2016

Aggregated Quantitative Multifactor Dimensionality Reduction

Rebecca E. Crouch
University of Kentucky, rebecca.crouch@uky.edu
Digital Object Identifier: https://doi.org/10.13023/ETD.2016.525

STUDENT AGREEMENT:

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

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

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

REVIEW, APPROVAL AND ACCEPTANCE

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

Rebecca E. Crouch, Student

Dr. Richard Charnigo, Major Professor

Dr. Constance Wood, Director of Graduate Studies

AGGREGATED QUANTITATIVE MULTIFACTOR DIMENSIONALITY REDUCTION

DISSERTATION

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

By

Rebecca Crouch

Lexington, Kentucky

Director: Dr. Richard Charnigo, Professor of Statistics

Lexington, Kentucky

ABSTRACT OF DISSERTATION

AGGREGATED QUANTITATIVE MULTIFACTOR DIMENSIONALITY REDUCTION

We consider the problem of making predictions for quantitative phenotypes based on gene-to-gene interactions among selected Single Nucleotide Polymorphisms (SNPs). Previously, Quantitative Multifactor Dimensionality Reduction (QMDR) has been applied to detect gene-to-gene interactions associated with elevated quantitative phenotypes, by creating a dichotomous predictor from one interaction which has been deemed optimal. We propose an Aggregated Quantitative Multifactor Dimensionality Reduction (AQMDR), which exhaustively considers all k-way interactions among a set of SNPs and replaces the dichotomous predictor from QMDR with a continuous aggregated score. We evaluate this new AQMDR method in a series of simulations for two-way and three-way interactions, comparing the new method with the original QMDR. In simulation, AQMDR yields consistently smaller prediction error than QMDR when more than one significant interaction is present in the simulation model. Theoretical support is provided for the method, and the method is applied on Alzheimer's Disease (AD) data to identify significant interactions between *APOE4* and other AD associated SNPs.

KEYWORDS: QMDR, quantitative Traits, MDR, phenotypes, SNP, interaction

REBECCA CROUCH
Student's Signature

DECEMBER 7, 2016
Date

AGGREGATED QUANTITATIVE MULTIFACTOR DIMENSIONALITY REDUCTION

By

Rebecca Crouch

RICHARD CHARNIGO
Director of Dissertation

CONSTANCE WOOD
Director of Graduate Studies

DECEMBER 7, 2016
Date

ACKNOWLEDGEMENTS

This dissertation is the product of a collaborative effort and was made possible by the support and enthusiasm demonstrated by all of the members of my dissertation committee. I would like to express my sincerest gratitude to the individuals who assisted with this research endeavor.

A debt of gratitude is owed to my advisor, Dr. Richard Charnigo for his leadership, encouragement, and thoughtful insight pertaining to all aspects of this work. Without his guidance and persistent assistance, this dissertation would not have been possible.

I would like to express my deepest appreciation to committee member Dr. David Fardo, whose expertise in genetic applications of statistics and willingness to help were invaluable.

I would like to thank committee member Dr. Katherine Thompson for the computational expertise she provided.

In addition, thank you to committee members Dr. Solomon Harrar and Dr. William Griffith for providing advice and insight regarding the theoretical considerations in this work.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Multifactor Dimensionality Reduction

The detection and characterization of susceptibility genes for common complex diseases such as atrial fibrillation, autism, breast cancer, and hypertension (among many others) has long been a concern in the field of genetics. These diseases often have inheritance patterns that are complex [2]. Due to advancements in the availability and cost-efficiency of genotypic data, Single Nucleotide Polymorphisms (SNPs) have been used to explore variation in susceptibility to such diseases [1]. Abundance of such data has raised concerns regarding traditional, parametric statistical methods. For example, logistic regression is frequently used for models involving categorical predictor variables and categorical response variables [2]. However, using genotypes as predictors in these models and a phenotype (or clinical outcome) as the response is not ideal, as logistic regression is ill-equipped to handle high-dimensional data which may result in contingency table cells which are empty [3]. It is possible to remedy this problem with large sample sizes, but of course, this may come at high expense to researchers. To avoid this, alternative non-parametric solutions have been developed in recent years, which can be applied to smaller sample sizes from case-control and discordant-sib-pair studies [2].

One such alternative solution, Multifactor Dimensionality Reduction (MDR) was developed by Ritchie et al. [2], and inspired by the combinatorial partitioning method [4]. MDR is a method for characterizing and identifying nonlinear complex gene-to-gene interactions which may be related to susceptibility to complex diseases. MDR converts high dimensional genotypic data into a single predictive variable by identifying gene-to-gene interactions as "high risk" or "low risk" [1]. MDR is appealing be-

cause it can identify non-linear epistasis (significant gene-to-gene interactions) with no assumptions about underlying distributions of predictors (genotypes) or phenotypic outcomes (apart from the dichotomous nature of these outcomes), and no requirement for large sample sizes [5].

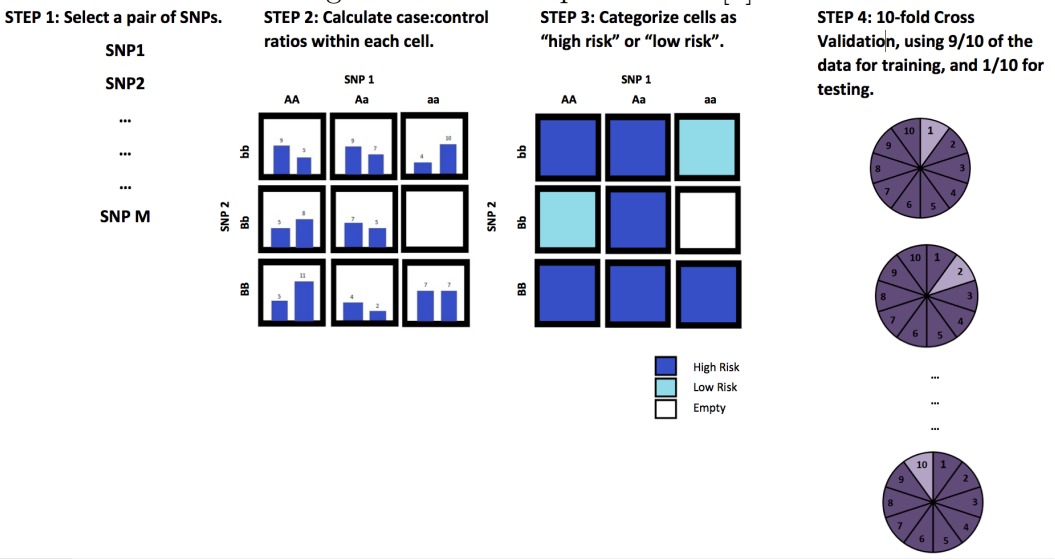
The MDR method can be implemented in four general steps. Note that a balanced case-control study is typical for Ritchie's original MDR approach. In step 1, a set of genotypic (SNPs) and/or categorical environmental factors are selected from all possible factors. In step 2, an interaction among these factors is selected and represented by cells or multifactor classes which represent genotypic combinations within that interaction. For example, a two-way interaction for SNPs with three states can be represented by a 3×3 table, where each cell in the table is a genotypic combination or multifactor class. For a case-control study, the ratio of the number of cases to the number of controls is calculated within each multifactor class. In a discordant-sib-pair study, the ratio of the number of affected sibs to the number of unaffected sibs is calculated within each multifactor class. In step 3, each of the multifactor cells is identified as "high risk" or "low risk". We define a "high risk" cell as any multifactor cell in which the cases:controls ratio is greater than or equal to some pre-defined threshold. Cells in which cases:controls is less than the threshold are defined as "low risk". In balanced case-control studies, a commonly used threshold is 1 [2]. Extensions to the original MDR method have focused on changing this threshold to accommodate unbalanced case control studies, in order to avoid assigning a "high risk" classification simply due to dominance of cases over controls in the data [6]. Then the "high risk" cells are pooled into a group, and the "low risk" cells are pooled into another group. The result is a one-dimensional model, that is, a single binary variable which takes a value of 1 if a subject is "high risk" and 0 otherwise.

In step 4, 10-fold cross-validation is used to estimate the prediction error of each one-dimensional model. In this process, the data are divided into ten equal subsets and MDR is performed on nine of the subsets (the training data), and used to make predictions about the remaining subset of subjects (validation data). Prediction error is estimated by calculating the proportion of subjects for which an incorrect prediction is made. The process is repeated ten times using one of the subsets for validation data, and the prediction error is averaged over the ten repetitions [2].

The four steps of the MDR method are repeated for each possible two-factor combination. Among all possible two-factor combinations, the model which maximizes the cases:controls ratio for the "high risk" group is selected as a candidate model. Then, this process is repeated in consideration of three-way gene-to-gene interactions, and a candidate three-way model is chosen. Eventually, a candidate model is selected for each of two to n way gene-to-gene interactions (resulting in n-1 candidate models). Among the set of candidate models, the one which minimizes the prediction error averaged over the ten repetitions of cross validation is selected as the best overall model. Finally, the optimal model is evaluated for statistical significance using 1000 count permutation testing. In this process, the distribution of the cross validation consistencies (under the null hypothesis that the gene-to-gene interaction is not associated with disease status) is estimated empirically by shuffling the response variable (permuting), and performing 10-fold cross validation. For each permutation, consistency of the selected model across cross validation sets (the number of times the selected model is identified as optimal in each possible 9/10 of the subjects) is calculated. The average cross validation consistency from the original data set is compared to the distribution of average consistencies found empirically from 1000 permutations, under the null hypothesis of no association. A model is said to be statistically significant if the relative frequency of average consistencies obtained by permutation which are

greater than the cross validation consistency obtained from the original data is less than .05 [2]. The MDR method is illustrated in figure 1.1, which has been adapted from a similar figure produced by Ritchie et al. [2].

Figure 1.1: The steps of MDR [2].



Software for implementing the MDR algorithm has been developed by Hahn et al. [5], Bush et. al. [7], Winham and Motsinger-Reif [8] and Moore [9] and the method has been widely used in the analysis of SNP data relating to common complex diseases. Despite the advantages of MDR in detecting and characterizing gene-to-gene interactions which are associated with disease susceptibility, the method does have some limitations. MDR can be computationally intensive when more than ten SNPs are included in a data set [2]. For example, consider a data set which includes M SNPs. There are $\binom{M}{k}$ possible k-way combinations. For large values of M, the number of combinations which are to be considered can become overwhelming. In some cases, MDR can be combined with a filter preprocess to select the optimal number of SNPs to include in MDR [1]. The original MDR approach is also strictly applicable to phenotypes with binary outcomes (absence or presence of a disease, for example). Many extensions to the MDR method have been investigated in order to make the

method applicable to quantitative outcomes, and to replace "high risk" and "low risk" susceptibility categories with a quantifier. Some of these extensions are discussed in this work.

