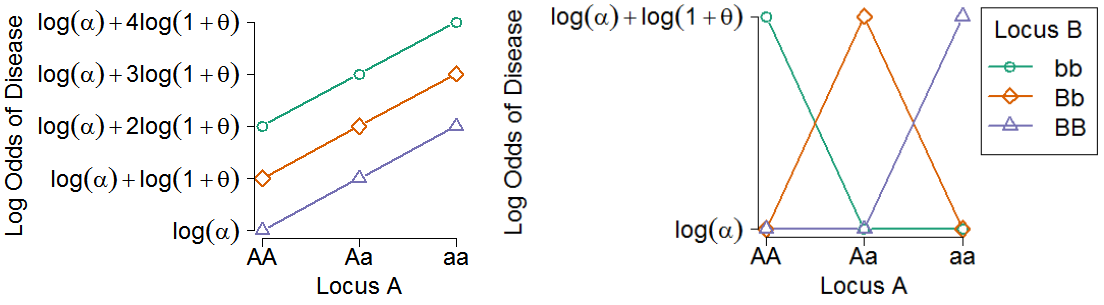
$$log\left(\frac{P(Y=1)}{1-P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 \tag{1.1}$$

where X_1X_2 is the interaction term. A 4 degrees-of-freedom (df) test on the difference in the log-likelihood of 1.1 and the log-likelihood of:

$$log\left(\frac{P(Y=1)}{1-P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{1.2}$$

is a test of the interaction (Wan et al., 2010a). The interaction of two variables can be determined by referencing an interaction plot (see Figure 1.1). In Figure 1.1a, both SNPs at locus A and B have a main effect which is apparent in the plot since the odds of disease increases for every additional copy of the minor allele. The slope for each genotype at locus B is the same. The SNPs are not interacting because the change in log odds of disease for every additional copy of the minor allele in A does not depend on the genotypes at locus B. An illustration of two SNPs interacting can be seen in Figure 1.1b.

Some of the standard approaches to detect epistatic interactions include logistic regression, chi-square tests, and permutation tests. The problem with traditional methods is the potential increase in Type I error when statistical tests are performed multiple times. The gold standard is to use permutation testing to control the number of false positives. The trade-off to using permutation testing is the computational burden. Other correction methods, such as Bonferroni are preferred to make the process more computationally feasible. However, Bonferroni is known to be extremely conservative and can miss causal SNPs.



(A) (B)

FIGURE 1.1: The interaction plot in (A) is an example of when two SNPs are not interacting (lines are parallel) and (B) is an example of two SNPs interacting (lines are not parallel).

TABLE 1.1: The two-way table for the control group, $Y = 0$, is on the left and the two-way table for the case group, $Y = 1$, is on the right. The rows represent the genotypes at Locus A and the columns represent the genotypes at Locus B. N_{rc} and D_{rc} for $r, c \in 1, 2, 3$ are the number of observations for each combination of genotypes.

$Y = 0$	BB	Bb	bb
AA	N_{11}	N_{12}	N_{13}
Aa	N_{21}	N_{22}	N_{23}
aa	N_{31}	N_{32}	N_{33}

$Y = 1$	BB	Bb	bb
AA	D_{11}	D_{12}	D_{13}
Aa	D_{21}	D_{22}	D_{23}
aa	D_{31}	D_{32}	D_{33}

TABLE 1.2: Table 1.1 is reconstructed to test if the distribution of the combination of genotypes at Locus A and B differ between the cases and controls.

	$Y = 0$	$Y = 1$
AA,BB	N_{11}	D_{11}
AA,Bb	N_{12}	D_{12}
AA,bb	N_{13}	D_{13}
Aa,BB	N_{21}	D_{21}
Aa,Bb	N_{22}	D_{22}
Aa,bb	N_{23}	D_{23}
aa,BB	N_{31}	D_{31}
aa,Bb	N_{32}	D_{32}
aa,bb	N_{33}	D_{33}

1.2.3 Methods to Detect Epistatic Interactions

Chi-square tests are commonly used as an exhaustive search in two-locus association studies. Typically an 8 df versus a 4 df chi-square test is used. See Tables 1.1 and 1.2. FastChi (Zhang et al., 2009) presents an efficient approach to using the standard test and Chi8 (Al-jouie et al., 2015) uses Graphics Processing Units (GPUs) to calculate χ^2 for all pairwise SNPs. Chatterjee et al. (2006) proposed a multi-way genetic interaction test using Tukey's 1 df model of interaction. BOOST (BOolean Operation-based Screening and Testing) fits a log-linear model and uses a 4 df χ^2 test to check for two-way interactions after the screening stage.

Machine learning methods, such as multifactor dimensionality reduction (MDR), Random Forests (RF), Neural Networks (NN), and Support Vector Machines (SVM) are alternatives to traditional statistical approaches for detecting SNP-SNP interactions. Upstill-Goddard et al. (2012) reviewed early machine learning approaches such

as multifactor-dimensionality reduction (MDR), neural networks, Random Forests, support vector machines, and more recent models such as extensions to MDR and Random Forests for detecting SNP-SNP interactions. They listed the strengths and limitations of the early and recent Machine Learning methods. An attractive feature of using machine learning algorithms is their ability to handle high-dimensional data. The downfalls to using the machine learning algorithms, compared to using an exhaustive search, e.g. Pearson χ^2 test, are computational time and lack of protection against false detection.

A possible approach to the overall problem is to conduct a two-stage method where the first stage reduces the number of SNPs or pairs of SNPs and the second stage uses an existing machine learning algorithm to rank them. However, some researchers use machine learning algorithms as a feature screening tool before applying an existing method, or they combine machine learning algorithms to identify interactions (Lin et al., 2012; De Lobel et al., 2010).

Tree-based approaches are known to do well in detecting SNP-SNP interactions compared to some existing methods (Goldstein et al., 2010). However, tree-based methods can fail to detect a joint effect if there are no strong main effect (Winham et al., 2012). Failure to detect a joint effect can happen because, at each node, a tree chooses the single variable with the best split, so it is unlikely to choose a variable without a strong individual effect. In Random Forests this is somewhat ameliorated by the random choice of variables at each node because Random Forests can choose variables that might not have strong individual effects.

There are only a few review papers that make an in-depth comparison of the methods using simulated data. To the best of our knowledge, there are only two, Wang et al. (2011) and Shang et al. (2011).

Wang et al. (2011) evaluate five different methods: TEAM (Zhang et al., 2010), BOOST (Wan et al., 2010a), SNPRuler (Wan et al., 2010b), SNPHarvester (Yang et al., 2009), and Screen and Clean (Wu et al., 2010) using simulated data with and without a main effect. The methods are evaluated by detection power, type-1 error rate, scalability, and completeness. Overall, BOOST and TEAM are the two methods recommended if computational cost is not of concern and users want powerful results.

BOOST performs the best in selecting interactions without main effects, was the fastest with 100, 1000, and 10000 SNPs with 2000 observations in each data set, and never wrongly prunes the most significant SNP pairs in models with or without main effects. Wang et al. (2011) found that BOOST had the highest type I error rate and

may not be able to detect interactions when there is a weak interaction effect, but the single SNP association term fits the model well. TEAM performed the best on data with main effects and was second best in detecting interaction without main effects. However, TEAM had the second largest Type I error rate. The high error rate may be due to TEAM having higher statistical power.

Shang et al. (2011) identified 36 different methods, excluding tweaked and specialized methods, and categorized the methods according to three different search strategies, i.e., exhaustive, stochastic, and heuristic. Shang et al. (2011) choose to compare five representative methods from the categories: TEAM, BOOST, SNPRuler, AntEpiSeeker (Wang et al., 2010) and epiMode (Tang et al., 2009) using simulated data of different sizes, nine commonly used epistasis models, and data with and without noise.

The types of noise simulated are due to missing data, genotyping error, and phenocopy. The five methods are compared by detection power, robustness, sensitivity, and computational complexity. Overall, Shang et al. (2011) recommended AntEpiSeeker and BOOST as the most efficient and effective methods.

AntEpiSeeker performed the best on detecting epistatic interactions with marginal effects and had good performance on models with no main effects. In comparison to other methods, AntEpiSeeker has far better robustness to all types of noise on marginal effect models, has good detection power on models with no marginal effect, is able to perform well on detecting multiple epistasis for models with main effects, can handle large-scale data sets in a reasonable amount of time, and is able to deal with higher order models. AntEpiSeeker was found to be sensitive to SNPs with a strong association with the phenotype.

BOOST performed the best on identifying epistatic interactions with no marginal effects. The detection power of models with no main effects was much higher than models with main effects. BOOST was robust to genotyping error and phenocopy on models with no marginal effects, is the fastest among the methods they compared, and can detect multiple epistatic interactions based on a sensitivity analysis. BOOST could not be evaluated on how well it performed with SNPs with missing values since it removes any SNPs with missing values. Shang et al. (2011) found that it is more sensitive to model type compared to SNPRuler based on detection power analysis, it is sensitive to sample size and SNP number based on a sensitivity analysis, and is limited to detecting two-way interactions.

1.2.4 Data Simulation Parameters

There are many different disease models that simulated data can come from, and those data vary depending on various parameter settings. Some of the possible effects that can be taken into account are:

1. Main Effect (Marginal Effects) - the strength of single SNP association.
2. Prevalence - the proportion of a population with the disease.
3. Minor Allele Frequency (MAF) - second most frequent occurring allele in a given population.
4. Heritability - the proportion of variation in the phenotype explained by genetic variation.
5. Linkage Disequilibrium - the correlation that occurs between two SNPs.

There are existing simulation tools that can simulate epistatic interactions for case-control association studies. Shang et al. (2011) provides a tool, called epiSIM, offering simulations of single-locus and epistasis models associated with the phenotype. Wang et al. (2011) uses simulated data (data without main effect) provided by Dartmouth Medical School at `http://discovery.dartmouth.edu/epistatic_data/`. The website provides 70 different models composed of combinations of two MAF settings (0.2, 0.4), seven heritability settings (0.01, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4), and five different tables in terms of probability of disease for a given genotype. Each model is simulated using four different samples size of 200, 400, 800, and 1600. A thousand SNPs were used for each data set.

Others have simulated their data using different models. Guo et al. (2014b) used a multiplicative model, an epistasis model that has been used to describe handedness and the color of swine, a classical epistasis model, a well known XOR model, and a three-locus epistasis model. For data with the main effect, Wang et al. (2011) simulated data based on three common epistasis models. For each model, they used an MAF of 0.2 and 0.5, and three different main effect values of 0.2, 0.3, and 0.5. They used 2000 samples and 1000 SNPs for each data set. The data sets are available at `http://compbio.ddns.comp.nus.edu.sg/~wangyue/`. Shang et al. (2011) used nine commonly used two-locus epistasis models: three models displaying marginal effects and six models without marginal effects.

1.2.5 Simulation Model

Figures 1.2a, 1.2c, and 1.2e are the three most commonly used interaction models presented by Marchini et al. (2005). The values in the cells of the tables are the odds of getting a disease where α is the baseline risk and θ represents the increased risk of a disease allele (either a or b). Some report their disease model in terms of penetrance, $P(D|g_i)$ where D and D^C are the disease status. Penetrance is the probability an individual will carry a given genotype, g_i. The relationship between the odds and penetrance is shown in Equation 1.3 and Equation 1.4.

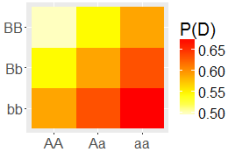
$$ODD_{g_i} = \frac{P(D|g_i)}{P(D^C|g_i)} = \frac{P(D|g_i)}{1 - P(D|g_i)} \tag{1.3}$$

$$P(D|g_i) = \frac{ODD_{g_i}}{1 + ODD_{g_i}}. \tag{1.4}$$

There are three possible genotypes for each SNP: the homozygous genotype (AA), the heterozygous reference genotype (Aa), and the homozygous variant genotype (aa). We

	BB	Bb	bb
AA	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
Aa	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^3$
aa	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^3$	$\alpha(1+\theta)^4$

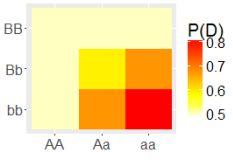
(A) Model Additive



(B)

	BB	Bb	bb
AA	α	α	α
Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
aa	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$

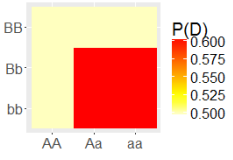
(C) Model Multiplicative



(D)

	BB	Bb	bb
AA	α	α	α
Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$

(E) Model Threshold



(F)

FIGURE 1.2: The two-way tables of two SNPs in (A), (C), and (E) are genetic models used to evaluate the performance of methods that can detect epistatic interactions. α is the baseline risk and θ is the increased risk for having a copy of the minor allele, a or b. The cells in the tables represents the odds of disease. (B), (D), and (F) are heatmaps of (A), (C), and (E), respectively in terms of probability of disease.

can determine what α and θ are by using the equation for marginal effect, λ and prevalence, p. Marginal effect and prevalence are defined as:

$$\lambda = \frac{P(D|Aa)/P(D|AA)}{P(D^C|Aa)/P(D^C|AA)} - 1 \tag{1.5}$$

$$p = \sum_i P(D|g_i) * P(g_i) \tag{1.6}$$

where A is the major allele and a is the minor allele of a SNP. See Sohn and Wee (2015) for a detailed derivation of Equation 1.5 and 1.6 and an explanation of the differences in disease models. Typical values of λ used in simulated models are 0.2, 0.3, and 0.5, typical values for prevalence are 0.005, 0.01, and 0.1, and typical settings for the MAF are 0.05, 0.1, 0.2, and 0.5 (Sohn and Wee, 2015).

Figures 1.2b, 1.2d and 1.2f illustrate the increased chance of disease for the additive 1.2a, multiplicative 1.2c, and threshold model 1.2e, respectively. The risk of disease increases when the disease allele, a or b occurs in the additive model, risk of disease increases when a pair of disease alleles across two loci occurs in the multiplicative model, and risks of disease increases for any existence of disease alleles across two loci in the threshold model (Sohn and Wee, 2015).

Figures 1.3a, 1.3b and 1.3c describe the existence of an interaction effect for the additive, multiplicative and threshold models, respectively.

1.3 Random Forests Literature Review and Background

Random Forests are tree-based classification and regression methods developed in 2001 by Leo Breiman (Breiman, 2001). Since then there have been 27,104 (4/4/2017)

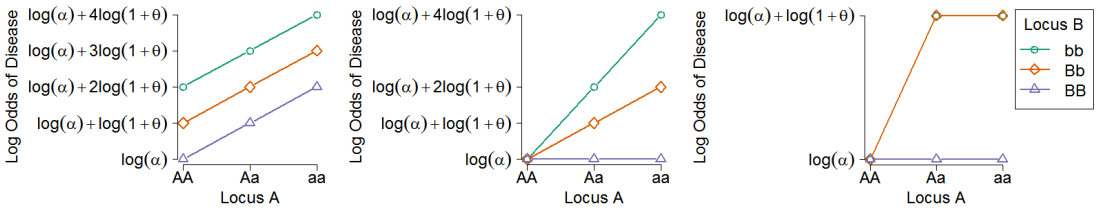


FIGURE 1.3: (A), (B), and (C) are interaction plots of the three models in Figure 1.2. The additive model (A) presents no interaction effect but both SNPs have main effects. The multiplicative model (B) presents both an interaction effect and main effects. The threshold model (C) has an interaction effect and slight main effects. In (C), lines for bb and Bb coincide.

citations (Google Scholar, 2017). Random Forests have been shown to do well in a wide variety of problems and have been applied in many disciplines. As a matter of fact, 179 classifiers implemented in Weka, R (R Core Team, 2015), C and Matlab were evaluated using 121 data sets (from the UC Irvine Machine Learning repository) by Fernández-Delgado et al. (2014) and Wainberg et al. (2016). Both groups found that overall Random Forests was one of the best classifiers in terms of accuracy.

Random Forests can be used for either regression or classification problems. In both cases, the goal is to predict the response variable using one or more predictor variables. More formally, suppose we have a training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ with N independent and identically distributed observations, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM})$ for $i \in 1, \ldots, N$. For regression problems the response variable is continuous, while for classification problems it is categorical, i.e. $y_i \in 1, \ldots, K$ for $i = 1, \ldots, N$ where K is the total number of classes.

Types of problems in which Random Forests have been used for classification are classifying images (Bosch et al., 2007), classifying invasive plant species (Cutler et al., 2007), and identifying SNPs associated with diseases (Goldstein et al., 2010) among many others. Regression in Random Forests has been used to: predict customer retention and profitability (Larivière and Van den Poel, 2005), and predict aqueous solubility (Palmer et al., 2007).

Random Forests is used as a screening tool to reduce the number unrelated predictor variables (Lunetta et al., 2004). Some of the extensions of Random Forests include Random Survival Forests (Ishwaran et al., 2008), Random Jungle (Schwarz et al., 2010), and Quantile Regression Forests (Meinshausen, 2006). For a more comprehensive review of the recent developments in Random Forests, both methodological and theoretical, see Biau and Scornet (2016).

The advantages of using Random Forests in comparison to other statistical methods are (Cutler et al., 2012):

1. High predictive accuracy.
2. Ability to handle continuous and categorical predictors.
3. Ability to handle more predictors than observations.
4. A novel method of determining variable importance.

5. Ability to model complex and unknown interactions among predictor variables.
6. Ability to handle correlated variables well.
7. Flexibility to perform several types of statistical data analysis including regression, classification, survival analysis, and unsupervised learning.
8. Ability to impute missing values.
9. Robustness to outliers in the predictor variables.
10. Insensitivity to monotone transformations of the predictor variables.
11. Scaling well for large sample sizes.
12. Dealing with irrelevant predictor variables.

1.3.1 Classification and Regression Trees

Before Leo Breiman developed Random Forests, he worked on classification and regression trees (CART) with Jerome H. Friedman, Charles J. Stone, and Richard A. Olshen (Breiman et al., 1984). The idea of Random Forests is built on the ideas of CART. CART are binary decision trees that can be used to predict categorical or continuous response variables. The tree initially contains all observations in the "root node." The node is split into two daughter nodes, which are themselves split and so forth.

In classification, the trees are grown until the nodes are pure (observations all belong to one class), there's only one case in the node, or a stopping criterion is met. The nodes at the bottom of the final tree are called "terminal nodes." The tree will most likely overfit the data set, i.e. the prediction error will be much larger on new data than it is on the data used to grow the tree. Therefore a pruning approach is used to find the optimal tree. The predicted values are determined by dropping each observation down the tree and computing the most frequent class in a terminal node for classification and by averaging the response for observations in the terminal node in regression.

A classification tree is built by initially determining the predictor variable that has the best split for separating the observations in the root node according to their class. A popular way to measure how well a potential split separates the observations is to use the Gini impurity, also known as the Gini index (Breiman et al., 1984). The Gini

impurity, G, for a node is calculated as:

$$G = \sum_k \sum_{l \neq k} p_k p_l = \sum_k p_k \sum_{l \neq k} p_l = \sum_k p_k(1 - p_k) \tag{1.7}$$

where p_k = proportion of class k in the node, and p_l = proportion of class l in the node. Note that in the two-class case $G = 2\hat{p_0}\hat{p_1}$ where $\hat{p_0}$ is the proportion of class 0 and $\hat{p_1}$ is the proportion of class 1 in the node. The variable that splits the set will minimize the splitting criteria, that is, the weighted average Gini impurity:

$$G_{split} = \frac{S_L \sum_{k=1}^{K} p_{kL}(1 - p_{kL}) + S_R \sum_{k=1}^{K} p_{kR}(1 - p_{kR})}{S_L + S_R} \tag{1.8}$$

where S_L = number of observations that go left (L), S_R = number of observations that go right (R), p_{kL} = proportion of class k in left node, and p_{kR} = proportion of class k in right node. Considering all possible splits, the one with the smallest value of G_{split} is used to split the root node, containing all observations, and the first split creates two daughter nodes that are more pure than the parent node. Daughter nodes are themselves split and the process continues.

1.3.2 Tree-Based Methods Naturally Fit Interactions

Tree-based methods are one of the families of machine learning methods that can naturally fit higher order interactions without the specification of which interactions to fit. Figure 1.4 is an example of a how a single tree can capture an interaction. Two SNPs are simulated from the multiplicative model in Figure 1.2c. The minor allele frequency is set to 0.5, and the main effect is set to 0.4. Details about the simulation models and settings can be found in Section 1.2.4. The two-way table of the two SNPs contains a bar plot of the number of controls and cases for each combination of the genotypes. Whether of not the background of each cell is shaded is determined by the classification tree in Figure 1.4b built on the two SNPs. The split that best separates the controls and cases is when the first SNP, x1, splits the genotype AA to the left, and Aa and aa to the right(between row 1 and 2 in Figure 1.4a). If an observation has the genotype AA for that SNP, it goes left and gets classified into the control group since the majority of the observations in the terminal node are in the control group (314 versus 194). Otherwise, the observations go right and the decision rule continues by determining the next best split given that the sets of observations have genotypes Aa or aa for SNP x1. The tree

(A) (B)

FIGURE 1.4: (A) is a contingency table of two SNPs generated from the multiplicative model in Figure 1.2c. Each cell in the table displays a bar plot of the response variable. A classification tree is grown on the two SNPs. The first best split is when the genotype is AA for x1, that is, between row 1 and 2 in (A). Observations with the genotype AA for x1 splits to the left and get classified into the control group, otherwise they split to the right and an additional split is made. There are potentially nine terminal nodes, but the tree is pruned to prevent overfitting.

could have a maximum of 9 terminal nodes, but the tree is pruned to prevent overfitting.

Tree-based methods can capture complex interactions in ways that a model-based method could not. However, the disadvantage of using a tree-based method, such as Random Forests, is the fact that if two SNPs are interacting and neither of them has a main effect, then the chances of the tree fitting the interacting pair decreases as the number of predictor variables increases. A second disadvantage may be the lack of precise indication of where the interacting pairs rank in terms of importance. A possible solution would be to recode each pair as a single covariate. This approach will work if the data has a moderate number of SNPs. An approach that can overcome the issues would be necessary for a tree-based method to be competitive against alternative methods.

1.3.3 Random Forests Algorithm

Random Forests are an extension of CART (Section 1.3.1), the difference being that Random Forests uses bootstrap samples (repeated sampling with replacement from the training set) and randomness in the tree-building procedure. Breiman (2001) defined Random Forests (RF) as "a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest."

The Random Forests algorithm begins with growing a forest of many trees; the default number of trees is $ntree = 500$ in R (Liaw and Wiener, 2002). Each tree is grown on an independent bootstrap sample from the data. At each node, a small number, say the square root of the number of variables ($\sqrt{M}$), of randomly selected variables are chosen. The best split on the selected variables is found, as described in Section 1.3.1. The variables are chosen independently at each node, and the trees are grown to maximum depth until the terminal node contains only one class in classification. Each tree is used to predict the observations that were not in the bootstrap sample ("out-of-bag" observations). The predicted class of an observation is calculated by the majority vote of the out-of-bag (OOB) predictions for that observation.

1.3.4 Random Forests Variable Importance

Random Forests provide two measures of variable importance, Gini importance, and permutation importance (Breiman, 2001). Because permutation importance is more computationally expensive than Gini importance, it is recommended for users to use Gini importance when permutation importance is infeasible (Breiman and Cutler, 2014).

Gini importance is based on the Gini criterion, see Equation 1.7 and Equation 1.8. The predictor variable that was used to form a split has the decrease in the Gini node impurity,

$$\Delta G = G - \frac{S_L G_L + S_R G_R}{S_L + S_R} \tag{1.9}$$

where G, defined in 1.7, is the Gini index of the parent node, and G_L and G_R are the Gini indices of the left (L) and right (R) daughter node, respectively. The Gini importance for the j^{th} predictor variable is the average of the impurity decrease of all nodes in a forest where the j^{th} predictor variable was selected for splitting.

To measure the permutation importance of some variable j consider a single tree and the observations that are OOB. The OOB observations are passed down the tree, and the OOB error rate for the tree is obtained. The OOB error rate is the number of OOB observations correctly classified divided by the number of OOB observations. Randomly permute the values of variable j for the OOB data, so each OOB observation gets a random value for variable j, and all the other observations are kept at their original values. Pass the modified OOB data down the tree and compute a new error rate. If the new error rate is about the same as before this means that the variable does not appear to be contributing to the accuracy of the classification. If the new error rate

is higher than before, the variable's values were useful for accurate classification. The variable importance measure in Random Forests is the difference between the proportion of those predicted correctly before and after permuting the j^{th} predictor X_j, that is,

$$imp_t(X_j) = \frac{\sum_{i \in B_t^c} I(Y_i = \hat{Y}_{ti})}{|B_t^c|} - \frac{\sum_{i \in B_t^c} I(Y_i = \hat{Y}_{ti}^*(j))}{|B_t^c|}. \tag{1.10}$$

where I denotes the indicator function, B_t^c is the OOB sample, $|B_t^c|$ is the number of OOB observations for tree $t = 1, \ldots, ntree$, $\hat{Y}_{ti}$ is the predicted class for observation i from tree t before permuting X_j and $\hat{Y}_{ti}^*(j)$ is the predicted class after permuting X_j. The variable importance measure is found by averaging over all trees:

$$imp(X_j) = \sum_{t=1}^{ntree} \frac{imp_t(X_j)}{ntree}. \tag{1.11}$$

The value of $imp(X_j)$ is used to rank the variables.

1.3.5 Detecting Interactions With Random Forests

An attractive feature of Random Forests and tree-based methods is their ability to automatically fit complex interactions (2-way or more) between predictors without the specification of which interactions to fit (Cutler et al., 2012). Random Forests is able to capture the interactions through the calculation of permutation importance. When one of the variables is permuted, this breaks the predictive power of the interaction. Thus, if an interaction between two important variables exists, it is likely that those variables will show up as important in a Random Forests.

1.3.6 Random Forests Proximities

One of the advantages of Random Forests is that they provide a way to identify outliers, interesting structures or clusters of subgroups of the same class through proximities. The proximity between the ith and jth observations is defined to be the number of times observations i and j are both out-of-bag and in the same terminal node divided by the number of trees in the forest for which i and j are in B_t^c (i.e. out-of-bag). Intuitively, the measure of how often a pair of out-of-bag observations occupies the same terminal node is a measure of how close in proximity that pair of observations is. In a

more mathematical sense, the proximity measure is:

$$prox(i,j) = \frac{\sum_{t=1}^{ntree} I(i \in B_t^c) I(j \in B_t^c) I(q_t(i) = q_t(j))}{\sum_{t=1}^{ntree} I(i \in B_t^c) I(j \in B_t^c)} \tag{1.12}$$

where $q_t(i)$ represents the terminal node observation i falls in, in tree t, for all $i \in 1, \ldots, N$. Two observations that are always in the same terminal node when both out-of-bag will have a proximity of 1 and observations that are never in the same terminal node when both out-of-bag will have a proximity of 0.

The proximity between an observation i from the training set (original data), and observation j from a test set (new data) is found by calculating the following:

$$prox(i,j) = \frac{1}{ntree} \sum_{t=1}^{ntree} I(q_t(i) = q_t(j)). \tag{1.13}$$

Identifying multivariate outliers if there are any and determining any possible structure in the data may be done by creating a multidimensional scaling (MDS) plot using the proximities (Breiman and Cutler, 2014). An MDS plot provides a scatter plot of the N observations in two or three dimensions based on the proximities. The MDS plot is primarily used as a data visualization tool to identify any clustering of points, where the points are viewed as a cluster if some points are close to each other and are not so close to other points that make up another cluster. An MDS plot represents objects that are close in proximity as points that are close to each other, while points that are quite far from each other have lower proximities. Pairs of objects that are very similar regarding the important variables in the classifier will be close to each other on the MDS plot and have proximities close to 1, while pairs of objects that are very dissimilar will be far away from each other on the MDS plot and have small proximities (close to 0).

1.3.7 Proximity-Weighted Nearest Neighbors

The motivation for introducing a new approach to approximating proximities between the ith and jth observation is to improve the Random Forests interpretation and missing value imputation. We introduce a new measure to calculate the proximities between two observations. The idea comes from knowing that Random Forests are like a nearest-neighbor classifier (Lin and Jeon, 2006). Using the proximities defined in Section 1.3.6 as weights for assigning a class label does not accurately reproduce the Random

Forests predictions. Using the new proximities in Section 1.3.8 as weights for a nearest-neighbor classifier, perfectly reproduces the Random Forests prediction accuracy.

1.3.8 New Proximities

The proximity matrix is initialized to the identity matrix. After fitting the t^{th} tree, the proximity matrix is updated for all observations $n \in B_t^c$ and $m \in B_t$:

$$prox(n,m) = \frac{\sum_{t=1}^{ntree} \left(\frac{I(m \in TN_{nt}) R_t(m)}{\sum_{w \in B_t} I(w \in TN_{nt}) R_t(w)} \right)}{\sum_{t=1}^{ntree} I(n \in B_t^c)} \tag{1.14}$$

where B_t is the boostrap sample, TN_{nt} is the set of observations in B_t that are in the same terminal node as observation n, including any repeats, and $R_t(m)$ is the number of times observation m is in B_t. Set $prox(n,n) = 1$.

1.4 No Free Lunch Theorem

Random Forests have been shown to work well on a wide variety of problems without much effort in tuning the parameters. However, Random Forests do not perform the best in every scenario. According to the no free lunch theorem (Wolpert, 1996), there is no model that works best for every problem. Random Forests do not perform well in detecting interactions without main effects. Winham et al. (2012) found that the increase in dimensions, decreases the chance of detecting ground truth more dramatically when SNPs are interacting in comparison to non-interacting SNPs. Known as a black box classifier, Random Forests are not as interpretable as simpler models. The weaknesses of Random Forests give an opportunity and motivation to improve Random Forests.

This dissertation focuses on improving the detection of interactions and interpretability of Random Forests. Chapter 2 proposes a filtering method to detect SNP-SNP interactions. Chapter 3 optimizes the parameter settings of Random Forests. A two-stage approach, combining Chapters 2 and 3 is applied to data on bipolar disorder in Chapter 4. In Chapter 5, we introduce a new visualization method and apply it to a Morse Code data set. Chapter 6 lists potential future work.

CHAPTER 2
A NEW FILTERING METHOD TO DETECT EPISTATIC INTERACTIONS

2.1 Introduction

In this Chapter, we present a new filtering method to detect interactions. We focus on detecting two-way interactions. Although, the filtering method can be used to detect three-way interactions or more. We compare the performance of our new method with leading methods (BOOST and an exhaustive χ^2 test with 8 df) by computational speed and how well the methods detect ground truth.

2.2 Data Simulation Models

In this Chapter, we are using three commonly used epistatic models: additive, multiplicative, and threshold. For some models, combinations of parameters, and sample size, solving Equation 1.5 and Equation 1.6 for α and θ to simulate data does not have a solution. Therefore we restrict the possible parameters used to simulate data for the MAF, main effect, and prevalence. The parameters have the possible values:

1. MAF: 0.1, 0.2, 0.5
2. Main Effects: 0.2, 0.3, 0.4
3. Prevalence: 0.5

A prevalence of 0.5 simulates a binary response variable with an approximately equal number of cases and controls. The number of observations is typically 2,000. The number of predictors simulated varies. The response variable is binary. Only two causative SNPs are embedded which may or may not be interacting. Without loss of generality, we label the causative SNPs x_1 and x_2. None of the simulations incorporate heritability and linkage disequilibrium. Ultimately the data sets look similar to Table 2.1.

TABLE 2.1: Example of data with a binary response from the GWA study. The 0's, 1's, and 2's represent the genotypes, AA, Aa, and aa, respectively where A is the major allele and a is the minor allele.

	y	snp1	snp2	snp3	...	snp750000
1	ctrl	2	0	1	...	0
2	case	0	0	0	...	0
3	case	1	0	2	...	2
⋮	⋮	⋮	⋮	⋮	...	⋮
2000	ctrl	0	1	0	...	1

2.3 Gini Index Versus χ^2 Test

In the process of exploring a new tree-based approach to detect interactions, we discovered that the Gini index and the χ^2 test with 8 degrees of freedom are equivalent when we have a binary response variable (disease/no disease). Imagine a tree with two levels with a three-way split at each node (Figure 2.1). Each terminal node represents the cases and controls for each combination of genotypes at two different loci. The Gini index is determined from the information obtained at those terminal nodes. We can translate the information obtained from the terminal nodes into a 9×2 table (Table 1.2). The

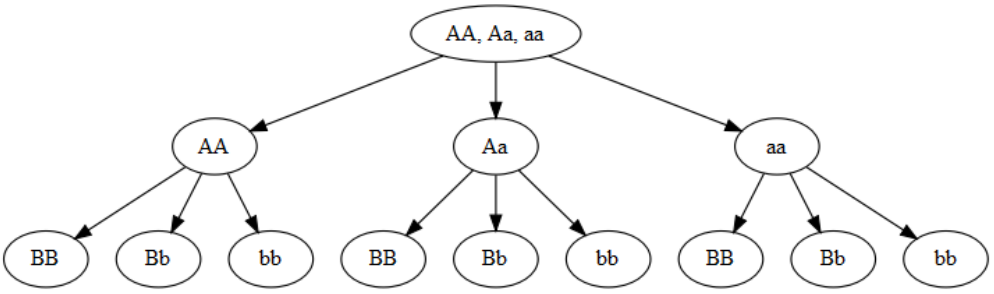


FIGURE 2.1: A tree is built on two SNPs. A three-way split on the first SNP creates three descendent nodes representing each one the SNP's genotypes. An additional three-way split on the second SNP creates a total of 9 terminal nodes. A weighted Gini index on the 9 terminal nodes gives us a measure of how strongly the two SNPs are interacting. The weighted Gini index is an equivalent measure to a χ^2 test with 8 degrees of freedom (Table 1.2).

Gini index for Figure 2.1 is calculated as:

$$\begin{aligned} G &= \frac{2\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{N_{ij}D_{ij}}{N_{ij}+D_{ij}}}{\sum_{i=1}^{3}\sum_{j=1}^{3}(N_{ij}+D_{ij})} \\ &= \frac{2\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{N_{ij}D_{ij}}{N_{ij}+D_{ij}}}{N} \end{aligned} \tag{2.1}$$

The discovery that the Gini index and the χ^2 test are equivalent in this special case has been reported and proved by Grabmeier and Lambe (2007). However, to the best of our knowledge, no existing method has incorporated the idea.

Grabmeier and Lambe (2007) did prove the equivalence of the χ^2 test and the Gini index, but they did not provide an equation to convert from one value to another. We propose the following equation that converts a χ^2 test statistic to a Gini value:

$$Gini = 2p_0p_1\left(1 - \frac{\chi^2}{N}\right) \tag{2.2}$$

where p_0 is the proportion of observations in the control group, p_1 is the proportion of observations in the case group, and N is the number of total observations. The equation is true only if χ^2 and $gini$ are calculated when the response is binary. The conversion is useful for when we want to determine a pool of SNPs to call statistical significant using Gini. For example, suppose $\alpha = 0.1$, $df = 8$, $N = 2000$, equal number of cases and controls, and $\chi^2_{crit} = 13.36157$, then we would have the following Gini critical value:

$$\begin{aligned} gini_{crit} &\approx 2(0.5)(0.5)\left(1 - \frac{13.36157}{2000}\right) \\ &\approx 0.49666. \end{aligned} \tag{2.3}$$

A Gini value less than $gini_{crit}$ is equivalent to statistical significance in the 8 df χ^2 test (Table 1.2).

2.4 Proposed Method

We propose an efficient filtering method that can be used to detect single SNPs and multiway interactions. It can detect interactions of dependent variables of mixed data types (it is not necessarily only applicable to three level categorical variables) for data with a binary response, but we only focus on detecting SNP-SNP interactions. The filtering method can be summarized in the following steps:

1. Randomly select pairs of SNPs.

   To understand why we randomly select pairs, we observe the pattern found in Figure 2.2 by executing steps 2 and 3 using Gini to measure the strength of the interaction between all possible pairs using simulated data for different combinations of parameters and models. We can see the pairs of SNPs that interact with the causative SNPs tend to have a smaller Gini index. Thus, rather than doing an exhaustive search, we can randomly select pairs.

2. Recode each pair of SNPs into a 9-level categorical variable.

   The pair of SNPs can be recoded efficiently using a linear transformation: $3\text{SNP}_1 + \text{SNP}_2$ where the SNPs are coded as 0's, 1's, and 2's. This will result in a variable with nine unique values ranging from zero to eight (see Figure 2.3). For each pair, we want to determine the best split that best separates the cases and controls on the 9-level variable. Typically this would be done by looking at all possible splits. There would be $2^{9-1} - 1 = 255$ possible splits in this case.

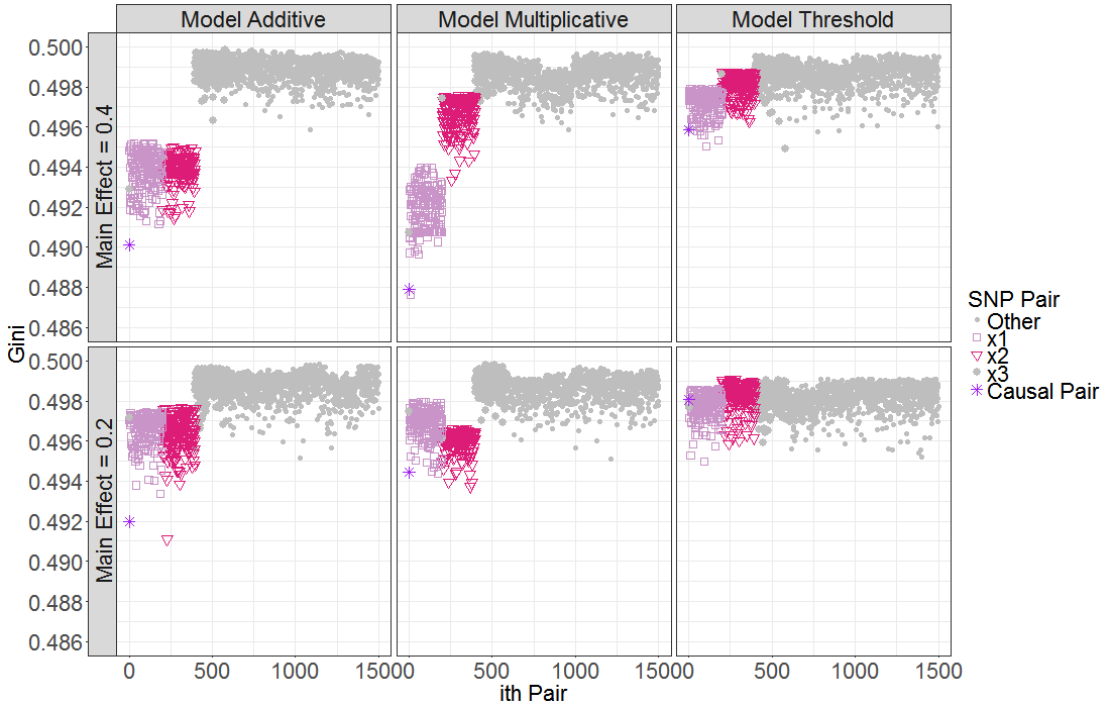


FIGURE 2.2: An exhaustive search to detect SNP-SNP interactions using Gini is applied to simulated data for when the MAF is 0.5, the main effect is 0.2 and 0.4, there are 2,000 observations, and 200 predictors (19,900 pairs) for three epistatic models. Only 1,500 pairs are shown. The pattern is similar for the remaining pairs that are not interacting with the causative SNPs, x1 and x2. Pairs interacting with the causative SNPs tend to have a smaller Gini index.