1.2 Quantitative Traits

An area of interest concerning MDR involves the extension of the method to account for quantitative outcomes (phenotypes). Rather than focusing on discrete, dichotomous outcomes, researchers may be interested in continuous outcomes such as body mass index, survival time, and tumor growth. In these cases, the original MDR method [2] cannot be applied. Several methods have been suggested in order to address this possible extension. One such method is Generalized MDR (GMDR), which is based on the score statistic of generalized linear models [10]. This extension of MDR allows continuous phenotypes by replacing the ratio of cases and controls which was used in MDR with a score-based statistic for a given cell. This statistic is used to classify gene-to-gene interactions (cells) as "high risk" or "low risk" [10]. Another approach to identifying interactions for quantitative traits is Model-based MDR (MB-MDR), which implements parametric regression within the MDR method [11].

Unfortunately, when considering quantitative phenotypes, GMDR and MB-MDR are not computationally efficient. Further, although an R package has been developed to implement Model-based MDR, this package does not consider gene-to-gene interactions with higher dimension than three [12]. To address the possible limitations of these two methods, Gui [12] suggests an extension of the original MDR method called Quantitative MDR (QMDR). The steps involved in implementation of QMDR are analogous to the original method. However, instead of comparing the cases:controls ratio within each multifactor cell to some threshold, we compare the

mean value within each multifactor cell to the overall mean. If the mean value for a genotypic combination is greater than the overall mean, the combination is considered to be in the "high-level" group, and if it is lower than the overall mean, the combination is "low-level". Once again, a new one-dimensional, binary variable is created by pooling subjects within "high-level" multifactor classes into a group and subjects within "low-level" multifactor classes into another.

Recall that in MDR, prediction error was used to evaluate the possible models. A 0-1 prediction error is meaningless for quantitative outcomes, so Gui suggests the use of a T-statistic calculated based on the difference in the "high-level" group mean and the "low-level" group mean as a training score to be maximized when selecting the best k-way interaction model. Cross-validation is performed in the same way, except we calculate a testing score based on assigning subjects in the testing data into "high-level" and "low-level" groups in regard to the classification obtained from the training data. We maximize the testing score to select the optimal model out of all candidate models. The final model is then evaluated for statistical significance with permutation testing. Gui tested the model against the GMDR method, and found that QMDR provided similar results to GMDR with decreased computation time [12].

1.3 Aggregated Multifactor Dimensionality Reduction

Another extension to Ritchie's original method was developed by Dai et al.[1], whose notation we adopt in the sequel. Aggregated-Multifactor Dimensionality Reduction (A-MDR) replaces the dichotomous "high risk" and "low risk" groups designated in the original MDR method with an epistasis enriched risk score. This risk factor aims to quantify disease susceptibility. The A-MDR method is described here. Assume we have SNPs which occur in one of three states: 0-homozygous, 1-heterozygous,

2-homozygous variant. For example, consider a SNP which contains one allele, with three possible states: AA, Aa, and aa. Then 0 is assigned to AA, 1 to Aa, and 2 to aa. In the first step of A-MDR, the predisposing risk factor is calculated as follows: Suppose we are interested in k-way gene-to-gene interactions among M different SNP factors. For one such k-way interaction, there are 3^k possible combinations, because $SNP_{(1)}$ has three possible states $(0, 1, 2)$, $SNP_{(2)}$ has three possible states, and so forth. There are also $\binom{M}{k}$ k-way gene-to-gene interactions among M SNP factors.

Let $j = 1, 2, ..., 3^k$ represent all possible genotypic combinations (or multifactor cells) within a k-way gene-to-gene interaction. Let $i = 1, 2, ..., \binom{M}{k}$ represent all possible k-way interactions among M SNPs. Let $X_{i,j}$ be the number of cases in the jth genotypic combination of the ith k-way gene-to-gene interaction. Similarly, let $Y_{i,j}$ be the number of control subjects in the jth genotypic combination of the ith k-way gene-to-gene interaction. These counts are used to create a threshold. A genotypic combination is classified as "highly susceptible" if the disease risk associated with that combination is greater than the threshold defined as:

$$p_0 = \frac{\sum_{j=1}^{3^k} X_{i,j}}{\sum_{j=1}^{3^k} X_{i,j} + \sum_{j=1}^{3^k} Y_{i,j}} \tag{1.1}$$

Note that this threshold itself is an extension of the original MDR method, because it does not assume a balanced case-control study. Gene-to-gene interactions are then classified into "high risk" and "low risk" groups as shown in the contingency table (table 1.1), which is reproduced from Dai et al. [1]. Note that $I[*]$ represents an indicator function which takes a value of 1 if $[*]$ is true and 0 otherwise. This table accounts for N subjects in a study. Dai suggests three different measures calculated

Table 1.1: Predisposing Risk Table, reproduced from Dai et al. [1]

	Case	Control	Total
High Risk	$n_{11} = \sum_{j=1}^{3^k} X_{ij} I[\frac{X_{ij}}{X_{ij}+Y_{ij}} > p_0]$	$n_{12} = \sum_{j=1}^{3^k} Y_{ij} I[\frac{X_{ij}}{X_{ij}+Y_{ij}} > p_0]$	n_{1+}
Low Risk	$n_{21} = \sum_{j=1}^{3^k} X_{ij} I[\frac{X_{ij}}{X_{ij}+Y_{ij}} \leq p_0]$	$n_{22} = \sum_{j=1}^{3^k} Y_{ij} I[\frac{X_{ij}}{X_{ij}+Y_{ij}} \leq p_0]$	n_{2+}
Total	n_{+1}	n_{+2}	N

from the contingency table (table 1.1), which can be used to evaluate the statistical significance of a k-way gene-to-gene interaction. These include the predisposing odds ratio (pOR), the predisposing chi-square ($pChi$) and the predisposing relative risk (pRR). To date, using inference based on normal or χ^2 asymptotic distributions with one degree of freedom (as is typically appropriate for contingency table analysis) has not been justified in this case. As a result, permutation testing has been be used to assess significance [1]. Formulas for the three measures are as follows:

$$pOR_i = \frac{n_{11}n_{22}/(n_{12}n_{21})}{F_0^{-1}(F(n_{11}n_{22}/(n_{12}n_{21})))} \tag{1.2}$$

$$pRR_i = \frac{\frac{n_{11}/(n_{11}+n_{12})}{n_{21}/(n_{21}+n_{22})}}{F_0^{-1}F(\frac{n_{11}/(n_{11}+n_{12})}{n_{21}/(n_{21}+n_{22})})} \tag{1.3}$$

$$pChi_i = \frac{\sum_{s=1}^{2}\sum_{t=1}^{2}\frac{(n_{st}-e_{st})^2}{e_{st}}}{F_0^{-1}F(\sum_{s=1}^{2}\sum_{t=1}^{2}\frac{(n_{st}-e_{st})^2}{e_{st}})}. \tag{1.4}$$

Note that $pChi$ is the based on the traditional calculation of a χ^2 test statistic, where n_{st} is the "observed" value in a contingency table cell, and e_{st} is the "expected" value

for the cell under the null hypothesis of no association. The expected value for each cell is calculated by $e_{st} = \frac{n_{s+}n_{+t}}{N}$ [1].

Regardless of the chosen measure (pOR, pRR or $pChi$), it is necessary to calculate $F(x)$ and $F_0^{-1}(x)$, where x is the numerator of the chosen measure provided in (1.2), (1.3) and (1.4). For example, if the measure of interest is pOR, then $x = n_{11}n_{22}/(n_{12}n_{21})$. Note that the denominators in each of these measure calculations are necessary only to facilitate the interpretation of 95% confidence intervals. That is, exempting the occurrence of a type II error, confidence intervals for each statistic (pOR, pRR or $pChi$) will contain 1 when H_0 (no interaction is present) is true. We can calculate the cumulative distribution of x under the alternative hypothesis that a gene-to-gene interaction is present in nature, and call it $F(x)$. Similarly, let $F_0(x)$ denote the cumulative distribution function under the null hypothesis that the interaction is not present, and F_0^{-1} be the inverse of $F_0(x)$. We estimate these cumulative distributions empirically. In order to estimate $F_0(x)$, the phenotypes are permuted among individuals in the data set while the SNPs (factors) are maintained for each individual. For each of these permutations, the chosen statistic (the numerator from (1.2), (1.3) or (1.4)) is computed. This process is repeated, say 1000 times, allowing a cumulative distribution function of the statistic under the null hypothesis to be estimated empirically. The logic here is that if no relationship between the predisposing risk and disease status is present, then pOR, pRR, or $pChi$ obtained from the original data will be similar to those obtained by permuting the response variable. $F(x)$ is estimated empirically through a process known as jackknife resampling. Subsets of the data are selected such that $80 - 90\%$ of the data are used to calculate the numerator from (1.2), (1.3) or (1.4), and again, a cumulative distribution function for a statistic is estimated from many, say 1000 repetitions of this process. We use $F(x)$ and $F_0^{-1}(x)$ in the calculation of pOR, pRR and pChi. $F_0(x)$ is used to

calculate a p-value for the statistic (the numerator from (1.2), (1.3) or (1.4)) obtained from the original data (call it z), by calculating the percentage of statistics from the permutations that are greater than z [1].

For subject n (noting that n may be a new subject for which a prediction is to be made or $n \in \{1, 2, ..., N\}$), an aggregated k-way epistasis enriched risk score is calculated as follows:

$$R(k,n) = \sum_{i=1}^{\binom{M}{k}} \left[I[pval_i < \hat{\alpha}] \sum_{j=1}^{3^k} I[n \in C_{ij}] I\left[\frac{X_{ij}}{X_{ij} + Y_{ij}} > p_0 \right] \right] \tag{1.5}$$

where $\hat{\alpha} = argmax[AUC|\alpha]$ for $0 \leq \alpha \leq .05$, C_{ij} represents the cell corresponding to genotypic combination j within gene-to-gene interaction i, and $pval_i$ is the p-value obtained from the permutation process performed on the chosen measure. Notice that the aggregated risk score only includes the significant k-way gene-to-gene interactions and those which are assigned to the "high risk" group. The area under a receiver operating characteristic curve (AUC) is maximized over possible values of $\alpha \in [0, .05]$ to maximize the risk score's ability to predict disease susceptibility [1].

1.4 Motivation and Outline of this Work

Current methodology allows for the aggregation of cumulative effects of multiple gene-to-gene interactions regarding dichotomous phenotypes. Progress has been made in the identification of significant gene-to-gene interactions under the consideration of quantitative phenotypes. In this work, we will examine a new approach to predicting quantitative phenotypes in which the effects from multiple significant interactions are combined to form a continuous, aggregated score to be used in a model for prediction.

In chapter 2, we will propose a new method, Aggregated Quantitative Multifactor Dimensionality Reduction (AQMDR), in which the dichotomous categorization from Quantitative Multifactor Dimensionality Reduction (QMDR) is replaced by a continuous score developed by accumulating effects from multiple (or all possible) k-way gene-to-gene interactions. We propose three distinct aggregated scores, which dictate which interactions are included in the model, and the weight assigned to each of these interactions. We apply these aggregated scores to simulated data, and evaluate the method in the realm of two-way interactions.

In chapter 3, we will extend the AQMDR method to three-way interactions, and look at accumulating effects from two-way and three-way interactions.