3. Split once on each pair.

   To avoid looking at all possible splits, we propose a new approach to determine the best split. A contingency table of the 9-level variable and the response variable gives the counts of the number of controls and cases for each combination of genotypes (Figure 2.3). If the number of controls is greater than the number of cases, then those observations split to the left and get classified into the majority group, the control group. Otherwise, the observations split to the right and get classified into the case group. This creates a stump which can translate to collapsing a 9×2 table into a 2×2 table. See Figure 2.4.

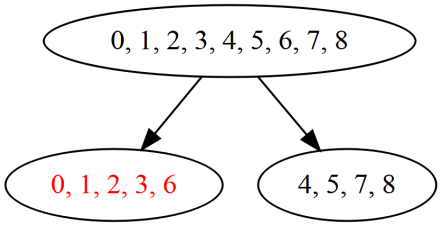
4. Compute Gini index (or χ^2).

   The Gini index or χ^2 test can be used to evaluate the strength of the interaction. Note that p-values will not be valid because we looked at the response to decide

$$3 \times \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{bmatrix} \implies$$

	$Y=0$	$Y=1$
0 (AA, BB)	**411**	**386**
1 (AA, Bb)	**226**	**185**
2 (AA, bb)	**26**	**21**
3 (Aa, BB)	**221**	**182**
4 (Aa, Bb)	59	166
5 (Aa, bb)	5	28
6 (aa, BB)	**33**	**20**
7 (aa, Bb)	3	26
8 (aa, bb)	0	2

FIGURE 2.3: An efficient linear transformation ($3\text{SNP}_1 + \text{SNP}_2$) is used to create a 9-level categorical variable to represent a pair of SNPs that may be interacting. Each of the values represent a combination of genotypes listed in the contingency table. An approximate single split is applied on the 9-level categorical variable. The levels where the number of controls is greater than the number of cases (in red) go left and the remaining levels go right.



	$Y=0$	$Y=1$
0, 1, 2, 3, 6	917	794
4, 5, 7, 8	67	222

FIGURE 2.4: A single split on the 9-level variable in Figure 2.3 sends the levels to the left when the number of controls is greater than the number of cases. Gini index measures the strength of the interaction. A contingency table can be used alternatively to represent the observations in the terminal nodes. Thus, equivalently, a χ^2 test can be used on a 2×2 contingency table, a collapsed table from Figure 2.3.

the split. Figure 2.2 shows the distribution of the Gini index of the first 1,500 pairs if we didn't randomly sample pairs.

5. Choose 10% of the pairs with the smallest Gini index (or largest χ^2 value).

6. Take the sum of the Gini index for each unique SNP randomly sampled in a pair, and rank the sums.
   SNPs that are chosen a higher proportion of times would suggest the SNPs have a main effect or they are interacting. The SNPs could be ranked by the number of times the SNPs corresponding Gini value occurs below a certain threshold. An improved measure would be to take the sum of Gini values in the bottom 10% for each unique SNP. SNPs with smaller sums in Gini are ranked higher.

7. Keep 50% of the SNPs with the smallest sum from Step 6.
   See explanation below.

8. (Optional) Fit Random Forests on the unique set of single SNPs interacting with at least one SNP from step 7.

We want to choose the number of SNPs to randomly sample such that the probability of detecting the causative SNPs is high, the number of times the causative SNPs are in the bottom 10% is high, and the number of unique SNPs in the bottom 10% is low. The number of SNPs randomly sampled is equal to mM, where m is a multiplier, and M is the number of predictors. The value for m is set to 5, 10, 20, 30, and 40, and M is set to 100, 500, 1000. The results are shown in Figures 2.5, 2.6, and 2.7.

The probability of detecting the causative SNPs in the bottom 10% in Gini is calculated by simulating a thousand data sets and averaging the Gini values of the interacting pairs in the bottom 10% for each unique SNP divided by the number of unique SNPs. As the number of randomly sampled SNP pairs increases, the probability of detecting the causative SNPs increases, but starts to stabilize for a multiplier of 10 when the number of predictors is 1000 (Figure 2.5). The threshold model for an MAF of 0.5 has a significantly lower probability of detection. As to why the particular setting for the threshold model is more of a difficult problem is not clear. Notice that overall the probability of detection is approximately above 0.5 for all parameter settings. Therefore, we chose to keep the top 50% best SNPs ranked by the sum Gini value from step 6.

Figure 2.6 illustrates that as the number of randomly selected pairs increases, the percentage of times the causative SNPs are in the bottom 10% increases. However, in Figure 2.7, as the number of randomly selected pairs increases, the average number of unique SNPs in the bottom 10% increases and gets quite close to the number of predictors when the multiplier, m, is 20. For each combination of parameter setting, model, a sample size of 2,000, and the number of predictors, the plots in Figures 2.6 and 2.7 are approximately the same. Overall, taking in considerations of the plots in Figures 2.5, 2.6, and 2.7 we decided that a multiplier, m, of 5 was optimal.

2.5 Methods to Compare to

Wang et al. (2011) and Shang et al. (2011) evaluated different approaches computationally and in terms of ability to detect interactions. Both Wang et al. (2011) and Shang et al. (2011) found that BOOST performed well overall.

Chi-square tests are commonly used as an exhaustive search in two-locus association studies, but Wang et al. (2011) and Shang et al. (2011) did not use the test for comparison. We compare our filtering method against BOOST and an exhaustive χ^2 test.

2.5.1 BOOST

BOOST is a two-stage exhaustive search method for detecting SNP-SNP interactions (Wan et al., 2010a). BOOST uses a Boolean representation of the genotype data and bitwise operations to obtain contingency tables for each pair of SNPs. The programming approach is what makes BOOST efficient. The two stages are defined as follows:

1. Screening Stage: Evaluate all pairwise interactions using $\chi^2 = 2(\hat{L}_S - \hat{L}_{KSA})$ where $\hat{L}_S$ is the log-likelihood of the saturated model in Equation 2.4 and $\hat{L}_{KSA}$ is the log-likelihood of the Kirkwood superposition approximation (KSA). The KSA is a non-iterative method for approximating $\hat{L}_H$, the log-likelihood of the homogeneous association model in Equation 2.5, that is computationally faster. Nonsignificant SNP pairs will be filtered out and those pairs that passed a significance threshold go on to stage 2.

2. Testing Stage: Test each pair where $2(\hat{L}_S - \hat{L}_{KSA}) < \tau$, ($\tau$ is the threshold) using the likelihood ratio statistic $2(\hat{L}_S - \hat{L}_H)$.

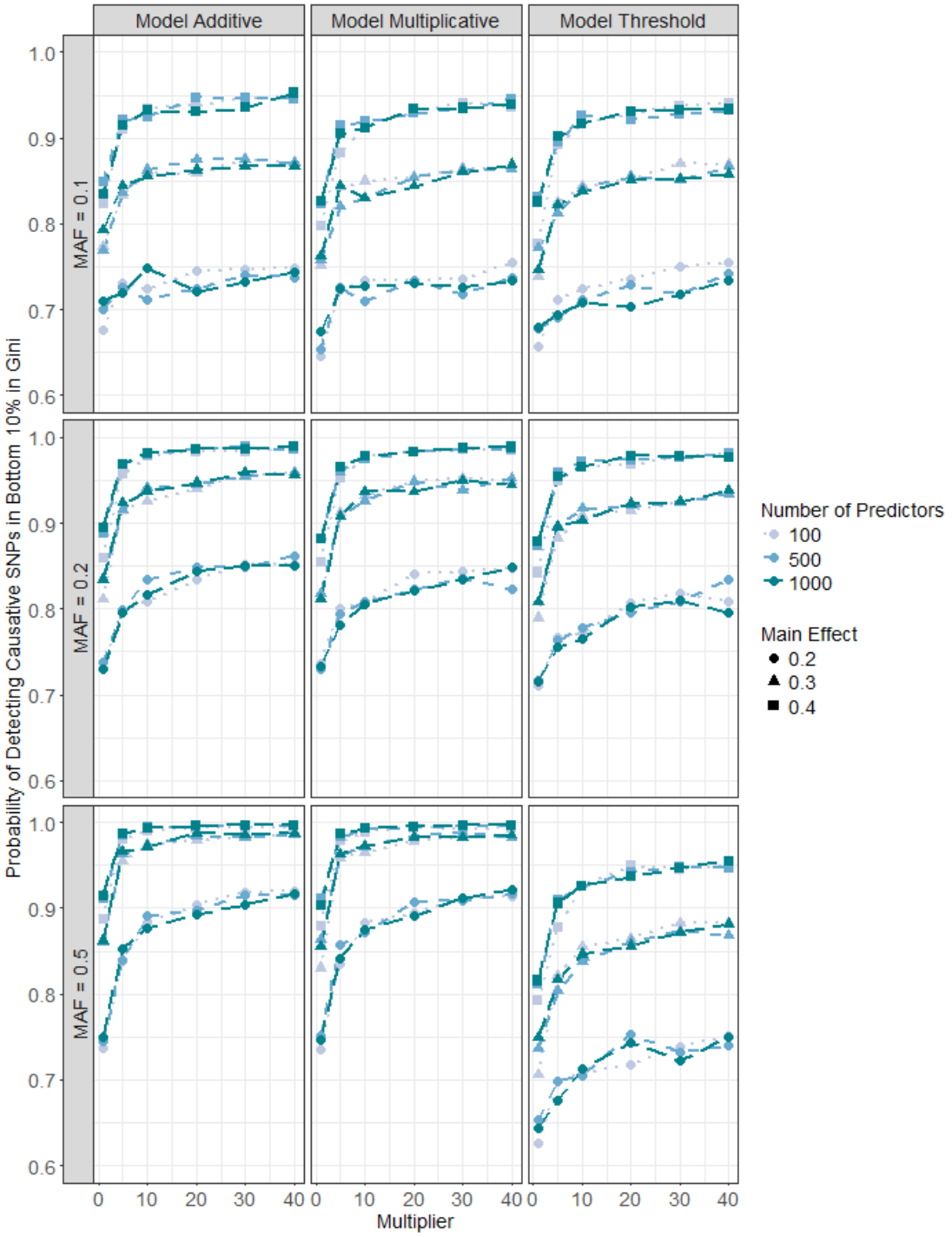


FIGURE 2.5: The number of pairs randomly sampled is the product of the multiplier and the number of predictors. A larger multiplier increases the chances of detecting the causative SNPs for each combination of parameters used to simulated the data but starts to become more steady after a multiplier of 10.

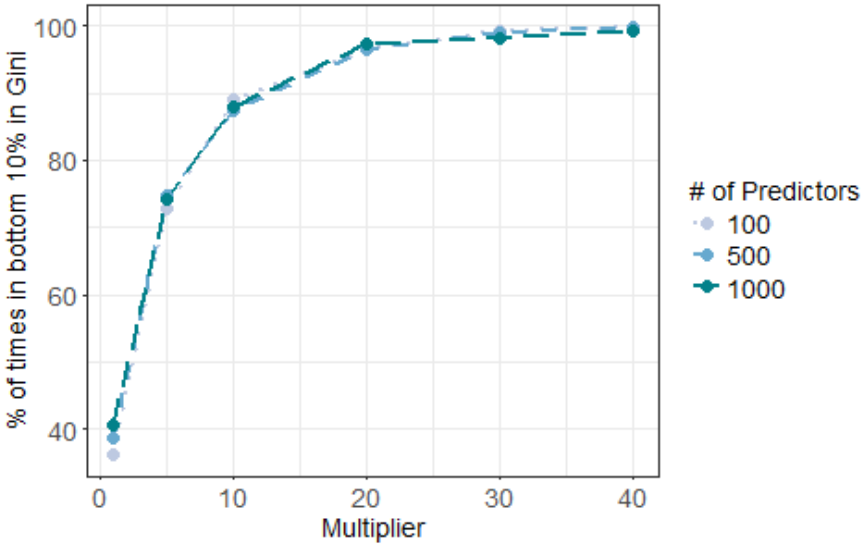


FIGURE 2.6: Data is simulated 100 times and the number of times the causative SNPs are in the bottom 10% in Gini is recorded. The number of pairs randomly sampled is the product of the multiplier and the number of predictors. For each set of predictors, as the number of randomly selected pairs increases, the percentage of the time the causative SNPs are in the bottom 10% in Gini increases. The overall pattern is similar for all parameter settings and models.

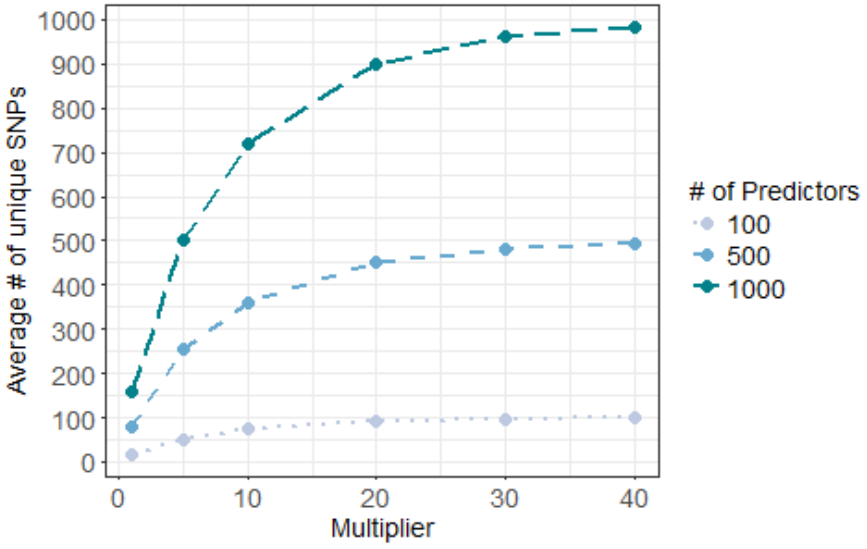


FIGURE 2.7: Data is simulated 100 times and the number of unique SNPs in the bottom 10% is determined in each iteration. As the number of randomly selected pairs increases, the number of unique SNPs increases quickly to the number of predictors used in the simulation. The pattern is consistent for all parameter settings and models.

BOOST defines interactions in the logistic sense. The main effect model is a logistic regression model fitted to a pair of SNPs:

$$log\frac{P(Y=1|X_p=i,X_q=j)}{P(Y=0|X_p=i,X_q=j)} = \beta_0 + \beta_i X_p + \beta_j X_q. \tag{2.4}$$

The interaction model fits logistic regression on both the main effect terms and the interaction term:

$$log\frac{P(Y=1|X_p=i,X_q=j)}{P(Y=0|X_p=i,X_q=j)} = \beta_0 + \beta_i X_p + \beta_j X_q + \beta_{ij} X_p X_q \tag{2.5}$$

where Y denotes the class label (0 for control and 1 for case) in both Equation 2.4 and 2.5. BOOST is written in C and the code is publicly available at `http://bioinformatics.ust.hk/BOOST.html`. Wan et al. (2010a) sets a default threshold τ to 30. $\tau = 30$ corresponds to an unadjusted p-value of $4.89 \times 10^{-}6$.

2.5.2 χ^2 Test With 8 Degrees of Freedom

There are multiple ways to carry out a χ^2 test to determine if two SNPs are interacting. The more efficient approach is to use a test of independence. Two possible approaches are to use a χ^2 test on Table 1.1 with 4 df or a χ^2 test on Table 1.2 with 8 df. In the 4 df test, we are testing whether the 3 variables, Locus A, Locus B, and the response, Y, are independent. In practice, we assume Locus A and Locus B are independent, so we use the 8 df test to see if the distribution of Y is independent of the observed Locus A Locus B combination. In our preliminary results, not shown here, an exhaustive χ^2 test on simulated data showed that the statistical test with 8 df is a better test for our purposes.

While neither of the review papers compared the recent methods that detect epistatic interactions to a χ^2 test with 8 degrees of freedom, the test does perform well in detecting interactions. The χ^2 test is dependent on whether or not the SNPs have a marginal effect. The χ^2 test with 8 df can detect marginal effects as well as interactions. For example, if Locus A = a is associated with an increased risk of disease, the first 3 rows of Table 1.2 will tend to show lower proportions of $Y = 1$ than the remaining rows and this should correspond to high χ^2.

An exhaustive search is performed on simulated data with 2,000 observations, with 200 predictors, an MAF of 0.5, with main effect 0.2 and 0.5 for each of the three epistatic models. Results are shown up to the $1,500^{th}$ pair in Figure 2.8. The pattern is similar

to the results when an exhaustive Gini index is applied (see Figure 2.2).

2.6 Performance Criteria

Different criteria can be used to evaluate how well a method detects interactions. Shang et al. (2011) and Wang et al. (2011) are two review papers that evaluated various methods by determining how well a method detects ground truth and by how well a method can handle a large number of SNPs. Shang et al. (2011) evaluated the robustness to noise using the degree of robustness (DOR), determined how sensitive a method is given a false discovery rate of 0.01 using the receiver operating characteristic (ROC) curve, and calculated computational complexity by measuring running time. Wang et al. (2011) evaluated the performance of epistasis detection by calculating the Type-I error rate, testing the scalability, and analyzing completeness.

One hundred data sets are used, each containing 2,000 observations and 1,000 SNPs, and one pair of ground-truth SNPs is embedded. We set λ, the marginal effect, to be

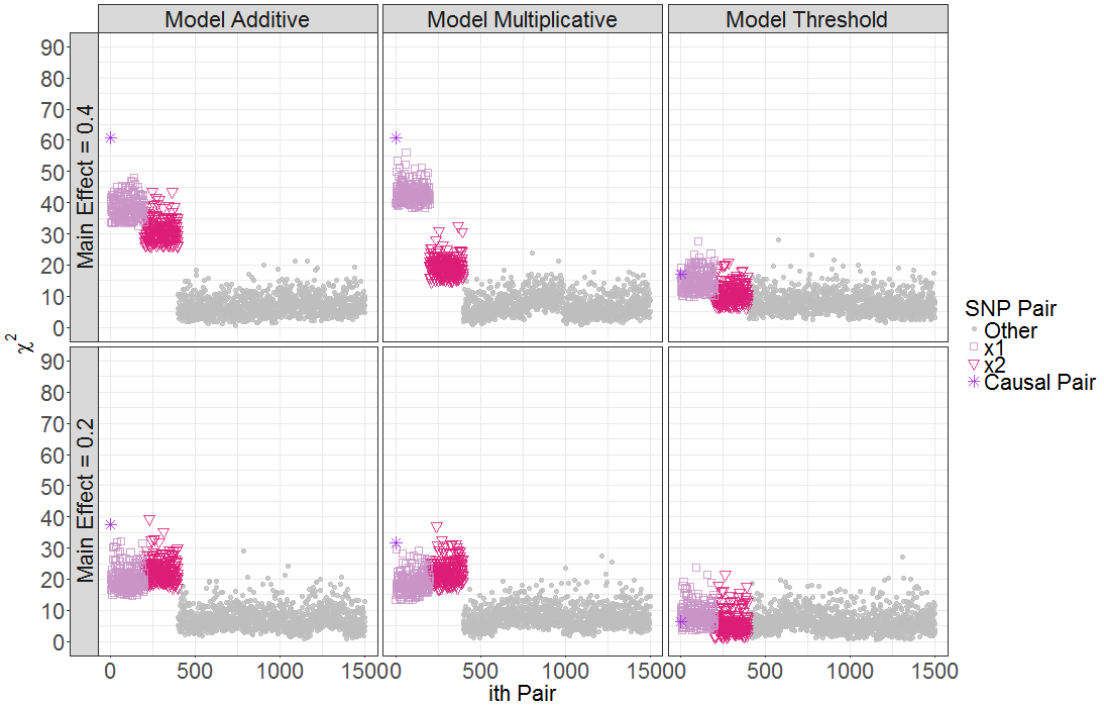


FIGURE 2.8: An exhaustive search to detect SNP-SNP interactions using a χ^2 test is applied to simulated data for when the MAF is 0.5, the main effects are 0.2 and 0.4, there 2,000 observations, and 200 predictors (19,900 pairs) for the three models. Only 1,500 pairs are shown. The pattern is similar for the remaining pairs that are not interacting with the causative SNPs, x1 and x2. Pairs interacting with the causative SNPs tend to have a larger χ^2 test statistic.