In chapter 4, we will explore quadratic models as opposed to simple linear models in the context of AQMDR.

In chapter 5, we will provide theoretical support for the AQMDR method.

In chapter 6, we will apply the methodology of this work to a real-world data set. We will seek to explore interactions between the epsilon 4 allele of the $APOE$ gene and several other genetic factors that are known to be associated with AD. In particular, we want to examine how these interactions influence the quantitative factors CSF Aβ and CSF tau.

Chapter 2

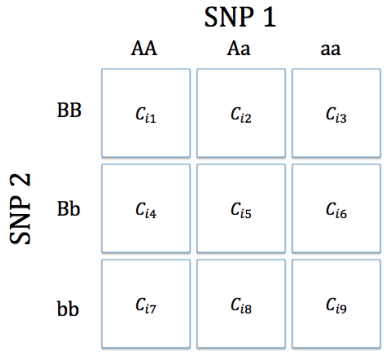
AQMDR and Considerations for Two-way Interactions

2.1 Aggregated Score for Quantitative Phenotypes

In the present work, an Aggregated Quantitative Multifactor Dimensionality Reduction (AQMDR) method is proposed, in which gene-to-gene interactions are considered exhaustively to generate aggregated scores. This score improves on Quantitative Multifactor Dimensionality Reduction [12] and replaces the dichotomous predisposing risk factor from the original QMDR. We introduce the new methodology, and propose three distinct aggregated scores. We evaluate the method through simulation study with focus on two-way gene-to-gene interactions.

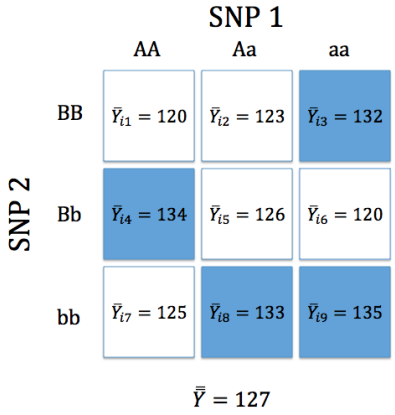
In the following sequence, we adopt the notation of Dai et al. [1] where appropriate. For concreteness, we present the proposed AQMDR method with respect to SNPs with three common states (0-homozygous reference, 1- heterozygous, 2 - homozygous variant). The method can be extended to incorporate interactions among other explanatory variables (e.g. environmental factors) which are categorical in nature. Suppose we are interested in k-way gene-to-gene interactions among M SNPs. There are 3^k possible genotypic combinations (recalling that each SNP exhibits three possible states). We denote these combinations as C_{ij}, where the $j = 1, 2, ..., 3^k$ index different genotypic combinations within one k-way gene-to-gene interaction and $i = 1, 2, ..., \binom{M}{k}$ represents the possible k-way interactions among M SNPs. For example, for a 2-way gene-to-gene interaction, we can think of the C_{ij} as cells in a 3×3 table. Figure 2.1 shows a two-way interaction between SNP 1 and SNP 2, each of which contain one allele with three possible states (AA, Aa, aa, BB, Bb, bb). Each of the nine cells represents a C_{ij} (the j^{th} genotypic combination within the i^{th} k-way gene-to-gene interaction).

Figure 2.1: A two-way interaction between SNP 1 and SNP 2



Departing from the notation of Dai et al. [1], suppose we consider N subjects and a continuous phenotype Y. Let $\bar{Y}_{ij}$ be the mean value of Y among subjects exhibiting genotypic combination j of the i^{th} k-way gene-to-gene interaction. Let $\bar{\bar{Y}}$ denote the overall mean of Y among the N subjects. Genotypic combinations are classified into groups based on $\bar{Y}_{ij}$. Combinations in which $\bar{Y}_{ij} > \bar{\bar{Y}}$ are classified as "high" and receive a value of 1 and all others are classified as "low" and receive a value of 0. In figure 2.2, the mean of Y among subjects in each cell is given ($\bar{Y}_{ij}$), as well as the grand mean, $\bar{\bar{Y}}$. Shaded cells represent those receiving a classification of "high" (1) and white cells represent those receiving a classification of "low" (0).

Figure 2.2: Cell classification of a two-way interaction.



After all genotypic combinations in the i^{th} k-way gene-to-gene interaction are identified as "high" or "low" combinations, the subjects within the "high" group are pooled together, and the "low" group subjects are pooled as well. Then these groups are used to calculate the following test statistic for the i^{th} k-way interaction:

$$t^* = \frac{\bar{Y}_{high} - \bar{Y}_{low}}{\sqrt{\frac{sd^2_{high}}{n_{high}} + \frac{sd^2_{low}}{n_{low}}}} \tag{2.1}$$

where sd_{high} is the standard deviation among phenotypes of the n_{high} individuals classified as "high", sd_{low} is the standard deviation among phenotypes of the n_{low} individuals classified as "low". Note that this test statistic was inspired by the QMDR method [12].

As the parametric distribution of this test statistic is unknown, permutation testing is used to evaluate the significance of the i^{th} k-way interaction. For permutation testing, we shuffle the Y outcomes while keeping genotypes fixed. For each permutation of the phenotype, "high" and "low" classifications are reassigned, and the t^* test statistic is recalculated. After many repetitions, say 1000, we calculate the relative frequency of the t^*'s calculated from permutation which are greater than or equal to the t^* calculated from the original data. This relative frequency is then used as a p-value.

This process is repeated for all $\binom{M}{k}$ k-way gene-to-gene interactions resulting in a one-dimensional binary variable identifying whether subjects are in the "high" or "low" group and a p-value for each interaction. These binary variables and p-values are then used to calculate an aggregated score for each subject. In this work, we define and compare three possible aggregated scores.

2.2 Aggregated Score with a Convex Weighting Function

The first aggregated score is characterized by a convex weighting function, designed to assign weights to gene-to-gene interactions based on the p-values obtained through permutation testing. With this weighting function, interactions with low p-values are assigned high weights, and those with high p-values receive lower weights. For interpretability and proper comparison among individuals, the risk score is presented on a scale between 0 and 1, where values close to 1 are assigned to individuals most at risk for elevated values of the phenotype. This aggregated score, referred to as the Convex Weighting Aggregated Score (CWAS) in this work, for individual n, where k-way interactions are considered is calculated as follows:

$$CWAS(k,n) = \frac{\sum_{i=1}^{\binom{M}{k}} (1 - pval_i^{1/2}) \sum_{j=1}^{3^k} I[n \in C_{ij}] I[\bar{Y}_{ij} > \bar{\bar{Y}}]}{\sum_{i=1}^{\binom{M}{k}} (1 - pval_i^{1/2})} \tag{2.2}$$

where $pval_i$ is the p-value for gene-to-gene interaction i obtained from permutation testing. Note that this aggregated score (as well as the other two aggregated scores to be defined in this work) may be calculated for an individual in the original data set, or for a new individual.

2.3 Aggregated Score with an Arbitrary Cutoff

The second proposed aggregated score is inspired by the score proposed for binary phenotypes by Dai et al. [1]. This score is based on an arbitrary cutoff for the Monte Carlo p-values for the gene-to-gene interactions obtained from permutation testing. Interactions with p-values less than this cutoff value are included in the aggregated score, and all others are omitted. Choosing the arbitrary cutoff is left to

the researcher, but cutoffs of .05 and .20 will be considered later in this work. This aggregated score will be referred to as the Arbitrary Cutoff Aggregated Score (ACAS) in this work. Let c be the cutoff chosen for inclusion in the aggregated score. The Arbitrary cutoff score is calculated as follows:

$$ACAS(k,n) = \frac{\sum_{i=1}^{\binom{M}{k}} I[pval_i < c] \sum_{j=1}^{3^k} I[n \in C_{ij}] I[\bar{Y}_{ij} > \bar{\bar{Y}}]}{\sum_{i=1}^{\binom{M}{k}} I[pval_i < c]} \tag{2.3}$$

Should an investigator encounter a situation in which none of the Monte Carlo p-values for the $\binom{M}{k}$ interactions are less than c (resulting in an undefined aggregated score), a larger value of c may be considered.

2.4 Hybrid Aggregated Score

The third and final proposed aggregated score is a hybrid of the first two aggregated scores. A weight of 1 is assigned to interactions which produced p-values less than a cutoff (again, call this cutoff value c), and the same convex weight function discussed in section 2.2 is applied to those interactions with p-values greater than the cutoff. In this work, we will refer to this aggregated score as the Hybrid Aggregated Score (HAS). HAS can be calculated as follows:

$$HAS(k,n) = \frac{A+B}{\binom{M}{k}} \tag{2.4}$$

where

$$A = \sum_{i=1}^{\binom{M}{k}} I[pval_i < c] \sum_{j=1}^{3^k} I[n \in C_{ij}] I[\bar{Y}_{ij} > \bar{\bar{Y}}] \tag{2.5}$$

and

$$B = \sum_{i=1}^{\binom{M}{k}} I[pval_i \geq c](1 - pval_i^{1/2}) \sum_{j=1}^{3^k} I[n \in C_{ij}] I[\bar{Y}_{ij} > \bar{\bar{Y}}] \tag{2.6}$$

2.5 The Aggregated Score as a Predictor

Recall that the original QMDR method results in a single, binary variable to be used as a predictor of phenotype values. This binary variable corresponds to the "high" or "low" categorization produced by the gene-to-gene interaction deemed to be optimal [12]. By calculating an aggregated score with AQMDR, we replace the dichotomous predictor with a continuous predictive variable. The aggregated score allows us to incorporate all k-way gene-to-gene interactions, or some subset of the interactions (significant interactions) to more adequately predict values of a quantitative phenotype. The potential advantage of QMDR over our method is that QMDR explicitly identifies which particular gene-to-gene interaction may be driving changes in phenotypes among individuals. However, if accurate prediction of a phenotype is the goal, AQMDR may provide a better predictor for linear regression with the phenotype. AQMDR may be most useful in situations where more than one gene-to-gene interaction influences the phenotype. The aggregated score may be extended to incorporate environmental factors, and the aggregated scores from multiple orders of interaction may be combined (for example, the accumulation of two-way and three-way gene-to-gene interactions). The latter idea will be explored in a later chapter of this work.

2.6 Empirical Assessment for Two-way Gene-to-gene Interactions

An extensive simulation study was performed to assess the performance of the AQMDR method in detection and making predictions based upon two-way interactions. The study is a factorial design in which we examined variations of the number of SNPs to be considered and the number of interactions present in nature. Let M be the number of SNPs under consideration, and x be the number of two-way interactions present. SNPs were generated under the assumption of Hardy-Weinberg equilibrium [13]. For example, consider a SNP, call it SNPA, with two alleles, each of which take states of A or a. We let the probability that an individual carries allele A on a single chromosome be $p = .5$, and the probability that a person carries allele a on a single chromosome be $q = .5$. Then the SNP states AA and aa each have probability $p^2 = .25$ and $q^2 = .25$ of occurring. SNP state Aa occurs with probability $2pq = .5$. Phenotypes for individuals were randomly generated by a $N(120, \sigma^2)$ distribution, where the mean of this normal distribution was increased by the presence of an interaction. For example, suppose that the two-way interaction of SNP1 and SNP2 was included in simulation. Then the mean of the random normal distribution used to generate the phenotype was increased by 30 in the event that an individual carried a genotypic combination such as (SNP1 = 2)(SNP2 = 2). This example simulation model is displayed in equation 2.7.

In the factorial simulation study, we varied the standard deviation, σ of the normal distribution used to generate phenotypes as well. See table 2.1 for the particular variations considered in this factorial simulation study. Each training and testing data set contained phenotypes and SNP states for 1000 simulated subjects. If one interaction was present, SNP1 $\times$ SNP2 was selected as the present interaction. If three interactions were present in simulation, SNP1 $\times$ SNP2, SNP1 $\times$ SNP3, and SNP1 $\times$ SNP4 were the interactions selected. If six interactions were present in simulation,

SNP1 × SNP2, SNP1 × SNP3, SNP1 × SNP4, SNP2 × SNP3, SNP2 × SNP4, and SNP3 × SNP4 were included. For each interaction, the homozygous variant (SNP=2) was used to define an interaction leading to an elevated phenotype, and all interactions were generated using the same magnitude of elevation. That is, the mean of the random normal was increased by 30 in the presence of each interaction.

$$Y_i = 120 + 30(I[SNP1_i = 2]I[SNP2_i = 2]) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \tag{2.7}$$

Table 2.1: The values of variants considered in the two-way interaction factorial simulation study.

M	x	σ
4	1	10
10	3	15
	6	20

The AQMDR method aims to produce more accurate predictions for quantitative phenotypes than current methodology. Thus, the focus of this simulation study is on quality of prediction. For each combination of variants (as shown in table 2.1) 100 independent training and testing data sets were generated. We applied the AQMDR method to each set of training data and subsequently made predictions for the testing data. Each of the proposed aggregated scores (CWAS, ACAS and HAS) were implemented in the AQMDR method and compared. ACAS and HAS were varied by using $c = .05$ and $c = .20$. For further comparison, we considered three alternative methods of prediction. The first method was the original QMDR method developed by Gui et al. [12] and discussed extensively in section 1.2 of this work. The second method involved using the overall mean for the phenotype in the training data as the prediction for all of the phenotype values in the testing data. For simplicity, we refer to this prediction method as the training mean (TM) method. The final alternative

prediction method was based on knowledge of the truth used for simulation. For example, if it was known that an interaction between SNP1 and SNP2 was present in the simulation, the presence of (SNP1=2)(SNP2=2) resulted in an increased mean for the normal distribution used to generate the phenotype. Thus, in this method (referred to as the oracle method), a binary variable in which individuals carrying a genotypic combination ((SNP1=2)(SNP2=2) in the previous example) receive a value of 1 is used as a linear predictor of the phenotype. Multiple present interactions result in multiple linear predictors. As this method is based on knowledge of the truth, we included it in this simulation study as a best-case-scenario (in the absence of dimensionality reduction) for comparison. After predictions for the phenotypes in the testing data were made based on the training data, mean squared prediction error (MSPE) was calculated for each testing set. Equation 2.8 shows the MSPE for n observations (y_i's) and the corresponding predictions ($\hat{y}_i$'s).

$$MSPE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{2.8}$$

The results of the factorial simulation study are displayed in table 2.2. Each value in the table is the average MSPE over 100 independent repetitions of generating training and testing data sets. The first three columns of the table indicate the variants of M, x and σ used in the simulation scenario for each row. Columns four through nine correspond to implementations of AQMDR using the indicated proposed aggregated scores. A value under the header such as .05 and .20 indicates the chosen cutoff value c in ACAS and HAS. The final three columns correspond to the alternative prediction methods, QMDR, the oracle method, and the training mean method, respectively.

As shown in table 2.2, when more than one gene-to-gene interaction is present, the average MSPEs are consistently lower for the AQMDR implementations when compared to the traditional QMDR method. In these cases, the AQMDR method demonstrated more adequate predictions. This is apparent, regardless of which of the aggregated scores were implemented in the AQMDR method. However, the lowest MSPEs for cases where $x > 1$ occurred when applying the Arbitrary Cutoff aggregated score with $c = .05$. These results provide evidence that AQMDR may be providing more predictive ability in scenarios where multiple gene-to-gene interactions affect the phenotype ($x > 1$).

Table 2.2: Average MSPE over 100 independent testing data sets for two-way interaction simulation study.

M	x	σ	CWAS	ACAS (.05)	ACAS (.20)	HAS (.05)	HAS (.20)	QMDR	Oracle	TM
4	1	10	127.19	126.59	127.27	128.30	128.09	107.78	100.55	153.96
4	1	15	255.00	254.51	254.79	255.92	255.80	247.02	225.65	277.17
4	1	20	437.61	437.03	437.40	439.31	438.87	431.89	403.66	456.85
10	1	10	133.91	131.25	132.91	131.82	133.47	109.41	100.61	153.67
10	1	15	262.52	258.09	260.87	259.29	261.85	247.03	225.28	277.70
10	1	20	444.32	438.73	440.95	440.42	442.51	435.32	402.44	454.40
4	3	10	203.78	203.78	203.78	203.78	203.78	225.70	99.57	320.45
4	3	15	331.22	331.26	331.26	331.26	331.26	354.72	225.45	448.18
4	3	20	505.97	506.35	506.20	506.36	506.20	525.42	402.17	620.46
10	3	10	223.39	220.25	223.18	220.87	223.68	226.41	100.41	320.52
10	3	15	349.97	345.91	350.19	346.67	350.87	352.16	224.93	446.17
10	3	20	528.69	523.60	528.67	524.69	529.72	527.73	400.41	623.86
4	6	10	423.23	423.23	423.23	423.23	423.23	561.48	101.14	675.00
4	6	15	554.04	554.04	554.04	554.04	554.04	693.65	227.01	805.64
4	6	20	719.00	719.00	719.00	719.00	719.00	856.36	403.00	970.46
10	6	10	338.23	329.28	336.23	330.49	337.40	561.73	101.42	674.75
10	6	15	458.38	450.28	456.92	451.36	457.97	683.20	225.75	800.40
10	6	20	631.19	622.48	630.08	623.71	631.07	844.66	402.48	954.56

To further explore the distinctions between methods in the simulation study, a single factor, within-subjects ANOVA was performed for each simulation scenario, followed by post hoc multiple pairwise comparisons among the aggregated scores and

alternative methods. For all simulation scenarios, the ANOVA yielded significance at $\alpha = .05$ level, suggesting that there are differences among the average MSPEs for the prediction methods. The prediction methods were ranked according to the results from multiple pairwise comparisons. The rankings from multiple pairwise comparisons with a Bonferroni correction based on the total number of comparisons ($\alpha = \frac{.05}{28}$) are displayed in table 2.3. For each simulation scenario (a row in the table), prediction methods were assigned rankings a, b , c, etc. where a is the ranking assigned to the method with the lowest average MSPE. Methods sharing a letter ranking were not determined to be statistically significant in a pairwise comparison.

Table 2.3: Within-subjects ANOVA with a Bonferroni correction, $\alpha = \frac{.05}{28}$ for two-way interaction simulation study.

M	x	σ	CWAS	ACAS (.05)	ACAS (.20)	HAS (.05)	HAS (.20)	QMDR	Oracle	TM
4	1	10	c	c	c	c	c	b	a	d
4	1	15	c	c	c	c	c	b	a	d
4	1	20	b	b	b	b	b	b	a	d
10	1	10	d	c	c	c	d	b	a	e
10	1	15	d	c	c	c	d	b	a	e
10	1	20	d	c	c	cd	d	b	a	e
4	3	10	b	b	b	b	b	c	a	d
4	3	15	b	b	b	b	b	c	a	d
4	3	20	b	b	b	b	b	c	a	d
10	3	10	c	b	b	b	c	d	a	e
10	3	15	c	b	b	b	c	d	a	e
10	3	20	c	b	b	b	c	c	a	d
4	6	10	b	b	b	b	b	c	a	d
4	6	15	b	b	b	b	b	c	a	d
4	6	20	b	b	b	b	b	c	a	d
10	6	10	b	b	b	b	b	c	a	d
10	6	15	b	b	b	b	b	c	a	d
10	6	20	b	b	b	b	b	c	a	d

First, notice that the AQMDR method performs significantly better than QMDR in all scenarios with $x > 1$. As expected, QMDR excels in those cases with a single interaction present. In scenarios with only four SNPs, the five proposed aggregated

scores within AQMDR are not significantly different. When $M = 10$, we see that the five aggregated scores begin to differ, and in all of these cases the ACAS with $c = .05$ always yields lower average MSPEs.

Table 2.4 displays the average testing R^2 values for each prediction method within each simulation scenario. When comparing the average R^2 values of QMDR with the AQMDR aggregated scores, we clearly see that AQMDR outperforms QMDR in all scenarios with $x > 1$. For example, the advantage of AQMDR is especially apparent when $M = 10$, $x = 6$ and $\sigma = 20$. In this case, QMDR produced an R^2 of approximately 11.51%, while AQMDR (with ACAS, $c = .05$) produced an R^2 of approximately 34.79%.

Table 2.4: Average testing R^2 values (as percentages) for two-way interaction simulation study.

M	x	σ	CWAS	ACAS (.05)	ACAS (.20)	HAS (.05)	HAS (.20)	QMDR	Oracle
4	1	10	17.39%	17.78%	17.34%	16.67%	16.80%	29.99%	34.69%
4	1	15	8.00%	8.18%	8.07%	7.67%	7.71%	10.88%	18.59%
4	1	20	4.21%	4.34%	4.26%	3.84%	3.94%	5.46%	11.64%
10	1	10	12.86%	14.59%	13.51%	14.22%	13.15%	28.80%	34.53%
10	1	15	5.47%	7.06%	6.06%	6.63%	5.71%	11.04%	18.88%
10	1	20	2.22%	3.45%	2.96%	3.08%	2.62%	4.20%	11.43%
4	3	10	36.41%	36.41%	36.41%	36.41%	36.41%	29.57%	68.93%
4	3	15	26.10%	26.09%	26.09%	26.09%	26.09%	20.85%	49.70%
4	3	20	18.45%	18.39%	18.42%	18.39%	18.42%	15.32%	35.18%
10	3	10	30.30%	31.28%	30.37%	31.09%	30.21%	29.36%	68.67%
10	3	15	21.56%	22.47%	21.51%	22.30%	21.36%	21.07%	49.59%
10	3	20	15.26%	16.07%	15.26%	15.90%	15.09%	15.41%	35.82%
4	6	10	37.30%	37.30%	37.30%	37.30%	37.30%	16.82%	85.02%
4	6	15	31.23%	31.23%	31.23%	31.23%	31.23%	13.90%	71.82%
4	6	20	25.91%	25.91%	25.91%	25.91%	25.91%	11.76%	58.47%
10	6	10	49.87%	51.20%	50.17%	51.02%	50.00%	16.75%	84.97%
10	6	15	42.73%	43.74%	42.91%	43.61%	42.78%	14.64%	71.80%
10	6	20	33.88%	34.79%	33.99%	34.66%	33.89%	11.51%	57.84%

In table 2.5, we see the ratios of each prediction method average testing R^2 value to that of the oracle method. For the AQMDR aggregated scores, these ratios tend to be larger for scenarios with multiple interactions present. The opposite is true of QMDR, where increases in x produce smaller ratios.