0.2, 0.3, and 0.4, prevalence at 0.5 since typically we would be working with real data sets that have approximately equal numbers of controls and cases, and the MAF to 0.1, 0.2, and 0.5. We evaluate the performance of our method against other methods using two different criteria: power and scalability.

2.6.1 Power

Performance is evaluated using general power (GP) and precise power (PP). General power is defined as the proportion of data sets in which all ground-truth interacting SNPs are ranked in the top L SNPs, i.e.,

$$GP_L = \frac{1}{Q}\sum_{i=1}^{Q} d_{gp,i} \tag{2.6}$$

where Q is the number of data sets and $d_{gp,i} \in \{0, 1\}$ is the detection indicator taking the value 1 if the pair is ranked in the top L^{th} position (or the ground-truth SNPs are ranked in the top L^{th} and $L + 1^{th}$ position), and 0 otherwise.

Precise power is defined as the proportion of data sets in which all ground-truth interacting SNPs are ranked highest by a method, i.e.,

$$PP = GP_1 \tag{2.7}$$

if SNP pairs are ranked and

$$PP = GP_2 \tag{2.8}$$

if SNPs are ranked.

2.6.2 Scalability

Interaction detection methods are assessed by how well they can scale. Each data set is generated to have 2000 samples with 100, 1000, 2000, 4000, 6000, 8000, and 10000 SNPs. Each method was timed used on a single core. I have interfaced BOOST in R version 3.3.2. Run time is done within R and determined using the microbenchmark package (Mersmann, 2015). A hundred iterations are used, and the median duration is recorded.

2.7 Results

The filtering method proposed is compared to BOOST and an exhaustive χ^2 test with 8 degrees of freedom. We refer to the new screening technique in terms of what we set the multiplier in Figures 2.9 and 2.10. The number of randomly chosen pairs used are $5M$, $10M$, $20M$, $30M$, and $40M$ and SNPs are ranked using the sum of Gini values of the pairs for each unique SNP in the bottom 10%. Two exhaustive searches using Gini are used to evaluate pairs. One lists the SNPs using the sum of the Gini values of the pairs for each unique SNP and the second ranks pairs using the Gini values themselves. The methods are assessed using precise power (Equation 2.8 and Equation 2.7) and general power (Equation 2.6). Results are found in Figure 2.9 and Figure 2.10, respectively. In

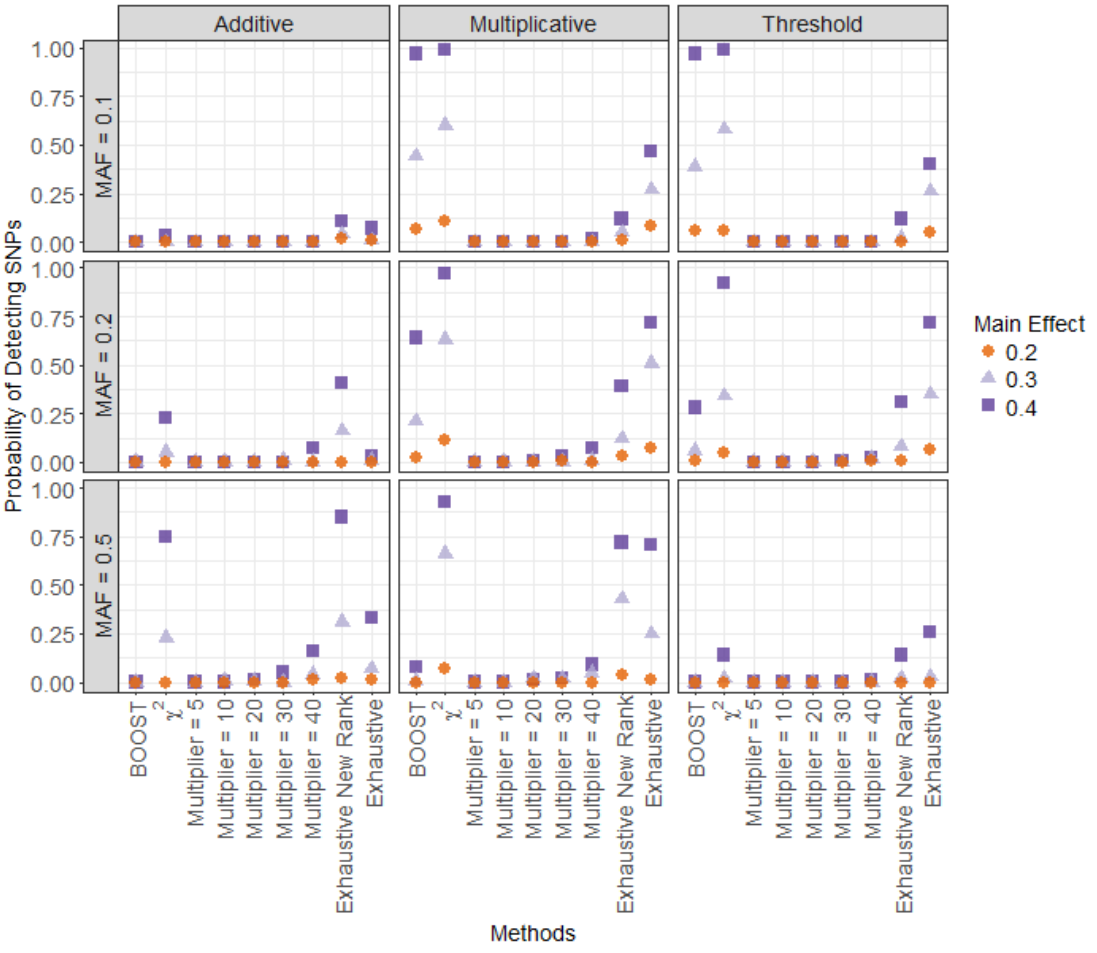


FIGURE 2.9: BOOST, an exhaustive χ^2 test with 8 degree of freedom, our filtering method for multipliers 5, 10, 20, 30, and 40, an exhaustive search using the sum of Gini values to rank SNPs, and an exhaustive Gini index search are evaluated using the definition of precise power in Equation 2.7 if SNPs are ranked by pairs and Equation 2.8 if single SNPs are ranked. The exhaustive χ^2 test performs best overall.

Figure 2.9, the χ^2 test performs the best overall. The Exhaustive Gini Search is second best, and BOOST is third best. The performance evaluation is expected since the best three methods test the strength of interaction for each possible pair.

In Figure 2.10, we used a more liberal cutoff. For methods that rank SNPs, the causative SNPs are considered detected if both SNPs are in the top 50, and if the approach lists pairs of SNPs, then pairs in the top 31,125 are captured. Results are shown in Figure 2.10. As expected, detection power has improved for all methods but more dramatically for the method that is not exhaustive. The probability of detection is consistently larger when there is a stronger main effect for each method. A larger

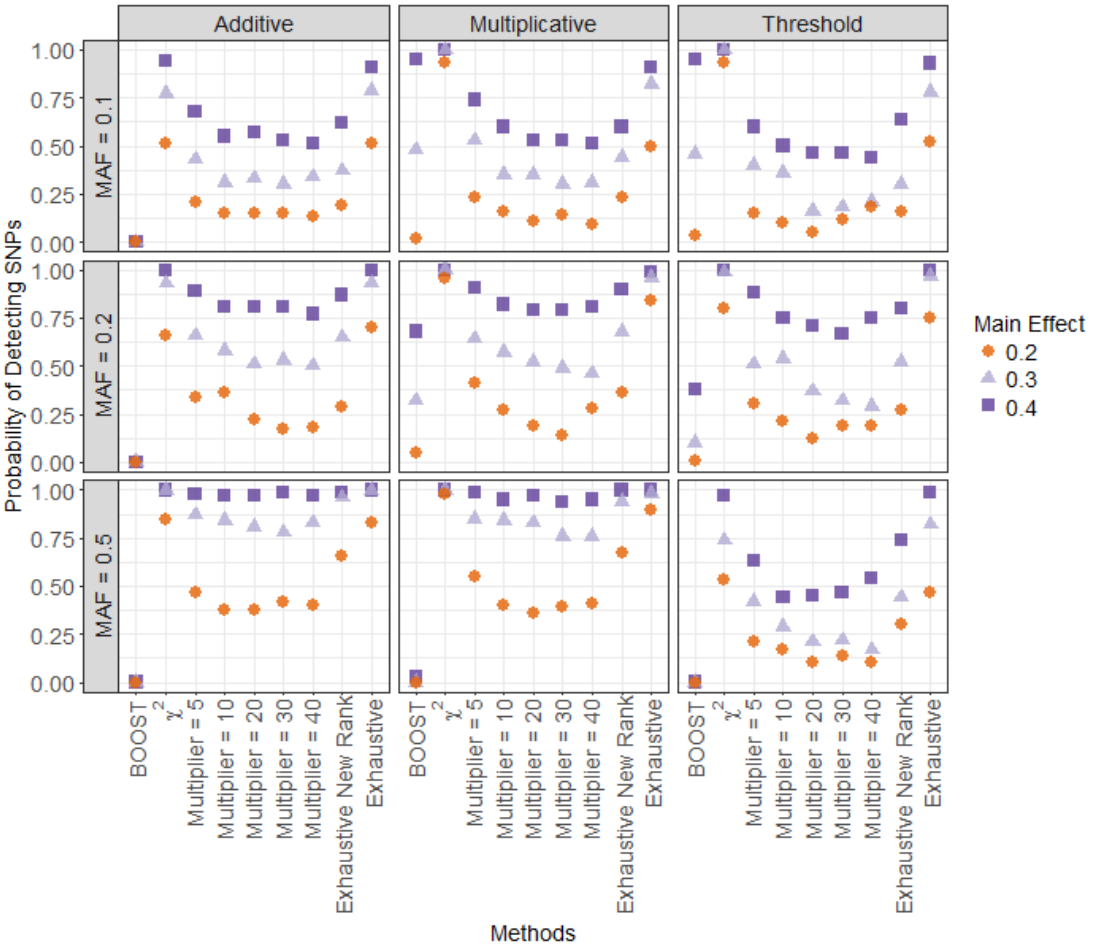


FIGURE 2.10: BOOST, an exhaustive χ^2 test with 8 degree of freedom, our filtering method for multipliers 5, 10, 20, 30, and 40, an exhaustive search using the sum of Gini values to rank SNPs, and an exhaustive Gini index search are evaluated using the definition of general power in Equation 2.6. Causative SNPs are detected if SNPs are ranked in the top 250 and top 31,125 if pairs are ranked. For each combination of the MAF, main effect, and model, each method improved overall compared to using precise power in Figure 2.9.

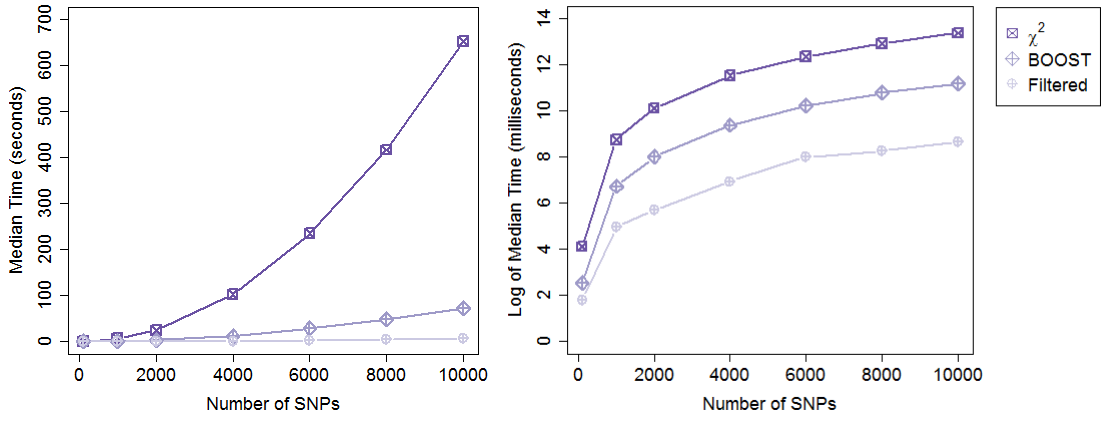
MAF did not necessarily have a greater probability of detection and is apparent when the MAF is 0.5 for the threshold model. In particular, the filtering method improved dramatically. For any of the multipliers used in the screening method, the performance in detection power is quite similar. Thus a multiplier of 5 seems to be adequate.

A comparison of the median time each method spent for a data set with 2,000 observations shows that the exhaustive χ^2 test is slower than BOOST and even slower than the filtering method with a multiplier of 5. See Figure 2.11. Figure 2.11a shows the median time in seconds each method took for 100, 1,000, 2,000, 4,000, 6,000, 8,000, and 10,000 predictors. Alternatively, Figure 2.11b shows the median time in log scale of milliseconds.

2.8 Conclusions

The number of SNP-SNP interactions grow exponentially as the number of SNPs increases poses a computational and methodological challenge. It is necessary that SNP interaction detection approaches are fast with high statistical power for a wide variety of epistatic models.

We consider only genome-wide case-control studies when the phenotype is binary. Data is simulated from three common epistatic models, additive, multiplicative, and threshold. We used five parameters, MAF, prevalence, main effect, sample size, and



(A) (B)

FIGURE 2.11: An exhaustive χ^2 test, BOOST, and the new filtering method (sampling $5M$ SNP pairs) are compared using the microbenchmark package in R. The median time is taken over 100 iterations for data sets with 2,000 observations and 1,000 predictors.

the number of predictors to generate data. A maximum of one causal interaction is embedded in each data set.

We compared our filtering method against leading methods that perform well in terms of power and computational time. Two survey papers compared SNP-SNP detection methods using simulated data, found that overall BOOST performed the best in terms of detection power and speed. Neither survey papers compared the methods against an exhaustive χ^2 test, but χ^2 tests are incorporated in quite a few new interaction detection methods. We compared our filtering method against BOOST and an exhaustive χ^2 test with 8 degrees of freedom.

The methods were compared in terms of computational time. We demonstrated that our screening technique is computationally feasible for hundreds of thousands of SNPs for thousands of observations and is faster than BOOST and an exhaustive χ^2 test.

The methods are evaluated in terms of detection ability using precise power and general power. See Equations 2.6, 2.7, and 2.8. The exhaustive χ^2 search outperforms BOOST and our method overall if the methods are evaluated using precise power. It is expected that techniques that don't look at each possible SNP pair to perform poorly if they are evaluated using precise power. General power, a more liberal evaluation, showed that our filtering method performs better than BOOST in some settings and performs worst against the exhaustive χ^2 test in the majority of the scenarios.

In summary, the following lists the advantages and disadvantages of using the filtering method:

1. Advantages

    - Can be used on dependent variables of mixed data types.
    - Can fit multiway interactions.
    - Fast.
    - Can detect single SNP associations simultaneously.
    - Can control false positives by using Bonferroni correction and Equation 2.2 to convert from a χ^2 critical value to a Gini critical value.
    - Gini can be interpreted as a probability.

2. Disadvantages

- Dependent on SNPs with main effects.
- Can't be used when the response variable is continuous.

The χ^2 test cannot handle continuous covariates, it is dependent on whether or not the SNPs interacting have marginal effects, and cannot be used if the response variable is continuous. The current implementation of BOOST cannot handle continuous phenotypes.

CHAPTER 3
OPTIMIZING RANDOM FORESTS FOR DATA FROM THE GENOME-WIDE ASSOCIATION STUDY

3.1 Introduction

In this Chapter we show that Random Forests can be improved in its ability to identify interactions and efficiency by a combination of changes in its parameters settings, by searching for sensible splits (Sections 3.2.1), using the optimal measure of variable importance (Section 3.2.2), and by determining the optimal *nodesize* (Sections 3.2.3).

3.2 Personalizing Random Forests

Random Forests are implemented as a general purpose machine learning method. The different parameter settings are the number of trees, the number of predictor variables to be randomly chosen at each node, and how deep to grow each tree by specifying the max size of the terminal node. There are two situations where the max size of the terminal node can be exceeded. One, if the variable cannot be split on, that is, if all the values of the randomly chosen variable are the same. Second, the node is pure, i.e., the response variable is the same for all observations.

The arguments of the three parameters are most commonly known as *ntree*, *mtry* and *nodesize*, respectively. The default setting for *ntree* is 500, for *mtry* it is $\sqrt{M}$ where M is the number of predictors, and *nodesize* is set to 1 for classification in the package randomForest in R (Liaw and Wiener, 2002). While these are the default settings and they are found to be optimal empirically in Breiman (2001), no theory supports it.

The parameter settings for Random Forests for data from the GWA study should be optimized to increase the chances of detecting causative SNP-SNP interactions in higher dimensions. Goldstein et al. (2010) used larger values for *mtry*, increased the number of trees, used the default node size of 1, and used the permutation importance to determine significant SNPs. Winham et al. (2012); Kim et al. (2009); Wright et al.

(2016) evaluated how well Random Forests detected causative interacting SNPs.

Winham et al. (2012) found that it was optimal to use 5,000 trees where the prediction error was low and the variable importance measures stabilized and $mtry = 0.1M$. In conclusion, they found that Random Forests failed to detect interaction effects in high-dimensional data when a strong marginal component was absent, increasing *ntree* did not improve the probability of interaction detection, and increasing *mtry* to $0.5M$ improved the prediction error, but not detection of interactions. However, Winham et al. (2012) did not optimize *nodesize*.

Wright et al. (2016) evaluated how well the Gini importance and permutation importance from Random Forests can detect gene-gene interactions and compared it to pairwise importance measures and a joint variable importance method. They stated that in the majority of the scenarios, the proportion of detected SNPs was larger for each of the four variable importance methods used when there were only marginal effects present in comparison to a model with an interaction effect only. Gini variable importance was able to detect interactions at least as well as permutation variable importance in the majority of the scenarios. This is important because Gini importance is computationally faster than permutation importance.

3.2.1 Making Logical Splits

In Random Forests, from a subset of *mtry* variables chosen at random, the variable that best separates the disease versus non-disease groups is selected to be split on at that node. In the original implementation, Random Forests looks at all combinations of a categorical variable to decide the optimal split. The total number of possible separations for a categorical variable with k levels is $2^{k-1} - 1$. An example of this for a single SNP is shown in Figure 3.1. Note that going left or right is equivalent, so we do not consider the splits where the nodes are reversed. However, in GWA study data, some combinations of the genotypes would be unlikely to be biologically sensible. For example, a split should not send AA and aa one way and Aa the other. Thus Random Forests has the potential to improve computationally by only considering biologically meaningful separations.

3.2.2 Permutation Versus Gini Variable Importance

One of the best uses of Random Forests is that it can handle data sets where the number of variables is much greater than the number of observations. Unlike some machine learning methods, Random Forests can rank variables using a permutation variable

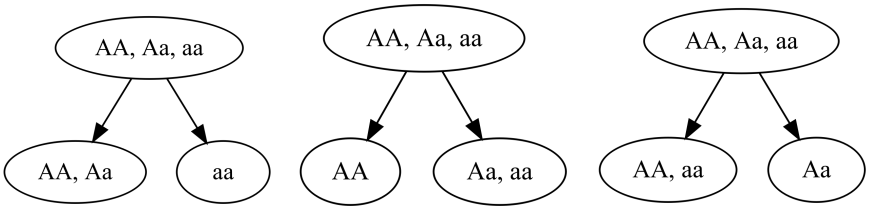


FIGURE 3.1: There are three distinct binary splits for a single SNP.

importance measure or a Gini variable importance measure.

Permutation variable importance is known to be a better measure of variable importance in comparison to Gini variable importance (Strobl et al., 2007). For larger data sets, it is advised to use Gini variable importance to reduce the number of variables due to its lower computational cost.

Figure 3.2 compares the chances of detecting the causative SNPs in each of the three models presented in Section 1.2.5. The number of SNPs is set to 20, 1000, 2500, and 5000, 2,000 observations are simulated, prevalence is set to 0.5, the MAF is set to 0.1 and 0.5, and the main effect is set to 0.2 and 0.4. For each combination of the settings, 100 data sets are simulated to determine the probability of detecting the causative SNPs, x1 and x2. The default parameters in Random Forests are used, i.e., *mtry* is set to $\sqrt{M}$, *nodesize* is set to 1, and *ntree* is set to 500. The Gini variable importance measure consistently outperforms the permutation variable importance measure. This is consistent to what Winham et al. (2012); Kim et al. (2009); Wright et al. (2016) have reported.

3.2.3 Node Size

The number of observations in the terminal node for each tree grown is an optional parameter usually set to one. We want to optimize *nodesize* in terms of detecting causative SNPs when there are 100 predictors, 2,000, 3,000, 4,000, 5,000, 10,000 observations, prevalence is 0.5, MAF is 0.1, 0.2, 0.5, and the main effect is 0.2, 0.3, and 0.4. For each combination of the settings, 100 data sets are simulated and the probability of detecting the causative SNPs, x1 and x2, is determined using Gini variable importance. *Nodesize* is set to varying proportions, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, of the sample size. Figure 3.3 shows the results for the first causative SNP and when the main effect is 0.2. The results are similar for the second causative SNP when the main effect is 0.3 and 0.4 (not shown here). Overall, the chance of detecting

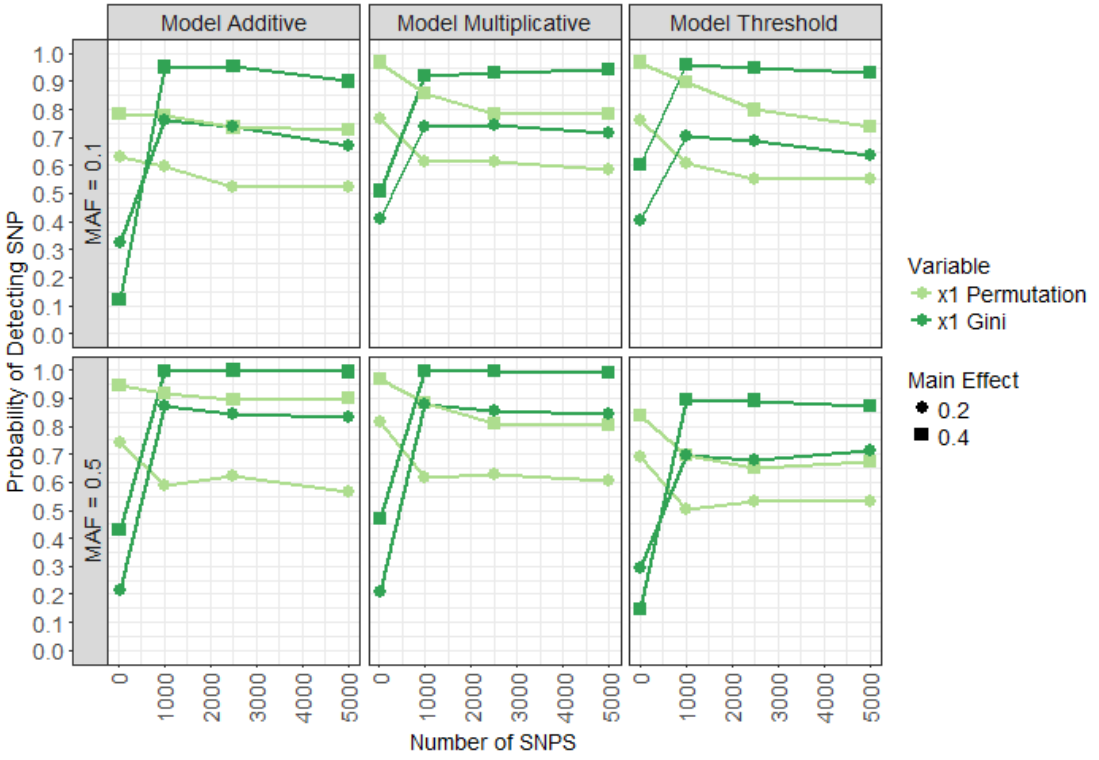


FIGURE 3.2: For each of the epistatic models, parameter setting of the MAF, and the main effect, the Gini variable importance measure outperforms permutation variable importance from Random Forests.

the causative SNPs is approximately the same for each possible combination of the data simulation parameters used. There is some variability in the probability for the multiplicative and threshold model for when MAF is 0.1. In this case, it would be optimal to use a smaller *nodesize*.

A sample size of 2,000 observations is further examined in Figure 3.4 to determine which *nodesize* is optimal. There are little differences in the probabilities. A *nodesize* equal to 5% or 10% is optimal for the majority of possible combinations used. A larger *nodesize* would grow trees that are more shallow, but it would be computationally faster so it may be a preferable to use 10% of the sample size for those with much larger data sets.

3.2.4 Number of Trees

As with any Random Forests classification model, the number of trees should be grown until both the class errors have become constant and the variable importance measures have stabilized. It is expected that the number of trees required should increase as the number of predictors increases. An example is shown in Figure 3.5. Two

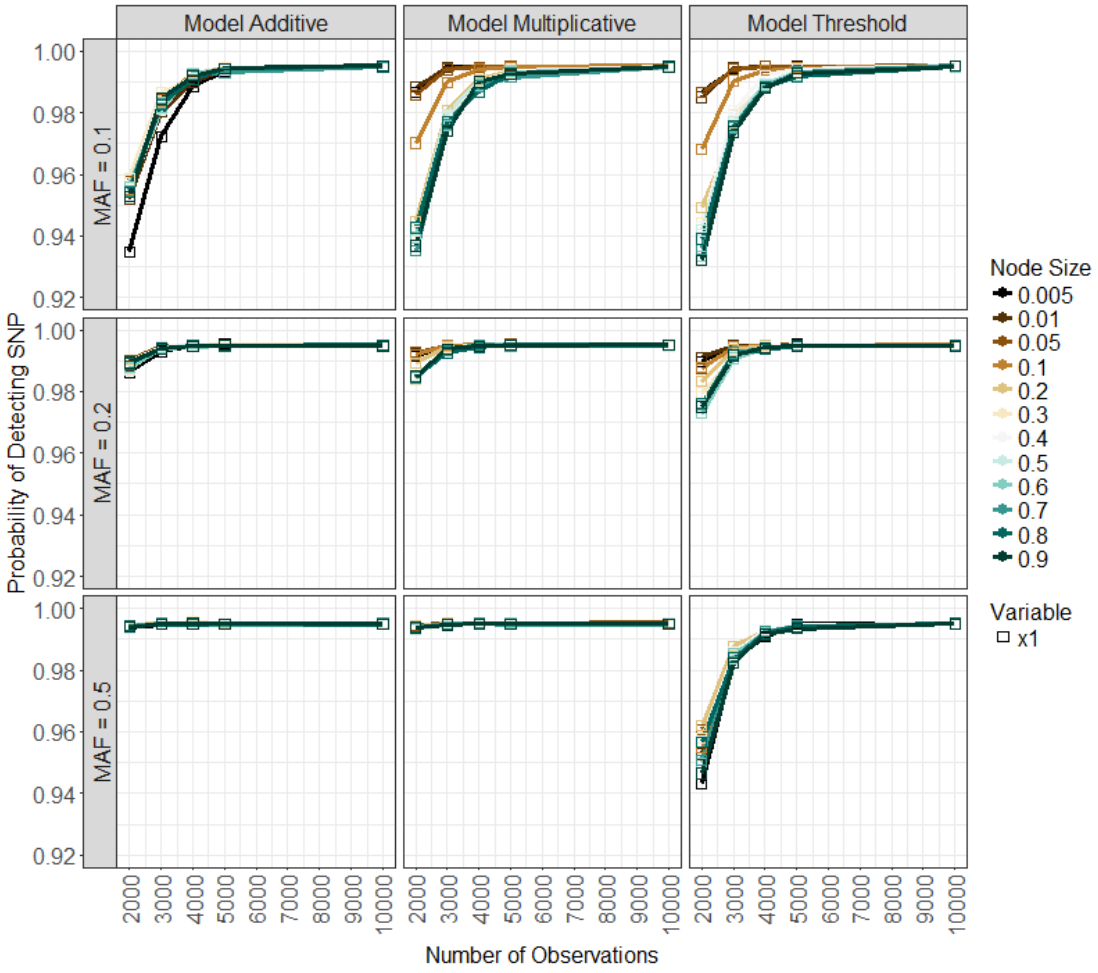


FIGURE 3.3: For each possible data simulation parameter setting, model, 100 predictors, sample size combination, and main effect of 0.2, the probability of detecting the causative SNPs is approximately the same, indicating that growing deeper trees is fitting a lot of noise.

thousand observations, a prevalence of 0.5, MAF of 0.1, a main effect of 0.2, and the threshold model are used to simulate data. Using Gini importance as a measure to identify relevant SNPs and varying the number of predictors and the number of trees grown, Figure 3.5a shows that the probability of detection becomes approximately constant for 1,000 predictors when 5,000 trees are grown and increasing the number of trees grown to 10,000 does not significantly improve the likelihood of detection. The chances of detection are similar with 2,000 predictors. If we fix the number of trees grown to 5,000 and increase the number of predictors, Figure 3.5b shows that as the number of predictors increases beyond 1,000 and 2,000, the probability of detecting the causative SNPs decreases.

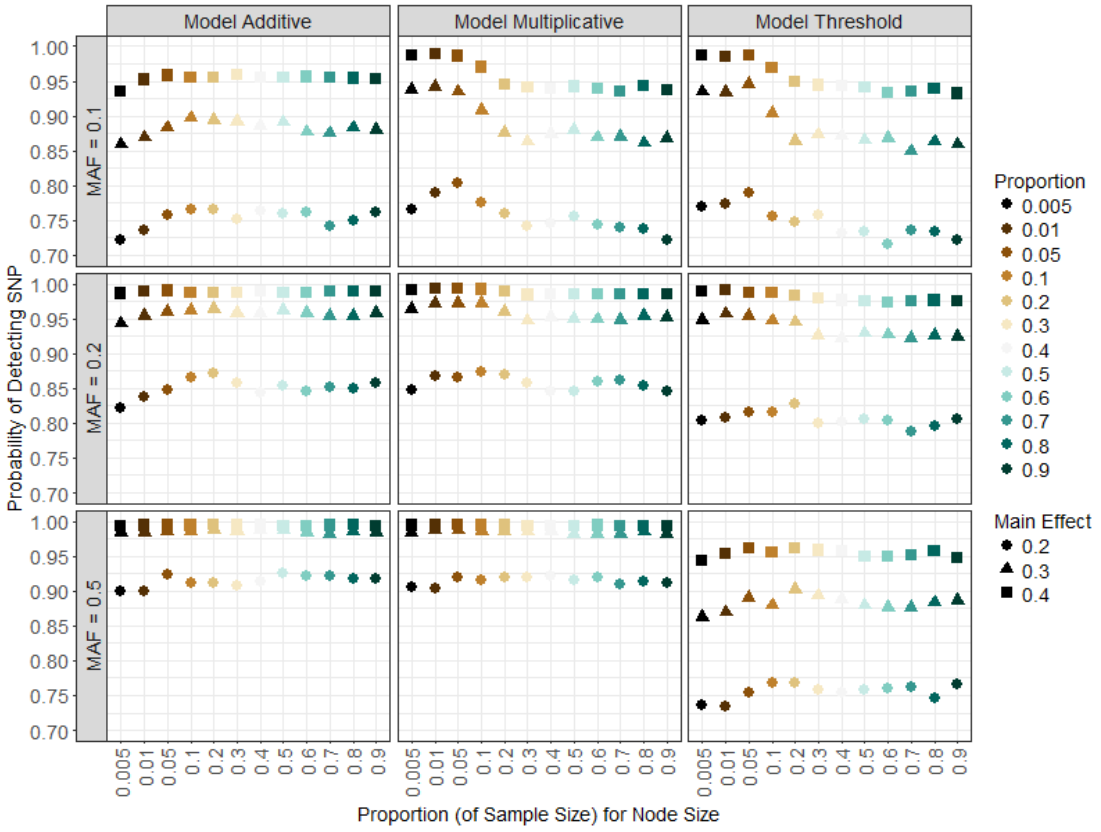


FIGURE 3.4: For each possible data simulation parameter setting, model, 100 predictors, and a sample size of 2,000, the probability of detecting the causative SNPs is close to optimal when the *nodesize* is either 5% or 10%.

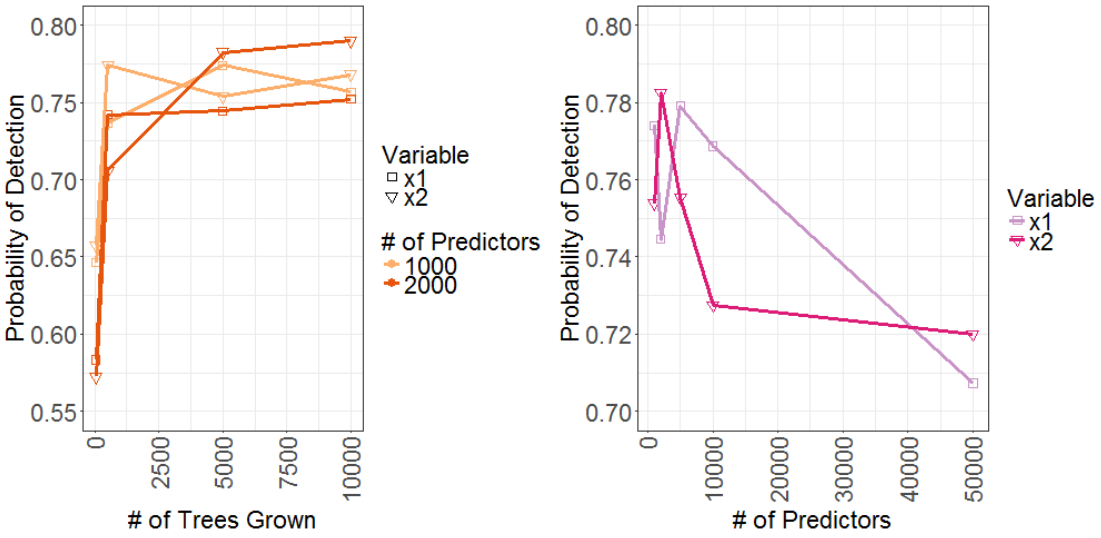
3.3 Conclusions

Statistical methods, such as logistic regression and linear regression are unable to handle data sets when the number of dependent variables exceeds the number of observations ($M >> N$). Machine learning methods, such as Random Forests is commonly used as an alternative approach to reduce the number of predictors or as a means to detect interactions.

Random Forests can naturally capture multiway interactions without specifying which ones to fit. However, studies have shown that Random Forests fails to identify genetic interactions in higher dimensions when interacting SNPs both have weak main effects. In this Chapter, we optimized Random Forests by:

1. The variable importance measure.

   Permutation variable importance is computationally expensive but is generally a better measure of variable importance in comparison to Gini variable importance. It is recommended that Gini variable importance should be used if the resources



(A) (B)

FIGURE 3.5: A hundred data sets are generated from the threshold model with a MAF of 0.1, main effect of 0.2, and 2,000 observations. Random Forests is applied to each data set with the default setting for *mtry* ($\sqrt{M}$) and node size equal to 10% of the sample size. The probability of detecting the causative SNPs is determined using Gini variable importance. In (A), as the number of trees grown increases, the probability of detecting the causative SNPs stabilizes and does not improve after 5,000 trees. This is similar for 2,000 predictors. Fixing the number of trees to 5,000 and increasing the number of predictors in (B), the probability of detecting the causative SNPs decreases.

are limited. We showed the performance of Random Forests detection of causal SNP interactions was best for each possible data simulation setting when the Gini variable importance measure ranked the SNPs. The explanation as to why Gini variable importance outperformed permutation variable is not well understood.

2. *nodesize*.

   Gini variable importance is further used to determine the best setting for *nodesize*. Our simulation results show that it is unnecessary to grow deep trees. A *nodesize* of 5% or 10% of the sample size captured the causative SNPs best.

3. The number of trees.

   We illustrate that the increase in the number of predictors requires more trees. The number of trees grown should be sufficient such that the class error becomes constant and the variable importance measure has stabilized. We showed that for 1,000 predictors, the probability of detection stabilized after 5,000 trees were grown and similarly for 2,000 predictors. Fixing the number of trees grown to 5,000, as

the number of predictors increases, the probability of detecting the causal SNPs decreased. It is recommended that the number of trees grown should be as much as the user can afford.

Also, we showed that the efficiency of Random Forests could improve by doing logical splits. The combination of using the mean decrease in Gini as a measure of variable importance and growing shallow trees improved the efficiency and predictive power in Random Forests.

CHAPTER 4

AN APPLICATION OF THE NEW FILTERING METHOD

4.1 Introduction

Bipolar disorder is a psychiatric disorder that affects approximately 1% of the population (Smith et al., 2011). Patients that suffer from bipolar disorder experience dramatic shifts between depression and mania. The risk of suicide for those individuals is as high as 17%. Evidence from studies of families, twins, and adoption show that there is a genetic component that plays a primary role in the psychiatric disorder (Barrett and Cardon, 2006; Consortium, 2005; Murray and Lopez, 1996). GWA studies have allowed us to search for possible genetic variants that increase the risk of bipolar disorder.

We apply our new filtering method on a bipolar disorder GWA study data set. The individuals are of European ancestry, comprised of 1,034 controls and 1,001 cases. We focus on the 21^{st} chromosome that has a total of 12,143 SNPs.

Previous analysis that looked at all chromosome (769,672 SNPs) has been published in Smith et al. (2009) and Smith et al. (2011). More information regarding the quality control of the bipolar data set is found in the Supplementary material provided from Smith et al. (2009). Neither paper looked at all possible SNP-SNP interactions. Smith et al. (2009) did not identify any SNPs that passed the level of significance of 5×10^{-8}. When they performed a fixed effects meta-analysis with SNPs found in common in the Wellcome Trust Case Control Consortium bipolar data set, they found no genome-wide significant association. Smith et al. (2011) conducted a GWA study involving a sample of individuals of European ancestry and African ancestry. They did not identify any SNPs that passed the level of significance of 5×10^{-8} in both samples. They reported the top two SNPs, rs5907577 and rs10193871, with the strongest statistical evidence for association with bipolar disease in the European ancestry sample, and rs2111504 and rs2769605 were identified in the African ancestry sample.

SNPedia (2017) lists SNPs that have been reported to increase the risk for bipolar disorder. None listed any statistically significant two-way SNP interactions. Hu et al. (2010) analyzed each possible pair of SNPs from the WTCCC bipolar disease GWA study data and did not find any statistically significant interactions. They selected potential SNP pairs based on a set of criteria to be used for a replication study and found two potential pairs, rs10124883-rs178069 and rs10124883-rs6004133. The replication study included 475 bipolar patients from the Chinese Han population. The first pair, rs10124883-rs178069, had a p-value of 0.026 and the second pair, rs10124883-rs6004133, had a p-value of 0.021. In addition, Hu et al. (2010) found the following risk pairs: rs10124883-rs6004133 (p-value = 0.027), rs10124883-rs165730 (p-value = 0.038), rs10124883-rs165596 (p-value = 0.035), and rs10124883-rs178069 (p-value = 0.031). Prabhu and Pe'er (2012) analyzed the WTCCC bipolar data set and found only one statistically significant pair: rs10925490 within RYR2 on chr1q43, and rs2041140 and rs2041141 within CACNA2D4 on chr12p13.33.