An interesting finding in the simulation study can be seen in those simulation scenarios with six interactions present in nature ($x = 6$). Focusing on the AQMDR implementations, one can see that average MSPEs for scenarios with four SNPs and six significant interactions present were consistently larger than the average MSPEs for scenarios with ten SNPs and six significant interactions (see table 2.2). In other words, the more complicated simulation scenario (ten SNPs) produced better predictions than the less complicated scenario (four SNPs). This phenomenon may be related to overestimation of an individual's phenotype value when he/she carries multiple (or all) of the gene-to-gene interactions present in nature.

Table 2.5: Method:Oracle R^2 Ratio for two-way interaction simulation study.

M	x	σ	CWAS	ACAS (.05)	ACAS (.20)	HAS (.05)	HAS (.20)	QMDR
4	1	10	0.50	0.51	0.50	0.48	0.48	0.86
4	1	15	0.43	0.44	0.43	0.41	0.41	0.59
4	1	20	0.36	0.37	0.37	0.33	0.34	0.47
10	1	10	0.37	0.42	0.39	0.41	0.38	0.83
10	1	15	0.29	0.37	0.32	0.35	0.30	0.59
10	1	20	0.19	0.30	0.26	0.27	0.23	0.37
4	3	10	0.53	0.53	0.53	0.53	0.53	0.43
4	3	15	0.53	0.52	0.52	0.52	0.52	0.42
4	3	20	0.52	0.52	0.52	0.52	0.52	0.44
10	3	10	0.44	0.46	0.44	0.45	0.44	0.43
10	3	15	0.43	0.45	0.43	0.45	0.43	0.42
10	3	20	0.43	0.45	0.43	0.44	0.42	0.43
4	6	10	0.44	0.44	0.44	0.44	0.44	0.20
4	6	15	0.43	0.43	0.43	0.43	0.43	0.19
4	6	20	0.44	0.44	0.44	0.44	0.44	0.20
10	6	10	0.59	0.60	0.59	0.60	0.59	0.20
10	6	15	0.60	0.61	0.60	0.61	0.60	0.20
10	6	20	0.59	0.60	0.59	0.60	0.59	0.20

Chapter 3

Higher Order Interactions

3.1 Empirical Assessment for Three-way Gene-to-gene Interactions

In 2001, Ritchie et. al [2] identified a four-way gene-to-gene interaction related to sporadic breast cancer. Based on this and other similar findings, we may be interested in the consideration of higher order interactions. Currently, we turn our attention toward three-way interactions in the context of the same AQMDR method and proposed aggregated scores discussed in chapter 2. Again, we performed a simulation study using a factorial design in which we examined variations of the number of SNPs to be considered (M) and the number of present interactions (x). SNPs were generated in the same manner discussed in chapter 2. Phenotypes for subjects were generated using a $N(120, \sigma^2)$ distribution. The mean of this normal distribution was increased by 30 with the presence of a three-way interaction. For example, suppose that an interaction between SNP1, SNP2 and SNP3 was included in a simulation scenario. In this case, the mean of the random normal distribution used to generate the phenotype was increased in the event that an individual carried a genotypic combination such as (SNP1 = 2)(SNP2 = 2)(SNP3 =2). This example simulation scenario is displayed in equation 3.1. In the factorial simulation study, we varied the standard deviation of the normal distributions used to generate the phenotypes. See table 3.1 for the particular variations considered in this study. Each training and testing data set contained phenotypes and SNP states for 1000 subjects.

$$Y_i = 120 + 30(I[SNP1_i = 2]I[SNP2_i = 2]I[SNP3_i = 2]) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad (3.1)$$

Table 3.1: The values of variants considered in the three-way interaction factorial simulation study.

M	x	σ
4	1	10
10	2	15
		20

For each combination of variants (M, x, σ), 100 independent training and testing data sets were generated. We applied the AQMDR method to each set of training data and subsequently made predictions for the testing data. The methods considered were the same as those described in chapter 2, including AQMDR (ACAS and HAS with cutoffs of $c = .05$ and $c = .20$, and CWAS), QMDR, the training mean method and the oracle method. As we considered only three-way interactions in this simulation study, the QMDR method was limited only to the selection of three-way interaction models. After predictions for the phenotypes in the testing data were made based on the training data, the mean squared prediction error (MSPE) was calculated for each testing set using equation 2.8. The results of the factorial simulation study are displayed in table 3.2. Each value in the table is the average MSPE over 100 independent testing data sets. If one interaction was present in simulation, SNP1 $\times$ SNP2 $\times$ SNP3 was used. If two interactions were present in simulation, SNP1 $\times$ SNP2 $\times$ SNP3 and SNP1 $\times$ SNP2 $\times$ SNP4 were used. For each interaction, the homozygous variant (SNP=2) was used to define an interaction leading to an elevated phenotype, and all interactions were generated using the same magnitude of elevation. That is, the mean of the random normal was increased by 30 in the presence of each interaction.

As we saw with two-way interactions, QMDR performs well when only one interaction is present ($x = 1$). In scenarios with more than one present interaction, AQMDR yields lower MSPEs than QMDR. This is apparent, regardless of which of the aggre-

gated scores were implemented in AQMDR. These results provide empirical evidence that AQMDR may provide more predictive ability than QMDR in scenarios where multiple three-way gene-to-gene interactions act on the phenotype.

Table 3.2: Average MSPE over 100 independent testing data sets for three-way interaction simulation study.

M	x	σ	CWAS	ACAS (.05)	ACAS (.20)	HAS (.05)	HAS (.20)	QMDR	Oracle	TM
4	1	10	117.26	117.26	117.26	117.26	117.26	105.21	100.17	208.32
4	1	15	250.27	250.32	250.32	250.32	250.32	249.87	225.50	332.00
4	1	20	435.61	436.10	436.09	436.29	436.27	441.55	400.84	508.61
10	1	10	147.45	144.75	146.96	144.89	147.07	106.28	99.91	208.83
10	1	15	271.20	265.96	271.13	266.25	271.34	246.01	223.67	330.63
10	1	20	456.11	447.47	455.27	447.94	455.64	438.04	400.56	507.22
4	2	10	152.08	152.08	152.08	152.08	152.08	201.06	101.15	345.97
4	2	15	278.18	278.18	278.18	278.18	278.18	323.70	223.56	470.72
4	2	20	457.59	457.59	457.59	457.59	457.59	497.20	397.00	644.16
10	2	10	175.67	173.43	174.89	173.50	174.96	200.34	100.27	346.25
10	2	15	303.76	301.19	302.78	301.28	302.87	323.12	225.83	468.38
10	2	20	487.30	483.83	485.95	483.97	486.08	508.05	401.86	646.56

As we observed with the two-way interactions, it appears that the arbitrary cut-off aggregated score (ACAS) with $c = .05$ tends to yield average MSPEs which are a bit more favorable than the other four implementations of AQMDR. To see this distinction a bit more clearly, a single factor, within-subjects ANOVA was performed for each of the twelve simulation scenarios, followed by post hoc multiple pairwise comparisons among the aggregated scores and alternative methods. In each simulation scenario, the ANOVA yielded significance at $\alpha = .05$ level, suggesting that there are differences among the average MSPEs for the methods. Prediction methods were ranked according to results from multiple pairwise comparisons (with a Bonferroni correction, $\alpha = \frac{.05}{28}$). These rankings are displayed in table 3.3. For each simulation scenario (represented by a row in the table), prediction methods were assigned rankings a, b, c, etc. where a is the ranking assigned to the method with the lowest average MSPE. Recall from chapter 2 that methods sharing a letter ranking were not

determined to be statistically significant in a pairwise comparison (that is, there is no significant difference between the methods).

Table 3.3: Within-subjects ANOVA with a Bonferroni correction, $\alpha = \frac{.05}{28}$ for three-way interaction simulation study.

M	x	σ	CWAS	ACAS (.05)	ACAS (.20)	HAS (.05)	HAS (.20)	QMDR	Oracle	TM
4	1	10	c	c	c	c	c	b	a	d
4	1	15	c	c	c	c	c	b	a	d
4	1	20	b	b	b	b	b	c	a	d
10	1	10	d	c	cd	c	d	b	a	e
10	1	15	d	c	d	c	d	b	a	e
10	1	20	d	c	d	c	d	b	a	e
4	2	10	b	b	b	b	b	c	a	d
4	2	15	b	b	b	b	b	c	a	d
4	2	20	b	b	b	b	b	c	a	d
10	2	10	b	b	b	b	b	c	a	d
10	2	15	b	b	b	b	b	c	a	d
10	2	20	b	b	b	b	b	c	a	d

As we saw in scenarios for two-way interactions, AQMDR performs significantly better than QMDR in scenarios in which more than one three-way interaction is present, and thus receives a higher letter ranking in all such scenarios. Recall from chapter 2 that QMDR always performed better than AQMDR in situations with $x = 1$, but for three-way interactions this is not always the case. With a smaller number of SNPs ($M = 4$), a single present interaction and $\sigma = 20$, we see that AQMDR performs better than expected and has a higher ranking than QMDR. However, when $M = 10$, we see that QMDR provides average MSPEs that are consistently lower than those of AQMDR. When $x = 2$, the AQMDR methods are ranked higher than QMDR in all cases, providing evidence that AQMDR may provide more adequate predictions than QMDR when multiple three-way interactions are present (regardless of which aggregated score is used). In nearly all scenarios, the five variations of AQMDR are not significantly different. In fact, differences among the AQMDR aggregated scores

are only present when we have ten SNPs and one present three-way interaction. In these cases, the arbitrary cutoff aggregated score and the hybrid aggregated score (with $c = .05$) perform a bit better than the other three variations of AQMDR, with the convex weighting aggregated score yielding the highest MSPEs.

Table 3.4 displays the average R^2 values (as percentages) for each prediction method within each simulation scenario. As we saw in the simulation study for two-way interactions, these R^2 values decrease as σ increases, and tend to be higher when there are four SNPs under consideration rather than ten. Table 3.5 displays the

Table 3.4: Average R^2 values (as percentages) for three-way interaction simulation study.