4.2 Quality Control

The real data set initially has a total of 2,035 observations and 12,143 SNPs in the 21^{st} chromosome. The snpStats package in R was used to carry out the data cleaning (Clayton, 2015). SNPs were removed for the following reason:

1. SNPs with 5% or more missing values
2. MAF is less than 0.1
3. SNPs with only two categories

There are 9,993 remaining SNPs. No observations are removed. Missing value imputation is done by substituting the most frequent occurring genotype.

4.3 Results

The filtering method presented in Chapter 2 is applied to the bipolar disorder data set. The data set is split into two data sets, a training set and a test set. The training set contains two-thirds of the observations that are randomly sampled from the data set, and the test set contains a third of the observations from the data set that are not in the training set. The following steps are applied to the training set:

1. There are approximately 9,993(9,992)/2 $\approx$ 50 million SNP pairs. $5(9,993) = 49,965$ SNP pairs are randomly selected.

2. Each pair of SNPs is recoded into a 9-level categorical variable. See Figure 2.3.

3. A single split is determined for each variable in Step 2.

4. Gini index is calculated on the split from Step 3 to measure the strength of the interaction.

5. Ten percent of the pairs in Step 4 with the smallest Gini index is retained.

6. The SNPs from Step 5 are ranked by taking the sum of the Gini index of the interacting SNP pairs for each unique SNP.

7. Fifty percent of the SNPs in Step 6 with the smallest sum value in Gini are kept.

After Step 7 is applied, there were 2,515 SNPs remaining. Random Forests is fit on the unique set of single SNPs. *mtry* is set to the default ($\sqrt{2,515} \approx 50$), *nodesize* is set to 10% of the sample size ($.1(2,035) \approx 204$), and 25,000 trees are grown. The SNPs are ranked using Gini variable importance. See Figure 4.1. The number of SNPs considered important in predicting bipolar disorder is determined by the jumps in the mean decrease in Gini. In this case, either 1, 2, 3, or 5 SNPs are a reasonable number of SNPs to keep.

The position, MAF, and the χ^2 test used to test for a main effect using the test set on the top 5 important SNPs with the largest mean decrease in Gini are listed in Table 4.1. Only one SNP, 8578860, passes the Bonferroni adjusted χ^2 critical value.

The 5 important SNPs determined from Random Forests in Figure 4.1 are validated using the test set. Random Forests do not provide information about which pairs of SNPs are interacting. We carry out an exhaustive χ^2 test with 8 df search on the 5 important SNPs identified. The pairs and it's corresponding χ^2 test statistic are listed in Table 4.2. The adjusted χ^2 critical value using Bonferroni is 13.52 for a level of significance of 0.05. Of the 10 pairs, none are considered statistically significant.

4.4 Conclusions

We analyzed a bipolar disorder GWA study data set. Of the 12,143 SNPs from the 21^{st} chromosome, 9,628 are removed. We split the data set into two sets, a training set

TABLE 4.1: Additional information is listed for the top 5 most important SNPs determined using Gini variable importance from Random Forests. The SNPs are validated using a χ^2 test with 2 df on the test set. The SNPs are considered significant if it passes a Bonferroni adjusted threshold of 3.32.

	SNP	Mean Decrease in Gini	Position	MAF	χ^2 2 df ($\chi^2_{crit} = 3.32$)
1	8578860	0.41	37724743	0.28	4.01
2	8579214	0.39	42829968	0.44	1.57
3	2019016	0.29	37810614	0.28	3.07
4	2019010	0.24	37771610	0.27	2.50
5	2008118	0.24	24093103	0.08	1.74

and a test set. The training set contains two-thirds of the observations, and the test set contains the remaining observations. We applied our filtering method on the training data, and 2,515 SNPs are fit into a Random Forests model. Gini variable importance ranked the SNPs, and 5 potential SNPs were identified as important predictors. The 5 important SNPs are validated using the test set. An exhaustive χ^2 test is used to determine which pairs are statistically significant. Of the 10 pairs, none of the pairs were identified statistically significant after Bonferroni correction. However, a χ^2 test on the single SNPs identified a statistically significant SNP after Bonferroni correction. Smith et al. (2009) and Smith et al. (2011) analyzed similar data with all chromosomes and did not identify and statistically significants SNPs.

Previous studies that found significant SNP-SNP interactions looked at all chromosomes and did not find any statistically significant risk pairs in the 21^{st} chromosome (Hu et al., 2010; Prabhu and Pe'er, 2012).

TABLE 4.2: The top 5 most important SNPs determined using Gini variable importance from Random Forests are validated using an exhaustive χ^2 test with 8 df search on a test set. Pairs are considered significant if it passes a Bonferroni adjusted threshold of 13.52.

	SNP Pair	χ^2 8 df ($\chi^2_{crit} = 13.52$)
1	8578860 - 8579214	7.63
2	8578860 - 2019016	6.35
3	8578860 - 2019010	10.34
4	8578860 - 2008118	7.15
5	8579214 - 2019016	5.76
6	8579214 - 2019010	6.38
7	8579214 - 2008118	11.38
8	2019016 - 2019010	6.79
9	2019016 - 2008118	6.48
10	2019010 - 2008118	7.44

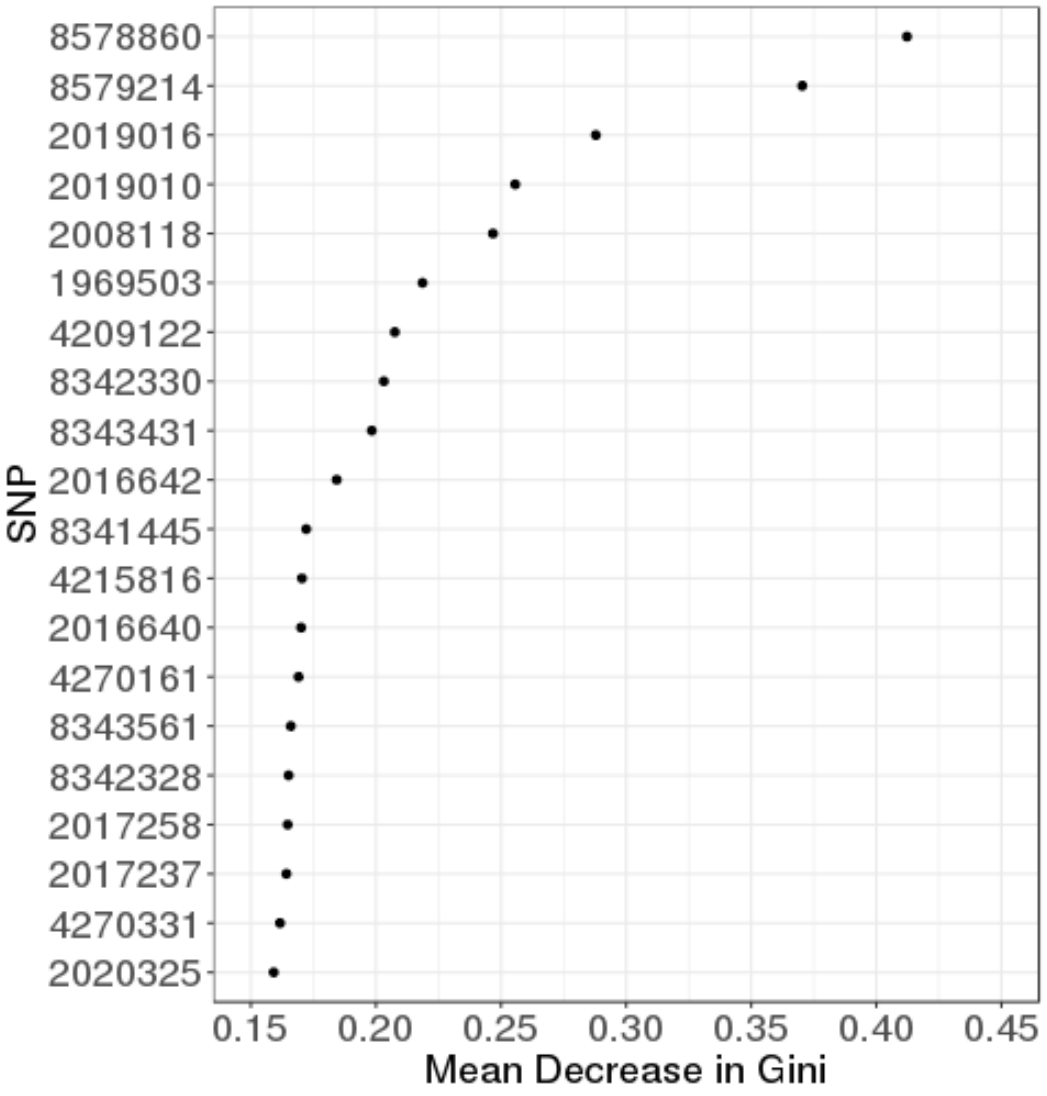


FIGURE 4.1: Random Forests is applied to a bipolar disorder data set with 2,515 SNPs, 1,034 controls, and 1,001 cases. The SNPs are ranked using Gini variable importance. The top 30 SNPs with the largest mean decrease in Gini are shown. The jumps in the mean decrease in Gini provide a guideline as to how many SNPs are considered important predictors of bipolar disorder.

CHAPTER 5
IMPROVING THE INTERPRETATION OF RANDOM FORESTS THROUGH PROXIMITIES

5.1 Introduction

In Chapter 1 Section 1.3.8 we introduced an improved calculation of the proximities in Random Forests. The new proximities perfectly reproduce the Random Forests predictions. The old implementation of Random Forests provides symmetric proximities, and a Multidimensional Scaling plot is used to visualize the proximities in two or three-dimensional space. The new proximities are asymmetric and require an appropriate visualization method for interpretation.

There are existing models that can visualize asymmetric data in two-dimensional space. We can apply those visualization methods to asymmetric proximities.

In Section 5.1.1, we introduce examples of asymmetric data. In Section 5.1.2, we present a matrix decomposition that is often used in visualizing asymmetric proximities. In Section 5.4, we describe an interesting Morse code data set from Rothkopf (1957). We apply two visualization methods from Sections 5.2 and 5.3 to the Morse code data in Section 5.4. In Section 5.5, we introduce a new visualization method and compare the results to those from the existing methods.

5.1.1 Asymmetric Data

Suppose we were to measure distances between the capitals in each state in the United States of America. The distance, say between Boise and Salt Lake City, is the same if we were to measure it from Boise to Salt Lake City or from Salt Lake City to Boise. In matrix form, the rows being the initial starting point and the columns representing the ending point, we would have a square symmetric matrix containing distances between all the capitals. Multidimensional Scaling (MDS) could then be applied to give us a 2-dimensional configuration of the relative distances between the capitals (Cox and

Cox, 2000). Unlike distances, in many different scenarios, we could obtain a square proximity matrix that deviates from symmetry. For example, suppose each person in a classroom was to rate each other's friendship. Joe may not rate his friendship with Jane the same as Jane would rate their friendship, resulting in asymmetry. In another example, subjects were given two Morse code signals and asked to judge whether the two were the same. The i, j element of the proximity matrix contains the number of subjects who thought signals i and j were the same when presented with signal i followed by signal j. The proximity matrix is asymmetric because the presentation order makes a difference (Rothkopf, 1957). Other examples include relationships among managers of a firm (Okada et al., 2005), brand switching among margarine brands (Okada and Tsurumi, 2012), dyadic interactions in a social system (Solanas et al., 2006), branch rivalry of Spanish financial sector restructuring (Sagarra et al., 2014), relationships among soft drink brands (Okada, 2014), and evaluating the effect of a new brand (Okada and Tsurumi, 2014). The traditional approach to model asymmetric data is to assume that the asymmetry is due to some error. If that is the case, we can visualize the asymmetric data by symmetrizing the matrix by replacing the asymmetric pairs by their average and performing MDS on the symmetrized matrix. However, it may be the case that the differences we observe in the asymmetric pairs are not due to error and that the nature of the asymmetry is important. Applying MDS would then be inappropriate, and we would lose information.

5.1.2 Decomposition of Asymmetric Data

Every square asymmetric matrix $\mathbf{P}$ can be uniquely decomposed into a symmetric component and a skew-symmetric component:

$$\mathbf{P} = \mathbf{S} + \mathbf{A} \tag{5.1}$$

where $\mathbf{S}$ is a symmetric matrix of averages $\mathbf{S} = (\mathbf{P} + \mathbf{P}')/2$ and $\mathbf{A}$ is a skew-symmetric matrix defined as $\mathbf{A} = (\mathbf{P} - \mathbf{P}')/2$. The result can be found in numerous linear algebra textbooks and also in Borg and Groenen (2005). We can think of the symmetric matrix $\mathbf{S}$ as departing from asymmetry and the skew-symmetric matrix $\mathbf{A}$ as departing from symmetry. A property of the decomposition of the asymmetric matrix is that the sum-of-squares of the asymmetric matrix $\mathbf{P}$ can be partitioned into the sum-of-squares of the

symmetric matrix and the sum-of-squares of the skew-symmetric matrix, that is,

$$\sum_{i,j} p_{i,j}^2 = \sum_{i,j} s_{i,j}^2 + \sum_{i,j} a_{i,j}^2. \tag{5.2}$$

where $\mathbf{S} = [s_{i,j}]_{i,j=1,\ldots,n}$ and $\mathbf{A} = [a_{i,j}]_{i,j=1,\ldots,n}$ (Borg and Groenen, 2005). Since we can decompose the asymmetric matrix, we can model the two components individually.

Asymmetric proximities can always be decomposed into a symmetric and a skew-symmetric component. Thus, existing methods for asymmetric data do one of three things. They use a special visualization technique just on the asymmetric component, represent the asymmetry along with the symmetric component simultaneously, or directly model the asymmetric matrix. The Gower model visualizes the asymmetric component by applying a singular value decomposition (Constantine and Gower, 1978). Newer models such as modeling skew-symmetry by distances were proposed by Borg and Groenen (2005). Models that incorporate both the symmetric and skew-symmetric component do so by applying MDS on the symmetric component and then embedding the non-symmetric component to the MDS configuration. Borg and Groenen (1996) embedded skew-symmetries as drift vectors. Last, we can analyze asymmetric proximities by directly fitting a model to the asymmetric data. Gower (1977) suggested using a simple distance model called unfolding.

5.2 Gower Model

The Gower model analyzes asymmetric data by applying the singular value decomposition (svd) to the skew-symmetric component. From Gower (1977), the skew-symmetric component can be written in matrix form as $A = V\Sigma JV'$ where $\mathbf{V}$ is an orthogonal matrix, Σ is a diagonal matrix containing singular values $\sigma_1, \sigma_1, \sigma_2, \sigma_2, \ldots$, and $\mathbf{J}$ is a matrix containing the 2×2 matrix $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ along its diagonal. From the svd, we can obtain coordinates of the position of each observation and thus we can get a two-dimensional configuration.

5.3 Drift Vectors in MDS Plots

An alternative to modeling asymmetric data is to simultaneously model the symmetric and skew-symmetric component in one plot. Modeling both the symmetric and

skew-symmetric component can be done by applying MDS to the symmetric component of the asymmetric data then embedding drift vectors to each of the points. The approach makes it possible to see the relationship between the symmetric and skew-symmetric components. The length of the arrows is determined by calculating the unit length between a point and all the other points and multiplying each length by its corresponding skew-symmetry. Averaging the unit lengths will give the length of the vector for a single point. The direction angle relative to the y-axis is determined using trigonometry.

5.4 Morse Code Data Set

The Morse code data set was collected by Rothkopf (1957). The data that we will be using includes 36 Morse codes consisting of 26 letters and ten numbers (0 to 9). Each letter and number are represented by a Morse code that consists of either short or long beeps or both. Five hundred and ninety-eight subjects are presented with 36^2 pairs of Morse code signals which they listened to and were asked to judge whether the two signals were the same or different. Each number in Table 5.1 refers to the percentage of the 598 individuals who said that the two signals were the same. The rows in Table 5.1 are the Morse code signals that were presented first and the columns refer to the signals that were presented second. See Table 5.2 in Appendix 5.7 for the Morse code signal for each letter or digit. Large values away from the diagonal indicate confusion between the signals. Clearly, the relationship is not symmetric. Subjects are more confused for some orderings of the Morse code signals pairs than others. For example, highlighted in red, are the percentages of subjects who said that the pairs XB, BX, YZ, and ZY are the same. A larger percentage of individuals were confused when B was presented before X than vice versa and similarly when Z was presented before Y. However, a higher percentage of individuals thought the pairs BX and XB were the same in comparison to the pairs YZ and ZY. This can be explained by observing that the Morse code signals for B and X are quite similar.

We can determine how asymmetric the Morse code data are by first decomposing the asymmetric data by using Equation 5.1 and using Equation 5.2 to calculate the

TABLE 5.1: The values represent percentage of people out of the 598 that said the Morse code signals were the same (Rothkopf, 1957). See Table refmorse in Appendix refAppendixA for the complete list of the Morse code signals for each letter or digit.

Morse Code		A	B	C	...	X	Y	Z	1	2	3	...	8	9	0
. _	A	92	4	6	...	12	7	3	2	7	5	...	6	2	3
_ . . .	B	5	84	37	...	**84**	30	42	12	17	14	...	17	4	4
_ . _ .	C	4	38	87	...	32	82	38	13	15	31	...	24	18	12
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
_ . . _	X	7	**64**	45	...	91	48	26	12	20	24	...	16	17	6
_ . _ _	Y	9	23	62	...	44	86	**23**	26	44	40	...	33	23	16
_ _ . .	Z	3	46	45	...	36	**42**	87	16	21	27	...	47	15	15
. _ _ _ _	1	2	5	10	...	17	19	22	84	63	13	...	32	57	55
. . _ _ _	2	7	14	22	...	17	30	13	62	89	54	...	21	16	11
. . . _ _	3	3	8	21	...	22	25	12	18	64	86	...	17	8	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
_ _ _ _ .	8	3	23	40	...	24	35	50	42	29	16	...	89	61	26
_ _ _ _ .	9	3	14	23	...	11	21	24	57	39	9	...	56	91	78
_ _ _ _ _	0	9	3	11	...	12	15	20	50	26	9	...	52	81	94

sum-of-squares for the symmetric and skew-symmetric components. From Equation 5.1:

$$\mathbf{P} = \begin{pmatrix} 92 & 4 & 6 & \cdots \\ 5 & 84 & 37 & \cdots \\ 4 & 38 & 87 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$= \begin{pmatrix} 92.0 & 4.5 & 5.0 & \cdots \\ 4.5 & 84.0 & 37.5 & \cdots \\ 5.0 & 37.5 & 87.0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} + \begin{pmatrix} 0.0 & -0.5 & 1.0 & \cdots \\ 0.5 & 0.0 & -0.5 & \cdots \\ -1.0 & 0.5 & 0.0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

From Equation 5.2: $\sum_{i,j} p_{i,j}^2 = 671533 + 26821 = 698354$. Then the proportion of the sum-of-squares due to the symmetric component is $\frac{671533}{698354} = 0.96$ and the proportion of the sum-of-squares due to the skew-symmetric component is $\frac{26821}{698354} = 0.04$. The symmetric portion is dominating, and the proportion of the sum-of-squares explained by the skew-symmetric portion is minor. However, further analysis of the skew-symmetric portion or the asymmetric data as a whole may reveal interesting relationships about the Morse code data.