M	x	σ	CWAS	ACAS (.05)	ACAS (.20)	HAS (.05)	HAS (.20)	QMDR	Oracle
4	1	10	43.71%	43.71%	43.71%	43.71%	43.71%	49.50%	51.92%
4	1	15	24.62%	24.60%	24.60%	24.60%	24.60%	24.74%	32.08%
4	1	20	14.35%	14.26%	14.26%	14.22%	14.22%	13.18%	21.19%
10	1	10	29.39%	30.69%	29.63%	30.62%	29.57%	49.11%	52.16%
10	1	15	17.97%	19.56%	18.00%	19.47%	17.93%	25.59%	32.35%
10	1	20	10.08%	11.78%	10.24%	11.69%	10.17%	13.64%	21.03%
4	2	10	56.04%	56.04%	56.04%	56.04%	56.04%	41.89%	70.76%
4	2	15	40.90%	40.90%	40.90%	40.90%	40.90%	31.23%	52.51%
4	2	20	28.96%	28.96%	28.96%	28.96%	28.96%	22.81%	38.37%
10	2	10	49.26%	49.91%	49.49%	49.89%	49.47%	42.14%	71.04%
10	2	15	35.15%	35.70%	35.36%	35.68%	35.34%	31.01%	51.78%
10	2	20	24.63%	25.17%	24.84%	25.15%	24.82%	21.42%	37.85%

ratios of each prediction method average R^2 value to that of the oracle method. As we saw with two-way interactions, these ratios for the AQMDR aggregated scores tend to be larger when more than one three-way interaction is present, and the opposite is true for QMDR.

Table 3.5: Method:Oracle R^2 Ratio for three-way interaction simulation study.

M	x	σ	CWAS	ACAS (.05)	ACAS (.20)	HAS (.05)	HAS (.20)	QMDR
4	1	10	0.84	0.84	0.84	0.84	0.84	0.95
4	1	15	0.77	0.77	0.77	0.77	0.77	0.77
4	1	20	0.68	0.67	0.67	0.67	0.67	0.62
10	1	10	0.56	0.59	0.57	0.59	0.57	0.94
10	1	15	0.56	0.60	0.56	0.60	0.55	0.79
10	1	20	0.48	0.56	0.49	0.56	0.48	0.65
4	2	10	0.79	0.79	0.79	0.79	0.79	0.59
4	2	15	0.78	0.78	0.78	0.78	0.78	0.59
4	2	20	0.75	0.75	0.75	0.75	0.75	0.59
10	2	10	0.69	0.70	0.70	0.70	0.70	0.59
10	2	15	0.68	0.69	0.68	0.69	0.68	0.60
10	2	20	0.65	0.67	0.66	0.66	0.66	0.57

3.2 Combining Two-way and Three-way Interactions

Now we turn our attention toward the simultaneous consideration of two-way and three-way interactions. If two-way and three-way interactions are considered using the QMDR technique, candidate interactions are selected for $k = 2$ and $k = 3$. Then 10-fold cross-validation is used to choose a single, optimal interaction among these candidates to be used in the regression model. Rather than choosing between $k = 2$ and $k = 3$, we propose three distinct extensions to AQMDR in which the aggregated scores for two-way and three-way interactions are considered simultaneously. The three proposed approaches are as follows:

1. **Sequential Inclusion** - In this extension, the aggregated scores for two-way and three-way interactions are considered sequentially. That is, the aggregated score for $k = 2$ (two-way interactions) is developed and used as a predictor in a regression model for the phenotype. Then, the residuals from this regression are regressed on the aggregated score for $k = 3$ (three-way interactions).

2. **Simultaneous Inclusion** - In this extension, the aggregated scores for two-way and three-way interactions are considered simultaneously. That is, the

aggregated scores for $k = 2$ and $k = 3$ are both used as predictors in a single regression model for the phenotype.

3. **Principal Components** - In this extension, the first principal component of the aggregated score for $k = 2$ and the aggregated score for $k = 3$ is used as a predictor in a regression model for the phenotype.

To compare these three approaches with the QMDR method, we perform an extensive simulation study. In each simulation, we consider ten SNPs ($M = 10$) and SNP states are generated using Hardy-Weinberg Equilibrium [13]. For example, consider a SNP, call it SNPA, with two alleles, each of which take states of A or a. We let the probability that an individual carries allele A on a single chromosome be $p = .5$, and the probability that a person carries allele a on a single chromosome be $q = .5$. Then the SNP states AA and aa each have probability p^2 and q^2 of occurring. SNP state Aa occurs with probability $2pq$. In previous simulation studies, p was fixed at $p = .5$. For this study, the value of p for each SNP is generated using a continuous uniform distribution on (.5, .9). Phenotypes for individuals were randomly generated by a $N(120, 15^2)$ distribution, where the mean of this normal distribution was increased by 30 with the presence of an interaction. For example, suppose that the two-way interaction of SNP1 and SNP2 was included in simulation, and the three-way interaction of SNP3, SNP4 and SNP5 was included as well. Then the mean of the random normal distribution used to generate the phenotype was increased in the event that an individual carried a genotypic combination such as (SNP1 = 2)(SNP2 = 2), or the genotypic combination (SNP3 = 2)(SNP4 = 2)(SNP5 = 2). This example simulation scenario is defined as

$$\begin{aligned} Y_i = 120 &+ 30(I[SNP1_i = 2]I[SNP2_i = 2]) \\ &+ 30(I[SNP3_i = 2]I[SNP4_i = 2]I[SNP5_i = 2]) + \epsilon_i \end{aligned}$$

where $\epsilon_i \overset{iid}{\sim} N(0, 15^2)$.

In table 3.6, the "1-way" column indicates which of the ten SNPs considered in simulation (SNPA, SNPB,..., SNPJ) are included as a one-way interaction (main effect). For example, if SNPJ is included as a main effect in simulation, then "J" appears in this column. The "2-way" column indicates which two-way interactions were included. For example, if the interaction between SNPA and SNPB was included in simulation, then AB appears in this column. Similarly, the "3-way" column indicates which three-way interactions were included. For example, if the interaction between SNPC, SNPD and SNPE was included in simulation, then "CDE" appears in this column. The magnitude of all one-way, two-way and three-way interactions included in simulation was 30. That is, the presence of an interaction in simulation increased the mean of the random normal distribution by 30. When calculating aggregated scores in the AQMDR method, the arbitrary cutoff aggregated score with a cutoff of $c = .05$ was used. This aggregated score was selected because previous simulations indicated that this particular score may have a slight advantage over the other proposed scores. In table 3.6, the "Seq." column corresponds to the Sequential Inclusion method, the "Sim." column corresponds to the Simultaneous Inclusion approach, and the "P.C." column corresponds to the Principal Components method. Each value in these columns is the average MSPE calculated from 100 independent testing data sets. For discussion that follows, we have assigned a numerical value to each scenario in the first column of table 3.6.

There were some simulation scenarios in which we expected QMDR to perform well, specifically scenarios with a single present interaction (either a three-way interaction or a two-way interaction) such as scenarios 2, 4, and 6. QMDR is well-suited for these situations because the method selects a single "optimal" interaction among all possible two-way and three-way interactions. Looking at the MSPEs for these three

scenarios, we see that QMDR did indeed produce smaller average MSPEs than the three AQMDR implementations. Table 3.7 displays the results of post hoc multiple pairwise comparisons among the prediction methods, performed after single factor, within subjects ANOVA tests which all yielded significance at the $\alpha = .05$ significance level. The prediction methods were ranked according to the multiple pairwise comparisons (with Bonferroni correction, $\alpha = \frac{.05}{10}$), where a ranking of a is assigned to the method with the lowest MSPE. As in previous sections, methods sharing a letter ranking were not determined to be statistically different. If we look at scenario numbers 2, 4, and 6 in table 3.7, we see that the difference between QMDR and the three AQMDR methods is statistically significant (as they don't share a letter ranking).

QMDR also yielded significantly lower MSPEs in scenarios 3 and 5. In both of these scenarios, we have two three-way interactions present and no present two-way interactions. QMDR may have performed well in scenarios 3 and 5 because it selects one of the two present three-way interactions as the optimal interaction. Even though QMDR will ignore the other interaction which is not deemed optimal, it may produce better predictions than AQMDR. In fact, as the AQMDR methods include an aggregated score for two-way interactions when there are actually no two-way interactions present, the two-way aggregated scores may incorporate unnecessary two-way interactions due to confounding with the present three-way interactions. However, we did not see the same results when there were multiple two-way interactions and no three way interactions present. This can be seen in scenario 11, where the AQMDR methods yielded lower MSPEs than QMDR.

In more complicated simulation scenarios in which there were a combination of two-way and three-way interactions (such as scenarios 9, 10, 12, 13, 14, 15), the AQMDR approaches resulted in lower average MSPEs than QMDR. With the exception of scenarios 7 and 8, the differences between the AQMDR methods and QMDR were

statistically significant (see table 3.7). In scenarios 16-30, the first fifteen simulation scenarios were repeated with the addition of a single main effect. With the exception of scenario 16 (in which no two-way or three-way interactions were present), AQMDR provided more adequate predictions than QMDR for all of these scenarios, with statistically significant differences between the AQMDR approaches and QMDR. Note that in scenarios 7 and 8, the Simultaneous Inclusion implementation of AQMDR yielded a lower average MSPE than QMDR, but the Sequential Inclusion and Principal Components implementations both yielded higher average MSPEs than QMDR. Pairwise comparisons in scenarios 7 and 8 determined that Simultaneous Inclusion, Principal Components and QMDR were not significantly different, but Sequential Inclusion performed significantly worse than QMDR and the other two implementations of AQMDR.

In several scenarios (scenarios 3, 6, 11 and 18), there were no significant differences among the three AQMDR approaches for inclusion of two-way and three-way interactions. All four of these scenarios are scenarios in which there were either no two-way interactions present, or no three-way interactions present. In all other scenarios (with the exception of scenarios 1 and 16 in which no two-way or three-way interactions were present), Simultaneous Inclusion and Principal Components are slightly favorable over Sequential Inclusion. Simultaneous Inclusion and Principal Components were not significantly different in any of the simulation scenarios with the exception of scenario 1.

Table 3.8 displays the average R^2 values (as percentages) for the simulation scenarios. One can see that these average R^2 values tend to be higher when a main effect is included in the simulation (scenarios 16-30).

Table 3.6: Average MSPEs for two-way and three-way interactions combined.

No.	1-way	2-way	3-way	Seq.	P.C.	Sim.	QMDR	TM
1	-	-	-	240.02	236.45	239.69	229.69	225.49
2	-	-	ABE	273.58	261.44	261.20	229.65	427.74
3	-	-	ABE & CDF	272.03	260.41	261.85	229.85	429.08
4	-	-	EFG	270.43	260.03	261.68	230.93	429.75
5	-	-	EFG & HIJ	268.50	257.28	258.02	229.04	420.92
6	-	AB	-	276.57	268.59	267.84	230.91	381.19
7	-	AB	ABE	332.05	310.38	301.97	311.47	622.86
8	-	AB	ABE & CDF	327.36	307.80	301.70	304.53	614.35
9	-	AB	EFG	313.17	294.00	292.28	358.02	581.11
10	-	AB	EFG & HIJ	309.87	290.69	288.42	358.65	583.58
11	-	AB & CD	-	314.14	300.61	303.04	355.12	540.66
12	-	AB & CD	ABE	366.72	339.59	339.34	465.36	768.77
13	-	AB & CD	ABE & CDF	367.72	339.94	338.14	481.86	791.92
14	-	AB & CD	EFG	349.83	326.19	327.28	511.12	749.87
15	-	AB & CD	EFG & HIJ	348.32	325.15	326.83	511.89	738.48
16	J	-	-	234.79	228.98	230.71	227.02	431.37
17	J	-	ABE	277.43	262.18	261.74	359.98	637.26
18	J	-	ABE & CDF	278.91	262.92	261.85	366.59	638.97
19	J	-	EFG	275.66	261.64	261.74	363.91	639.04
20	J	-	EFG & HIJ	271.56	257.94	258.51	352.67	630.41
21	J	AB	-	295.75	276.33	271.15	296.91	586.28
22	J	AB	ABE	333.72	308.15	305.91	488.95	832.80
23	J	AB	ABE & CDF	329.56	304.58	302.56	481.66	819.86
24	J	AB	EFG	318.15	295.20	295.72	482.74	786.77
25	J	AB	EFG & HIJ	316.42	293.29	292.92	489.47	784.29
26	J	AB & CD	-	328.99	305.46	305.49	451.17	747.47
27	J	AB & CD	ABE	367.91	340.29	341.44	635.78	981.25
28	J	AB & CD	ABE & CDF	366.24	339.03	340.91	651.35	994.43
29	J	AB & CD	EFG	353.67	328.51	330.26	646.87	964.19
30	J	AB & CD	EFG & HIJ	351.97	326.85	329.04	637.61	947.39

Table 3.7: Within-subjects ANOVA with a Bonferroni correction, $\alpha = \frac{.05}{10}$

No.	1-way	2-way	3-way	Seq.	P.C.	Sim.	QMDR	TM
1	-	-	-	d	c	d	b	a
2	-	-	ABE	c	b	b	a	d
3	-	-	ABE & CDF	c	b	b	a	c
4	-	-	EFG	c	b	b	a	d
5	-	-	EFG & HIJ	c	b	b	a	d
6	-	AB	-	b	b	b	a	c
7	-	AB	ABE	b	a	a	a	c
8	-	AB	ABE & CDF	b	a	a	a	c
9	-	AB	EFG	b	a	a	c	d
10	-	AB	EFG & HIJ	b	a	a	c	d
11	-	AB & CD	-	a	a	a	b	c
12	-	AB & CD	ABE	b	a	a	c	d
13	-	AB & CD	ABE & CDF	b	a	a	c	d
14	-	AB & CD	EFG	b	a	a	c	d
15	-	AB & CD	EFG & HIJ	b	a	a	c	d
16	J	-	-	b	ab	ab	a	c
17	J	-	ABE	b	a	a	c	d
18	J	-	ABE & CDF	b	a	a	c	d
19	J	-	EFG	b	a	a	c	d
20	J	-	EFG & HIJ	b	a	a	c	d
21	J	AB	-	b	a	a	c	d
22	J	AB	ABE	b	a	a	c	d
23	J	AB	ABE & CDF	b	a	a	c	d
24	J	AB	EFG	b	a	a	c	d
25	J	AB	EFG & HIJ	b	a	a	c	d
26	J	AB & CD	-	b	a	a	c	d
27	J	AB & CD	ABE	b	a	a	c	d
28	J	AB & CD	ABE & CDF	b	a	a	c	d
29	J	AB & CD	EFG	b	a	a	c	d
30	J	AB & CD	EFG & HIJ	b	a	a	c	d

Table 3.8: Average R^2 values (as percentages) for two-way and three-way interactions combined

No.	1-way	2-way	3-way	Seq.	P.C.	Sim.	QMDR
1	-	-	-	-6.44%	-4.86%	-6.30%	-1.86%
2	-	-	ABE	36.04%	38.88%	38.93%	46.31%
3	-	-	ABE & CDF	36.60%	39.31%	38.97%	346.43%
4	-	-	EFG	37.07%	39.49%	39.11%	46.26%
5	-	-	EFG & HIJ	36.21%	38.88%	38.70%	45.59%
6	-	AB	-	27.45%	29.54%	29.74%	39.42%
7	-	AB	ABE	46.69%	50.17%	51.52%	49.99%
8	-	AB	ABE & CDF	46.71%	49.90%	50.89%	50.43%
9	-	AB	EFG	46.11%	49.41%	49.70%	38.39%
10	-	AB	EFG & HIJ	46.90%	50.19%	50.58%	38.54%
11	-	AB & CD	-	41.90%	44.40%	43.95%	34.32%
12	-	AB & CD	ABE	52.30%	55.83%	55.86%	39.47%
13	-	AB & CD	ABE & CDF	53.57%	57.07%	57.30%	39.15%
14	-	AB & CD	EFG	53.35%	56.50%	56.36%	31.84%
15	-	AB & CD	EFG & HIJ	52.83%	55.97%	55.74%	30.68%
16	J	-	-	45.57%	46.92%	46.52%	47.37%
17	J	-	ABE	56.47%	58.86%	58.93%	43.51%
18	J	-	ABE & CDF	56.35%	58.85%	59.02%	42.63%
19	J	-	EFG	56.86%	59.06%	59.04%	43.05%
20	J	-	EFG & HIJ	56.92%	59.08%	58.99%	44.06%
21	J	AB	-	49.55%	52.87%	53.75%	49.36%
22	J	AB	ABE	59.93%	63.00%	63.27%	41.29%
23	J	AB	ABE & CDF	59.80%	62.85%	63.10%	41.25%
24	J	AB	EFG	59.56%	62.48%	62.41%	38.64%
25	J	AB	EFG & HIJ	59.66%	62.60%	62.65%	37.59%
26	J	AB & CD	-	55.99%	59.13%	59.13%	39.64%
27	J	AB & CD	ABE	62.51%	65.32%	65.20%	35.21%
28	J	AB & CD	ABE & CDF	63.17%	65.91%	65.72%	34.50%
29	J	AB & CD	EFG	63.32%	65.93%	65.75%	32.91%
30	J	AB & CD	EFG & HIJ	62.85%	65.50%	65.27%	32.70%

Chapter 4

Model Selection

4.1 Introduction

In previous chapters of this work, we used a simple linear regression to model the quantitative outcome variable against the predictor derived from AQMDR, known as the aggregated score. The reasoning behind this model selection was in interest of consistent comparison with current methodology. However, a general strategy for this type of analysis may require model selection. We have shown that AQMDR is an adequate method for developing a predictor, but in some cases a simple linear regression including the aggregated score may be improved upon by performing polynomial regression, or by taking other transformations of the aggregated score. In this chapter, we provide a simple example of this model selection, and simulation results for the new model.

4.2 The Quadratic Model

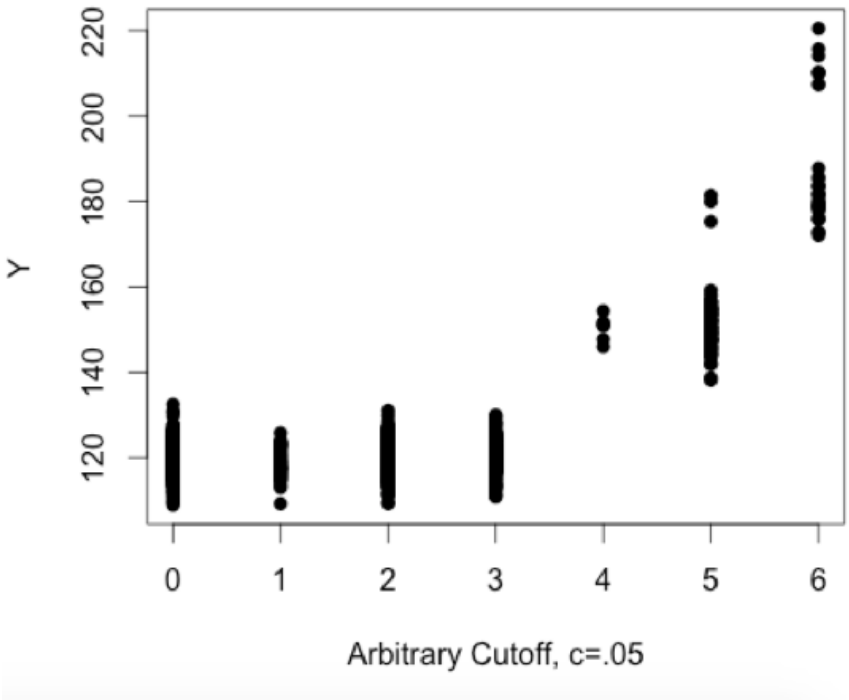
Consider an example scenario with 4 SNPs of interest ($SNP1, SNP2, SNP3, SNP4$) and an outcome variable which is simulated with a normal distribution, where the mean of this normal distribution was increased by the presence of interactions. That is,

$$\begin{aligned} Y_i = 120 &+ 30I[SNP1_i = 2]I[SNP2_i = 2] + 30I[SNP1_i = 2]I[SNP3_i = 2] \\ &+ 30I[SNP1_i = 2]I[SNP4_i = 2] + \epsilon_i \end{aligned} \tag{4.1}$$

where $\epsilon_i \overset{iid}{\sim} N(0, 16)$ and Y_i is the value of some quantitative trait for individual i. For 1000 simulated testing subjects, the aggregated score using an arbitrary cutoff of $c = .05$ (ACAS) based on 1000 training subjects was calculated. The scatterplot

of Y vs. the numerator of the arbitrary cutoff aggregated score for the testing data is displayed in figure 4.1. By examining this plot, one can see that a simple linear

Figure 4.1: Scatterplot of Y vs. the Numerator of the Aggregated Score



regression may not be appropriate in this case. In fact, the relationship between the response and the aggregated score seems to follow a parabolic pattern. Based on this visual analysis, a quadratic model may be a better fit. In fact, suppose we have the following:

Full model: $Y_i = \beta_0 + \beta_1 AggregatedScore_i + \beta_2 AggregatedScore_i^2 + \epsilon_i$

Reduced model: $Y_i = \beta_0 + \beta_1 AggregatedScore_i + \epsilon_i$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. If we examine the hypotheses $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$, we observe a p-value that is less than .001, and reject H_0. Thus, we determine that the full, quadratic model is a better fit than the reduced, simple linear model.