Figure 5.1 is a Gower diagram for the Morse code data. The plot is interpreted by drawing a triangle with vertices at the origin and two Morse code signals, say B and X. The area of the triangle is a measure of how asymmetric the Morse codes B and X are. In this case, the asymmetry is quite large because one of the orders is more often confused than the other order. To determine which order is more confused in the data, we look at clockwise rotations. The letter that comes first is B; therefore B is more often confused with X than vice versa. We can see why this is the case if we look back at what the Morse code looks like for B and X. The Morse code for B is _ . . . and the Morse code for X is _ . . _ . The two Morse code signals are quite similar. It may be that individuals are more confused if B is presented first then X because the last beep is short for B making it more difficult to know when the signal of B has ended.

When interpreting the Gower model for skew-symmetries, note that if points lie on the vertical or horizontal axes, which are the lines through the origin, then there is no asymmetry since the area of the triangle would be zero. See Gower (1977) for a more in-depth introduction of the Gower model.

A two-dimensional configuration of the model with drift vectors embedded in the MDS plot can be found in Figure 5.2 applied to the Morse code data. The configuration reveals a clear pattern, that is, most of the arrows are directed in the northwest direction. The vertical axis can be interpreted as the length of the Morse code signal. The letters, for example, E, T, M, N, A, and I near the top of the Figure have short Morse code signals, whereas, the digits near the bottom, ranging from two to eight have long Morse signals. The length of the arrows suggests that the shorter Morse code signals tend to be more confused with longer signals than vice versa.

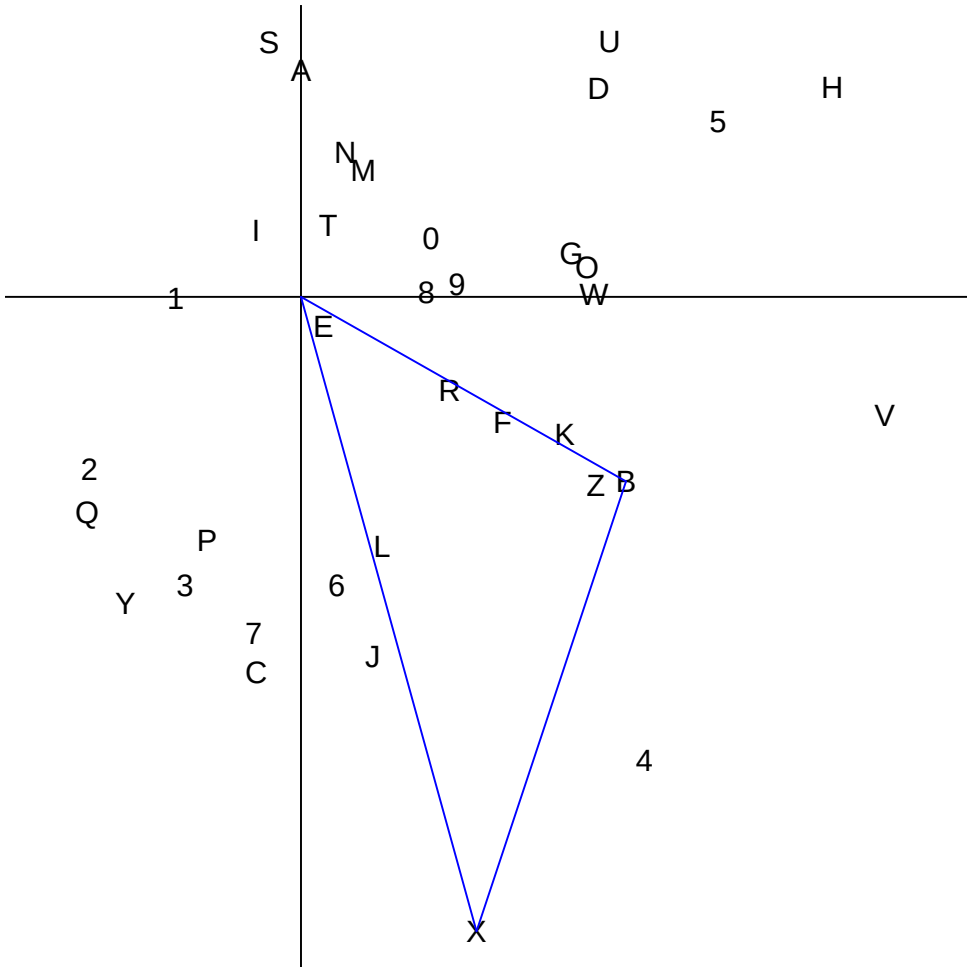


FIGURE 5.1: Gower diagram applied to the Morse code data. See Table 5.2 in Appendix 5.7 for the Morse code signal for each letter or digit. The origin occurs at the intersection of the black lines. Coordinates are obtained from the svd.

5.5 New Method

An alternative way to visualize asymmetric data is to address the asymmetry directly. We propose a simple new model that can be easily interpreted. Given an asymmetric matrix, such as the Morse code data:

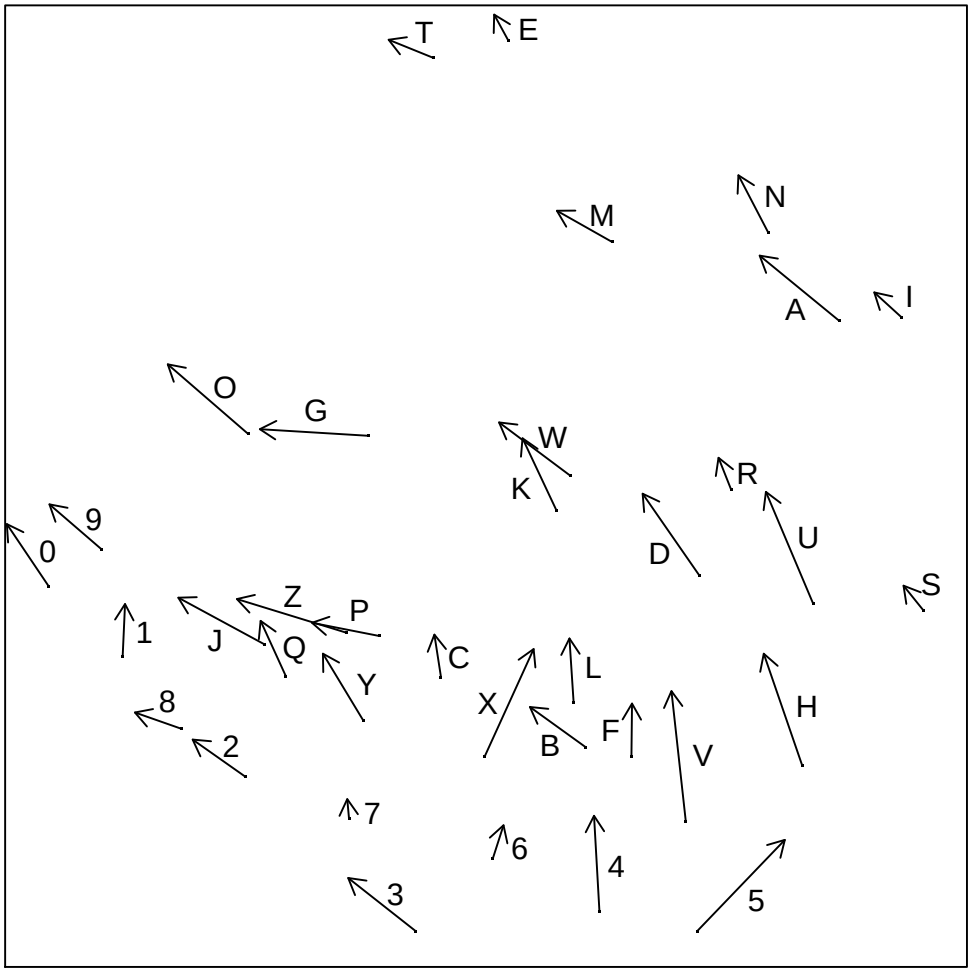


FIGURE 5.2: Multimensional scaling plot with drift vectors representing the asymmetric component applied to the Morse code data set. See Table 5.2 in Appendix 5.7 for the Morse code signal for each letter or digit.

$$\tilde{\mathbf{P}} = \begin{pmatrix} 0.92 & 0.04 & 0.06 & \cdots \\ 0.05 & 0.84 & 0.37 & \cdots \\ 0.04 & 0.38 & 0.87 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

we can symmetrize the matrix $\tilde{\mathbf{P}}$ by reflecting the upper triangle of $\tilde{\mathbf{P}}$ about it's diagonal (replacing the existing lower triangle) to obtain:

$$\mathbf{U} = \begin{pmatrix} 0.92 & 0.04 & 0.06 & \cdots \\ 0.04 & 0.84 & 0.37 & \cdots \\ 0.06 & 0.37 & 0.87 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Similarly, we can symmetrize $\tilde{\mathbf{P}}$ by taking the lower triangle of $\tilde{\mathbf{P}}$ and reflecting it to give us a new upper triangle:

$$\mathbf{L} = \begin{pmatrix} 0.92 & 0.05 & 0.04 & \cdots \\ 0.05 & 0.84 & 0.38 & \cdots \\ 0.04 & 0.38 & 0.87 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Since both matrices are symmetric, we can run MDS on each of them to give us two two-dimensional plots. It is preferable to visualize the result in one plot, so we include each of the configurations in one figure. Since the two configurations may be off by some rotation, we rotated to minimize the sum of the distances between each of the two positions of the entities. See Figure 5.3 for the visualization applied to the Morse code data. The points for the two possible solutions for each Morse code signal are connected giving us an overall idea of how big the difference is between the two MDS configurations. If the Morse code data set were close to symmetry, then we would expect the solid and open points to be quite close to each other. However, this is not the case, giving us a reason to believe that an asymmetric model is more appropriate for the Morse code data set.

Notice that the relative positions of the Morse code signals in Figure 5.2 and Figure 5.3 are quite similar, for example, letter E, T, M, N, A, and I are near each other. One of the major differences between Figure 5.2 and Figure 5.3 is the Morse code signal for 7. The length of the arrow for 7 in Figure 5.2 is quite short while the line segment is quite long in Figure 5.3. Looking back at the Morse code proximities, it shows that 15% of the subjects said the signals for 7 and 4 are the same, but when the signals were presented in reverse order, the percent of subjects claiming the signal are the same increases to 32%. The new method proposed may introduce new insight to asymmetric proximity data and perhaps may be more interpretable.

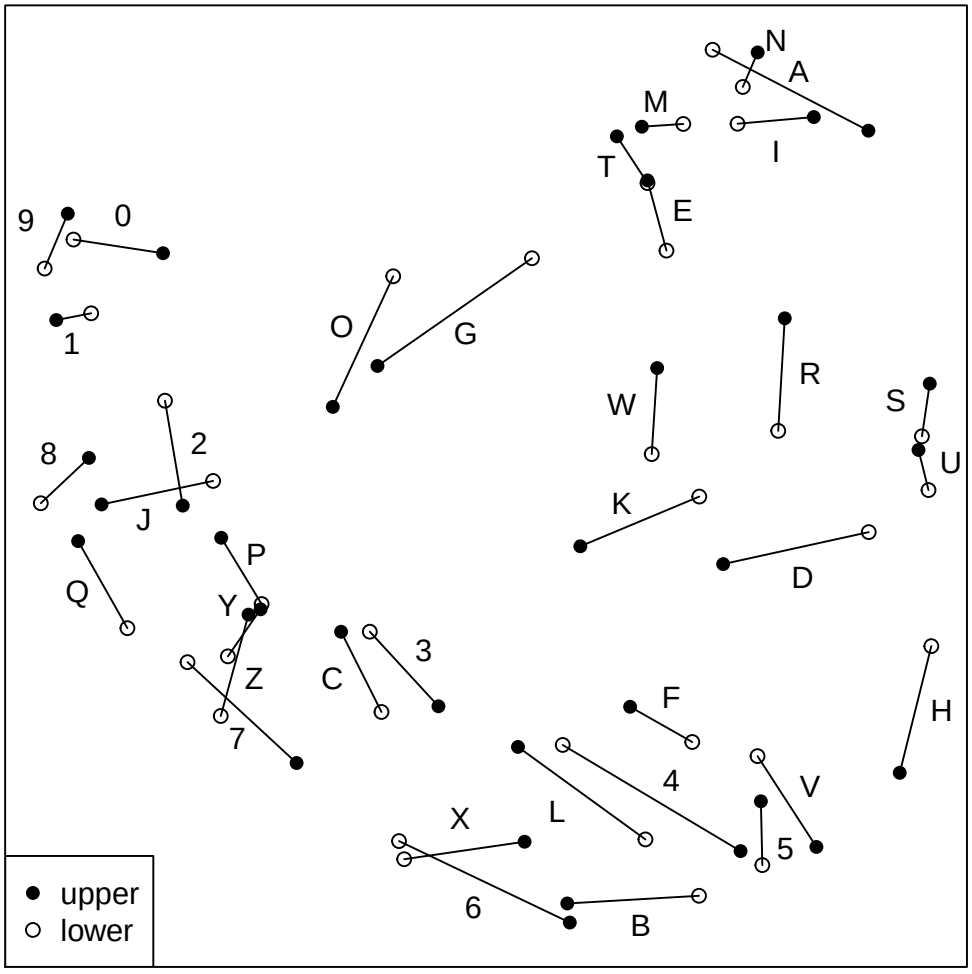


FIGURE 5.3: Two-dimensional configuration representation of the asymmetry in the Morse code data set. Solid dots represent the configuration of the MDS plot applied to the matrix symmetrized from the upper triangle, whereas, the open dots use the matrix symmetrized by the lower triangle. See Table 5.2 in Appendix 5.7 for the Morse code signal for each letter or digit.

5.6 Conclusions

The current implementation of Random Forests provides a symmetric proximity matrix informing us the proportion of times two observations fall in the same terminal node. Improvement in the interpretation of Random Forests is done through the proximities. The new proximities perfectly reproduce the prediction accuracy in Random Forests. However, the new proximities are asymmetric. It would not be suitable to visualize the asymmetric proximities using a MDS plot. Existing asymmetric models such as the Gower model or embedding the skew-symmetric component into an MDS

plot can be used to visualize asymmetric data.

We introduced a new visualization model for asymmetric data. The model can be applied to asymmetric data that weren't produced from a Random Forests model. The new visualization model directly fits a model to the asymmetric data. The asymmetric data is split into two components, the upper and lower triangle. Each component is symmetrized about it's diagonal, and MDS is applied to each matrix. Both configurations are combined into a single plot, and a line segment is drawn for each entity. The length of the line measure the size of the asymmetry.

We compare our new model against the Gower model and a model that embeds the skew-symmetric component into an MDS plot using the Morse code data set. The Gower model shows the asymmetry of pairs of entities, while our new method shows the asymmetry of the entity itself. The Gower model interprets asymmetry through the areas of the triangles. The model that embeds the skew-symmetric component interprets the asymmetry through the axes and the direction and length of the arrows. The Gower model and embedding drift vectors into an MDS plot would be impossible to understand for larger data sets. It would be easier to interpret the asymmetries for big data using our new model because it would be easier to identify large segments.

5.7 Appendix

TABLE 5.2: Table of the letters and digits corresponding Morse code signal.

Morse Code	Letter/Digit
. _	A
_ . . .	B
_ . _ .	C
_ . .	D
.	E
. . _ .	F
_ _ .	G
. . . .	H
. .	I
. _ _ _	J
_ . _	K
. _ . .	L
_ _	M
_ .	N
_ _ _	O
. _ _ .	P
_ _ . _	Q
. _ .	R
. . .	S
_	T
. . _	U
. . . _	V
. _ _	W
_ . . _	X
_ . _ _	Y
_ _ . .	Z
. _ _ _ _	1
. . _ _ _	2
. . . _ _	3
. . . . _	4
. . . . .	5
_	6
_ _ . . .	7
_ _ _ . .	8
_ _ _ _ .	9
_ _ _ _ _	0

CHAPTER 6
FUTURE WORK AND CONCLUSIONS

In Chapter 2 we presented a new filtering method. The filtering method could further be evaluated the following ways:

1. How well it can capture three-way interactions and compare it to methods proposed by González-Domínguez and Schmidt (2015), Leem et al. (2014), Moore et al. (2006), or Guo et al. (2014a).
2. How well it can handle the incorporation of genetic heritability and linkage disequilibrium.
3. How well it can capture multiple embedded epistatic interactions.
4. How well it can capture continuous and mixed dependent variable data type interactions.
5. Compare against faster methods than BOOST. In Niel et al. (2015), they stated that Genome-Wide Interaction Search (GWIS) is faster than BOOST. However, Goudey et al. (2013) found that Rapid (RApid Pair IDentification) was faster than GWIS.

Quite of few interaction detection methods incorporate a χ^2 test, but not all are equivalent. It would be interesting to evaluate the performance of a χ^2 test with 8 degrees of freedom (9×2 table) against 4 degrees of freedom ($3 \times 3 \times 2$ table).

In Chapter 4, there are many approaches to impute missing values to the real data set. Further analysis of the bipolar disorder data set could be done:

1. Carry out a replication study to determine if the interacting SNPs found are associated with bipolar disorder.

2. Compare results with BOOST and the exhaustive χ^2 test.
3. Look at all chromosomes.
4. Look at SNP-environment interactions.
5. Carry out a single SNP analysis using new filtering method.

The new proximities introduced in Chapter 1 and the new visualization method to visualize the proximities could be further explored by:

1. Applying the new visualization method to simulated data to better understand when asymmetry occurs in Random Forests.
2. Using interactive plots to help interpret the asymmetric data.
3. Providing a numerical value to measure how asymmetric an entity is.
4. Building a single tree to represent the asymmetric data.
5. Evaluating how well the proximities detect outliers.
6. Providing numerical values that identify outliers.

Methods compared against our filtering method presented in Chapter 2 and models compared against our visualization method presented in Chapter 5 are not available in the libraries found in R. Potential packages could be created in R:

1. BOOST.
2. New filtering method.
3. Gower Model.
4. Drift Vectors in MDS plot.
5. New Visualization Method.

REFERENCES

Al-jouie, A., Esfandiari, M., Ramakrishnan, S., Roshan, U., 2015. Chi8: a GPU program for detecting significant interacting SNPs with the chi-square 8-df test. BMC Research Notes 8 (1), doi: 10.1186/s13104–015–1392–5.

Barrett, J. C., Cardon, L. R., 2006. Evaluating coverage of Genome-Wide Association studies. Nature Genetics 38, 659–662.

Biau, G., Scornet, E., 2016. A Random Forest guided tour. TEST 25 (2), 197–227.

Borg, I., Groenen, P., 1996. Asymmetries in multidimensional scaling. SoftStat '95: Advances in Statistical Software, Gustav Fisher, Stuttgart, pp. 31–35.

Borg, I., Groenen, P. J., 2005. Modeling asymmetric data. In: Modern Multidimensional Scaling: Theory and Applications, 2nd Edition. Springer Science & Business Media, New York.

Bosch, A., Zisserman, A., Munoz, X., 2007. Image classification using Random Forests and ferns. In: 2007 IEEE 11th International Conference on Computer Vision. IEEE, pp. 1–8.

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Breiman, L., Cutler, A., 2014. Random Forests. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (last accessed May 12, 2015).

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Wadsworth and Brooks, Monterey, CA.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., Wacholder, S., 2006. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. The American Journal of Human Genetics 79 (6), 1002–1016.

Clayton, D., 2015. snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.22.0.
URL http://bioconductor.org/packages/snpStats/

Consortium, I. H., 2005. A haplotype map of the human genome. Nature 437, 1299–1320.

Constantine, A., Gower, J., 1978. Graphical representation of asymmetric matrices. Applied Statistics 27 (3), 297–304.

Cox, T. F., Cox, M. A., 2000. Multidimensional scaling. CRC Press, London.

Cutler, A., Cutler, D. R., Stevens, J. R., 2012. Random forests. In: Zhang C., M. Y. (Ed.), Ensemble Machine Learning. Springer, Boston, MA, pp. 157–175.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J., 2007. Random Forests for classification in ecology. Ecology 88 (11), 2783–2792.

De Lobel, L., Geurts, P., Baele, G., Castro-Giner, F., Kogevinas, M., Van Steen, K., 2010. A screening methodology based on Random Forests to improve the detection of gene–gene interactions. European Journal of Human Genetics 18 (10), 1127–1132.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research 15 (1), 3133–3181.

Fisher, R. A., 1918. XV.–the correlation between relatives on the supposition of mendelian inheritance. Transactions of The Royal Society of Edinburgh 52 (2), 399–433.

Goldstein, B. A., Hubbard, A. E., Cutler, A., Barcellos, L. F., 2010. An application of Random Forests to a Genome-Wide Association dataset: methodological considerations & new findings. BMC Genetics 11 (1), doi:10.1186/1471–2156–11–49.

González-Domínguez, J., Schmidt, B., 2015. GPU-accelerated exhaustive search for third-order epistatic interactions in case–control studies. Journal of Computational Science 8, 93–100.

Google Scholar, 2017. Random Forests. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=mXSv_1UAAAAJ&citation_for_view=mXSv_1UAAAAJ:d1gkVwhDpl0C (last accessed April 4, 2017).

Goudey, B., Rawlinson, D., Wang, Q., Shi, F., Ferra, H., Campbell, R. M., Stern, L., Inouye, M. T., Ong, C. S., Kowalczyk, A., 2013. GWIS-model-free, fast and exhaustive search for epistatic interactions in case-control GWAS. BMC Genomics 14(Suppl 3) (S10), doi:10.1186/1471–2164–14–S3–S10.

Gower, J. C., 1977. The analysis of asymmetry and orthogonality. In: Barra, J., Brodeau, F., Romier, G., van Cutsen, B. (Eds.), Recent Developments in Statistics. Vol. 1. North Holland, Amsterdam, pp. 109–123.

Grabmeier, J. L., Lambe, L. A., 2007. Decision trees for binary classification variables grow equally with the Gini impurity measure and Pearson's chi-square test. International Journal of Business Intelligence and Data Mining 2 (2), 213–226.

Guo, X., Meng, Y., Yu, N., Pan, Y., 2014a. Cloud computing for detecting high-order Genome-Wide epistatic interaction via dynamic clustering. BMC Bioinformatics 15 (102), doi:10.1186/1471–2105–15–102.

Guo, X., Yu, N., Gu, F., Ding, X., Wang, J., Pan, Y., 2014b. Genome-Wide interaction-based association of human diseases-a survey. Tsinghua Science and Technology 19 (6), 596–616.

Hu, X., Liu, Q., Zhang, Z., Li, Z., Wang, S., He, L., Shi, Y., 2010. SHEsisEpi, a GPU-enhanced Genome-Wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. Cell Research 20, doi:10.1038/cr.2010.68.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., 2008. Random survival forests. The Annals of Applied Statistics 2 (3), 841–860.

Kim, Y., Wojciechowski, R., Sung, H., Mathias, R. A., Wang, L., Klein, A. P., Lenroot, R. K., Malley, J., Bailey-Wilson, J. E., 2009. Evaluation of Random Forests performance for Genome-Wide Association studies in the presence of interaction effects. In: BMC Proceedings. Vol. 3. BioMed Central, pp. doi:10.1186/1753–6561–3–S7–S64.

Larivière, B., Van den Poel, D., 2005. Predicting customer retention and profitability by using Random Forests and regression forests techniques. Expert Systems with Applications 29 (2), 472–484.

Leem, S., Jeong, H.-h., Lee, J., Wee, K., Sohn, K.-A., 2014. Fast detection of high-order epistatic interactions in Genome-Wide Association studies using information theoretic measure. Computational Biology and Chemistry 50, 19–28.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.
URL http://CRAN.R-project.org/doc/Rnews/

Lin, H.-Y., Ann Chen, Y., Tsai, Y.-Y., Qu, X., Tseng, T.-S., Park, J. Y., 2012. TRM: a powerful two-stage machine learning approach for identifying SNP-SNP interactions. Annals of Human Genetics 76 (1), 53–62.

Lin, Y., Jeon, Y., 2006. Random Forests and adaptive nearest neighbors. Journal of the American Statistical Association 101 (474), 578–590.

Lunetta, K. L., Hayward, L. B., Segal, J., Van Eerdewegh, P., 2004. Screening large-scale association study data: exploiting interactions using Random Forests. BMC Genetics 5 (1), doi:10.1186/1471–2156–5–32.

Marchini, J., Donnelly, P., Cardon, L. R., 2005. Genome-Wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics 37 (4), 413–417.

Meinshausen, N., 2006. Quantile regression forests. The Journal of Machine Learning Research 7, 983–999.

Mersmann, O., 2015. microbenchmark: accurate timing functions. R package version 1.4-2.1.
URL https://CRAN.R-project.org/package=microbenchmark

Moore, J. H., Gilbert, J. C., Tsai, C.-T., Chiang, F.-T., Holden, T., Barney, N., White, B. C., 2006. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. Journal of Theoretical Biology 241 (2), 252–261.

Murray, C. J., Lopez, A. D., 1996. Evidence-based health policy–lessons from the global burden of disease study. Science 274 (5288), 740–743.

National Library of Medicine, 2017. What are single nucleotide polymorphisms (SNPs)? https://ghr.nlm.nih.gov/primer/genomicresearch/snp (last accessed April 27, 2017).

Niel, C., Sinoquet, C., Dina, C., Rocheleau, G., 2015. A survey about methods dedicated to epistasis detection. Frontiers in Genetics 6 (285).

Okada, A., 2014. Analysis of asymmetric relationships among soft drink brands. In: Gaul, W., Geyer-Schulz, A., Baba, Y., Okada, A. (Eds.), German-Japanese Interchange of Data Analysis Results. Springer, Cham, pp. 147–156.

Okada, A., Imaizumi, T., Inoue, H., 2005. Asymmetric multidimensional scaling of relationships among managers of a firm. In: Baier, D., Decker, R., Schmidt-Thieme, L. (Eds.), Data Analysis and Decision Support. Springer, Berlin, pp. 100–107.

Okada, A., Tsurumi, H., 2012. Asymmetric multidimensional scaling of brand switching among margarine brands. Behaviormetrika 39 (1), 111–126.

Okada, A., Tsurumi, H., 2014. Evaluating the effect of new brand by asymmetric multidimensional scaling. In: Vicari, D., Okada, A., Ragozini, G., Weihs, C. (Eds.), Analysis and Modeling of Complex Data in Behavioral and Social Sciences. Springer, Cham, pp. 201–209.

Palmer, D. S., O'Boyle, N. M., Glen, R. C., Mitchell, J. B., 2007. Random Forest models to predict aqueous solubility. Journal of Chemical Information and Modeling 47 (1), 150–158.

Prabhu, S., Pe'er, I., 2012. Ultrafast Genome-Wide scan for SNP–SNP interactions in common complex disease. Genome Research 22, 2230–2240.

R Core Team, 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Rothkopf, E. Z., 1957. A measure of stimulus similarity and errors in some paired-associate learning tasks. Journal of Experimental Psychology 53 (2), 94–101.

Sagarra, M., Busing, F. M., Mar-Molinero, C., Rialp, J., 2014. Assessing the asymmetric effects on branch rivalry of Spanish financial sector restructuring. Advances in Data Analysis and Classification, doi: 10.1007/s11634–014–0186–2.

Schwarz, D. F., König, I. R., Ziegler, A., 2010. On safari to random jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics 26 (14), 1752–1758.

Shang, J., Zhang, J., Sun, Y., Liu, D., Ye, D., Yin, Y., 2011. Performance analysis of novel methods for detecting epistasis. BMC Bioinformatics 12 (1), doi:10.1186/1471–2105–12–475.

Slatkin, M., 2008. Linkage disequilibriumunderstanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9, 477–485.

Smith, E. N., Bloss, C. S., Badner, J. A., Barrett, T., Belmonte, P. L., Berrettini, W., Byerley, W., Coryell, W., Craig, D., Edenberg, H. J., et al., 2009. Genome-Wide Association study of bipolar disorder in European American and African American individuals. Molecular Psychiatry 14 (8), 755–763.

Smith, E. N., Koller, D. L., Panganiban, C., Szelinger, S., Zhang, P., Badner, J. A., Barrett, T. B., Berrettini, W. H., Bloss, C. S., Byerley, W., et al., 2011. Genome-Wide Association of bipolar disorder suggests an enrichment of replicable associations in regions near genes. PLoS Genetics 7 (6), doi:10.1371/journal.pgen.1002134.

SNPedia, 2017. https://www.snpedia.com/index.php/Bipolar_disorder (last accessed April 20, 2017).

Sohn, K.-A., Wee, K., 2015. A comment on two-locus epistatic interaction models for Genome-Wide Association studies. Journal of Bioinformatics and Computational Biology 13 (6), doi:10.1142/S0219720015710043.

Solanas, A., Salafranca, L., Riba, C., Sierra, V., Leiva, D., 2006. Quantifying social asymmetric structures. Behavior Research Methods 38 (3), 390–399.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in Random Forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8 (25), doi:10.1186/1471–2105–8–25.

Tang, W., Wu, X., Jiang, R., Li, Y., 2009. Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy. PLoS Genetics 5 (5), doi:10.1371/journal.pgen.1000464.

Upstill-Goddard, R., Eccles, D., Fliege, J., Collins, A., 2012. Machine learning approaches for the discovery of gene–gene interactions in disease data. Briefings in Bioinformatics, doi:10.1093/bib/bbs024.

Wainberg, M., Alipanahi, B., Frey, B. J., 2016. Are Random Forests truly the best classifiers? Journal of Machine Learning Research 17 (110), 1–5.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., Yu, W., 2010a. 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.

Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N. L., Yu, W., 2010b. Predictive rule inference for epistatic interaction detection in Genome-Wide Association studies. Bioinformatics 26 (1), 30–37.

Wang, Y., Liu, G., Feng, M., Wong, L., 2011. An empirical comparison of several recent epistatic interaction detection methods. Bioinformatics 27 (21), 2936–2943.

Wang, Y., Liu, X., Robbins, K., Rekaya, R., 2010. Antepiseeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. BMC Research Notes 3 (1), doi:10.1186/1756–0500–3–117.

Winham, S. J., Colby, C. L., Freimuth, R. R., Wang, X., de Andrade, M., Huebner, M., Biernacka, J. M., 2012. SNP interaction detection with Random Forests in high-dimensional genetic data. BMC Bioinformatics 13 (1), doi:10.1186/1471–2105–13–164.

Wolpert, D. H., 1996. The lack of a priori distinctions between learning algorithms. Neural Computation 8 (7), 1341–1390.

Wright, M. N., Ziegler, A., König, I. R., 2016. Do little interactions get lost in dark Random Forests? BMC Bioinformatics 17 (1), doi:10.1186/s12859–016–0995–8.

Wu, J., Devlin, B., Ringquist, S., Trucco, M., Roeder, K., 2010. Screen and clean: a tool for identifying interactions in Genome-Wide Association studies. Genetic Epidemiology 34 (3), 275–285.

Yang, C., He, Z., Wan, X., Yang, Q., Xue, H., Yu, W., 2009. SNPHarvester: a filtering-based approach for detecting epistatic interactions in Genome-Wide Association studies. Bioinformatics 25 (4), 504–511.

Zhang, X., Huang, S., Zou, F., Wang, W., 2010. TEAM: efficient two-locus epistasis tests in human Genome-Wide Association study. Bioinformatics 26 (12), i217–i227.

Zhang, X., Zou, F., Wang, W., 2009. FastChi: an efficient algorithm for analyzing gene-gene interactions. In: Pacific Symposium on Biocomputing. NIH Public Access, pp. 528–539.

CURRICULUM VITAE

Anna Quach

anna.quach@aggiemail.usu.edu

EDUCATION

Ph.D. in Mathematical Sciences with an emphasis in Statistics 2012–2017

Utah State University, Logan, Utah

MS Statistics 2010–2012

Utah State University, Logan, Utah

B.Sc. Applied Mathematics 2008–2012

Utah State University, Logan, Utah

AE Electrical Engineering 2008–2010

AE Mechanical Engineering

AE Civil Engineering

College of Southern Idaho, Twin Falls, Idaho

PhD DISSERTATION

Title: Extensions and Improvements to Random Forests for Classification

Advisor: Adele Cutler

Description:

- Developed a new visualization method for asymmetric proximities.
- Improved Random Forests on data from the Genome Wide Association Study.
- Developed an effective filtering method to reduce high-dimensional genetic data.

MASTERS THESIS

Title: Interactive Random Forests Plots

Advisor: Adele Cutler

Description:

- Created a function, called irfplot (interactive random forests plot) that specifically uses Random Forests to produce interactive graphs.
- Used the interactive Random Forests plot to explore the nutrition data set from the Cache County Memory Study.

PUBLICATIONS

- Quach, A., Symanzik, J., Forsgren, N. (2015): Soul of the Community: An Attempt to Assess Attachment to a Community, *Computational Statistics*, Accepted
- Wengreen, H., Munger, R. G., Cutler, A., Quach, A., Bowles, A., Corcoran, C., Tschanz, J. T., Norton, M. C., Welsh-Bohmer, K. (2013): A Prospective study of Dietary Approaches to Stop Hypertension- and Mediterranean-style dietary patterns and age-related cognitive change: the Cache County Study on Memory, Health and Aging. *The American Journal of Clinical Nutrition*, 98, 1263–1271
- Wengreen, H., Quach, A., Cutler, A., Munger, R., Corcoran, C. (2012): Whole-grain intake and risk of all-cause mortality among elderly men and women: the Cache, County Study on Memory, Health and Aging. The FASEB Journal, 26, 119–120
- Wengreen, H., Corcoran, C., Cutler, A., Munger, R., Quach, A., Tschanz, J., Ward, R. (2012): Erythrocyte omega-3 fatty acid concentrations and cognitive function: The Cache County Study on Memory and Aging. *Alzheimer's & Dementia*, 8(4), 449
- Quach, A. T. (2012). Interactive Random Forests Plots. All Graduate Plan B and other Reports. 134. http://digitalcommons.usu.edu/gradreports/134

CONFERENCE PROCEEDINGS

- Quach, A., Cutler, A. (2016): Archetypal Analysis: Three Case Studies, In *JSM Proceedings*, Statistical Learning and Data Science Section. Alexandria, VA: American Statistical Association.
- Quach, A., Cutler, A. (2015): A New Visualization Method for Asymmetric Proximity Data, In *JSM Proceedings*, Statistical Graphics Section. Alexandria, VA: American Statistical Association.

- Quach, A., Symanzik, J., Valesquez, N. F. (2013): Soul of the Community: A First Attempt to Assess Attachment to a Community, In *JSM Proceedings*, Statistical Graphics Section. Alexandria, VA: American Statistical Association.

AWARDS AND HONORS

USU Department of Mathematics and Statistics 2016 – 2017
Ph.D. Researcher of the Year

USU Department of Mathematics and Statistics 2013 – 2014
Excellence in Research Award

USU Department of Mathematics and Statistics 2009 – 2010
Outstanding Undergraduate Mathematics Award

TRAVEL AWARDS

XSEDE HPC Workshop: BIG DATA 2017
$92.51, Department of Mathematics and Statistics, USU

Women in Statistics and Data Science 2016
$800, Department of Mathematics and Statistics, USU

The Joint Statistical Meeting, Chicago, Illinois 2016
$800, Department of Mathematics and Statistics, USU
$300, RGS Graduate Student, USU

Summer Institute for Big Data at University of Washington 2016
$1850 (3 courses + $500), University of Washington
$700, Department of Mathematics and Statistics, USU

Statistical and Applied Mathematical Sciences Institute (SAMSI) 2016
Spring Opportunities Workshop for Women in Math Sciences
Hotel + $400, SAMSI
$161, Department of Mathematics and Statistics, USU

The Joint Statistical Meeting, Seattle, Washington 2015
$800, Department of Mathematics and Statistics, USU
$300, RGS Graduate Student, USU

The Joint Statistical Meeting, Montreal, Canada 2013

$600, Department of Mathematics and Statistics, USU

$500, Women and Gender Studies, USU

$400, RGS Graduate Student, USU

PRESENTATIONS AND POSTERS AT CONFERENCES AND WORKSHOPS

Invited Talks:

"Tips and Tricks for Building a Random Forests Classifier" 2017

AI With the Best

"Random Forests: The Vanilla of Machine Learning" 2016

AI With the Best

Oral Presentations:

"Archetypal Analysis and Its Application" 2016

The Joint Statistical Meeting, Chicago, Illinois

"Modeling Asymmetric Proximities" 2015

The Joint Statistical Meeting, Seattle, Washington

Poster and Oral Presentations:

"Detecting SNP–SNP Interactions With Random Forests" 2016

Women in Statistics and Data Science

"Detecting Epistatic Interactions" 2016

Statistical and Applied Mathematical Sciences Institute

Spring Opportunities Workshop for Women in Math Sciences

Posters:

"Soul of the Community" 2013

The Joint Statistical Meeting, Montreal, Canada

SERVICE: MANUSCRIPT REVIEW

Computational Statistics 2016

Computational Statistics 2016

Computational Statistics 2014

Computational Statistics	2013
Journal of Computational and Graphical Statistics	2013
The R Journal	2013

TEACHING EXPERIENCE

Instructor

Introduction to R	Spring 2017
Introduction to Statistics	Fall 2016
Business Statistics	Summer 2016
Introduction to Statistics	Fall 2015
Introduction to Statistics	Summer 2015
Introduction to Statistical Methods	Fall 2014
Introduction to Statistics	Summer 2014

Recitation Leader

Introduction to Statistics with Elements in Algebra	Spring 2016

Head Teaching Assistant

Introduction to Statistics	Spring 2014
Business Statistics	Fall 2013

SAS Tutor

Design of Experiments	Summer 2013

Grader

Statistical Computing	Spring 2016
Applied Multivariate Statistics	Spring 2014

RESEARCH EXPERIENCE

PhD Dissertation 2012 – 2017

Research Assistant for Dr. Guifang Fu 2015

- Analyzed leaves of various shapes with the goal to characterize leaves with similar shapes.
- Used Archetypal Analysis to identify extremes or unusual leaves and express all other leaves as a mixture of the identified extremes.

- Identify which genes affect the shape of leaves using linear regression and Random Forests.

Research Assistant for Dr. Heidi Wengreen 2010 – 2013

- Determined if there is an association between dietary pattern, cognitive decline and Alzheimer's disease.
- Determined the association between fatty acids and risk of cognitive decline.
- Examined the association between whole-grain consumption and all-cause mortality.

Master Thesis Project 2010 – 2012

SOCIETY MEMBERSHIPS

American Statistical Association (ASA)	2013 – present
ASA Utah Chapter	2010 – present
Society for Industrial and Applied Mathematics (SIAM)	2016 – present
Kaggle USU	2016 – present
Caucas for Women in Statistics (CWS)	2016 – present

REFERENCES

Adele Cutler, Professor
Utah State University
Department of Mathematics and Statistics
3900 Old Main Hill
Logan, Utah 84322-3900
Phone: 435.797.2761
Email: adele.cutler@usu.edu

Jürgen Symanzik, Professor
Utah State University
Department of Mathematics and Statistics
3900 Old Main Hill
Logan, Utah 84322-3900
Phone: 435.797.0696

Email: symanzik@math.usu.edu

Heidi Wengreen, Associate Professor

Utah State University

Department of Nutrition, Dietetics, and Food Sciences

8700 Old Main Hill

Logan, Utah 84322-8700

Phone: 435.797.1806

Email: heidi.wengreen@usu.edu