With the exception of the variance used in simulation of phenotypes, simulation model 4.1 is very similar to the simulation scenarios used in previous chapters of this

work. In what follows, we recreate the simulation scenarios presented in chapter 2 and chapter 3, but this time we also use the following quadratic data analysis model to replace the simple linear data analysis model previously employed in AQMDR:

$$Y_i = \beta_0 + \beta_1 AggregatedScore_i + \beta_2 AggregatedScore_i^2 + \epsilon_i \tag{4.2}$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

4.3 Simulation Results: Two-way Interactions

Recall the simulation scenarios for two-way interactions presented in chapter 2 of this work. The study was a factorial design in which we examined variations of the number of SNPs to be considered and the number of interactions present in nature. Let M be the number of SNPs under consideration, and x be the number of two-way interactions present. SNPs were generated under the assumption of Hardy-Weinberg equilibrium [13]. For example, consider a SNP, call it SNPA, with two alleles, each of which take states of A or a. We let the probability that an individual carries allele A on a single chromosome be $p = .5$, and the probability that a person carries allele a on a single chromosome be $q = .5$. Then the SNP states AA and aa each have probability $p^2 = .25$ and $q^2 = .25$ of occurring. SNP state Aa occurs with probability $2pq = .5$. Phenotypes for individuals were randomly generated by a $N(120, \sigma^2)$ distribution, where the mean of this normal distribution was increased by the presence of an interaction. For example, suppose that the two-way interaction of SNP1 and SNP2 was included in simulation. Then the mean of the random normal distribution used to generate the phenotype was increased by 30 in the event that an individual carried a genotypic combination such as (SNP1 = 2)(SNP2 = 2). This example simulation model is displayed in equation 4.3. In the factorial simulation study, we varied the standard deviation, σ of the normal distribution used to generate

phenotypes as well. See table 2.1 for the particular variations considered in this factorial simulation study. Each training and testing data set contained phenotypes and SNP states for 1000 simulated subjects.

$$Y_i = 120 + 30(I[SNP1_i = 2]I[SNP2_i = 2]) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{4.3}$$

For each combination of x, M, and σ^2 variants, 100 independent training and testing data sets were generated. After predictions for the phenotypes in the testing data were made based on the training data, mean squared prediction error (MSPE) was calculated for each testing set. The results of the simulation study are presented in table 4.1. Note that this table is identical to table 2.2 except for the addition of a new column called "Quad". This column contains the average MSPEs obtained from implementing AQMDR (with ACAS, $c = .05$) and using the model expressed in equation 4.2. In all scenarios, the Quadratic AQMDR MSPEs are lower than those produced by the five original AQMDR methods (CWAS, ACAS (.05), ACAS (.20), HAS (.05) and HAS (.20)). The Quad MSPEs are also lower than MSPEs produced by QMDR in all scenarios. In many scenarios, AQMDR with a quadratic model produced results that are very close to the Oracle approach.

To further explore the distinctions between methods in the simulation study, a single factor, within-subjects ANOVA was performed for each simulation scenario, followed by post hoc multiple pairwise comparisons among AQMDR, Quadratic AQMDR, QMDR and the oracle method. For simplicity, we selected ACAS ($c = .05$) as the representative variation of the original AQMDR, to be compared with QMDR, the Oracle Method, and Quadratic AQMDR (Quad). For all simulation scenarios, the ANOVA yielded significance at $\alpha = .05$ level, suggesting that there are differences among the average MSPEs for the prediction methods. The prediction methods were ranked according to the results from multiple pairwise comparisons. The rankings

Table 4.1: Average MSPE over 100 independent testing data sets for two-way interaction simulation study (with Quadratic AQMDR included).

M	x	σ	CWAS	ACAS (.05)	ACAS (.20)	HAS (.05)	HAS (.20)	QMDR	Oracle	TM	Quad
4	1	10	127.19	126.59	127.27	128.30	128.09	107.78	100.55	153.96	101.93
4	1	15	255.00	254.51	254.79	255.92	255.80	247.02	225.65	277.17	231.77
4	1	20	437.61	437.03	437.40	439.31	438.87	431.89	403.66	456.85	416.76
10	1	10	133.91	131.25	132.91	131.82	133.47	109.41	100.61	153.67	103.18
10	1	15	262.52	258.09	260.87	259.29	261.85	247.03	225.28	277.70	232.19
10	1	20	444.32	438.73	440.95	440.42	442.51	435.32	402.44	454.40	417.65
4	3	10	203.78	203.78	203.78	203.78	203.78	225.70	99.57	320.45	114.06
4	3	15	331.22	331.26	331.26	331.26	331.26	354.72	225.45	448.18	241.28
4	3	20	505.97	506.35	506.20	506.36	506.20	525.42	402.17	620.46	424.07
10	3	10	223.39	220.25	223.18	220.87	223.68	226.41	100.41	320.52	149.30
10	3	15	349.97	345.91	350.19	346.67	350.87	352.16	224.93	446.17	277.99
10	3	20	528.69	523.60	528.67	524.69	529.72	527.73	400.41	623.86	458.92
4	6	10	423.23	423.23	423.23	423.23	423.23	561.48	101.14	675.00	190.14
4	6	15	554.04	554.04	554.04	554.04	554.04	693.65	227.01	805.64	320.52
4	6	20	719.00	719.00	719.00	719.00	719.00	856.36	403.00	970.46	500.87
10	6	10	338.23	329.28	336.23	330.49	337.40	561.73	101.42	674.75	109.62
10	6	15	458.38	450.28	456.92	451.36	457.97	683.20	225.75	800.40	235.75
10	6	20	631.19	622.48	630.08	623.71	631.07	844.66	402.48	954.56	413.61

from multiple pairwise comparisons with a Bonferroni correction based on the total number of comparisons ($\alpha = \frac{.05}{6}$) are displayed in table 4.2.

Table 4.2: Within-subjects ANOVA with a Bonferroni correction, $\alpha = \frac{.05}{6}$ for two-way interaction simulation study (with Quadratic AQMDR included).

M	x	σ	ACAS (.05)	QMDR	Oracle	Quad
4	1	10	c	b	a	a
4	1	15	d	c	a	b
4	1	20	d	c	a	b
10	1	10	c	b	a	a
10	1	15	d	c	a	b
10	1	20	c	c	a	b
4	3	10	c	d	a	b
4	3	15	c	d	a	b
4	3	20	c	d	a	b
10	3	10	c	d	a	b
10	3	15	c	d	a	b
10	3	20	c	c	a	b
4	6	10	c	d	a	b
4	6	15	c	d	a	b
4	6	20	c	d	a	b
10	6	10	b	c	a	a
10	6	15	c	d	a	b
10	6	20	c	d	a	b

For each simulation scenario (a row in the table), prediction methods were assigned rankings a, b , c, etc. where a is the ranking assigned to the method with the lowest average MSPE. Methods sharing a letter ranking were not determined to be statistically significant in a pairwise comparison. Note that in every scenario, Quadratic AQMDR receives a higher ranking than ACAS and QMDR, and in some scenarios receives a ranking equivalent to that of the Oracle Method.

4.4 Simulation Results: Three-way Interactions

Recall the simulation scenarios for three-way interactions presented in chapter 3. Again, we performed a simulation study using a factorial design in which we examined variations of the number of SNPs to be considered (M) and the number of present interactions (x). SNPs were generated in the same manner discussed in the previous

section. Phenotypes for subjects were generated using a $N(120, \sigma^2)$ distribution. The mean of this normal distribution was increased by 30 with the presence of a three-way interaction. For example, suppose that an interaction between SNP1, SNP2 and SNP3 was included in a simulation scenario. In this case, the mean of the random normal distribution used to generate the phenotype was increased in the event that an individual carried a genotypic combination such as (SNP1 = 2)(SNP2 = 2)(SNP3 =2). This example simulation scenario is displayed in equation 4.4. In the factorial simulation study, we varied the standard deviation of the normal distributions used to generate the phenotypes. See table 3.1 for the particular variations considered in this study. Each training and testing data set contained phenotypes and SNP states for 1000 subjects.

$$Y_i = 120 + 30(I[SNP1_i = 2]I[SNP2_i = 2]I[SNP3_i = 2]) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \quad (4.4)$$

For each combination of variants (M, x, σ), 100 independent training and testing data sets were generated. The methods considered were the same as those described in chapter 2, including AQMDR (ACAS and HAS with cutoffs of $c = .05$ and $c = .20$, and CWAS), QMDR, the training mean method and the oracle method. As we considered only three-way interactions in this simulation study, the QMDR method was limited only to the selection of three-way interaction models. After predictions for the phenotypes in the testing data were made based on the training data, the mean squared prediction error (MSPE) was calculated for each testing set. The results of the factorial simulation study are displayed in table 4.3. Each value in the table is the average MSPE over 100 independent testing data sets. Note that this table is identical to table 3.2 except for the addition of a new column called "Quad". This column contains the average MSPEs obtained from implementing AQMDR (with ACAS, $c = .05$) and using the model expressed in equation 4.2. Again we see that

in all scenarios, the Quadratic AQMDR MSPEs are lower than those produced by the five original AQMDR methods (CWAS, ACAS (.05), ACAS (.20), HAS (.05) and HAS (.20)). The Quad MSPEs are also lower than MSPEs produced by QMDR in all scenarios.

Table 4.3: Average MSPE over 100 independent testing data sets for three-way interaction simulation study (with Quadratic AQMDR included).

M	x	σ	CWAS	ACAS (.05)	ACAS (.20)	HAS (.05)	HAS (.20)	QMDR	Oracle	TM	Quad
4	1	10	117.26	117.26	117.26	117.26	117.26	105.21	100.17	208.32	102.91
4	1	15	250.27	250.32	250.32	250.32	250.32	249.87	225.50	332.00	233.68
4	1	20	435.61	436.10	436.09	436.29	436.27	441.55	400.84	508.61	419.16
10	1	10	147.45	144.75	146.96	144.89	147.07	106.28	99.91	208.83	114.14
10	1	15	271.20	265.96	271.13	266.25	271.34	246.01	223.67	330.63	240.71
10	1	20	456.11	447.47	455.27	447.94	455.64	438.04	400.56	507.22	428.03
4	2	10	152.08	152.08	152.08	152.08	152.08	201.06	101.15	345.97	131.69
4	2	15	278.18	278.18	278.18	278.18	278.18	323.70	223.56	470.72	257.36
4	2	20	457.59	457.59	457.59	457.59	457.59	497.20	397.00	644.16	434.38
10	2	10	175.67	173.43	174.89	173.50	174.96	200.34	100.27	346.25	124.85
10	2	15	303.76	301.19	302.78	301.28	302.87	323.12	225.83	468.38	254.09
10	2	20	487.30	483.83	485.95	483.97	486.08	508.05	401.86	646.56	437.34

A single factor, within-subjects ANOVA was performed for each of the twelve simulation scenarios, followed by post hoc multiple pairwise comparisons among AQMDR, Quadratic AQMDR, QMDR and the oracle method. For simplicity, we selected ACAS ($c = .05$) as the representative variation of the original AQMDR, to be compared with QMDR, the Oracle Method, and Quadratic AQMDR (Quad). In each simulation scenario, the ANOVA yielded significance at $\alpha = .05$ level, suggesting that there are differences among the average MSPEs for the methods. Prediction methods were ranked according to results from multiple pairwise comparisons (with a Bonferroni correction, $\alpha = \frac{.05}{6}$). These rankings are displayed in table 4.4.

Table 4.4: Within-subjects ANOVA with a Bonferroni correction, $\alpha = \frac{.05}{6}$ for three-way interaction simulation study (with Quadratic AQMDR included).

M	x	σ	ACAS (.05)	QMDR	Oracle	Quad
4	1	10	d	c	a	b
4	1	15	c	c	a	b
4	1	20	c	d	a	b
10	1	10	d	c	a	b
10	1	15	d	c	a	b
10	1	20	d	c	a	b
4	2	10	c	d	a	b
4	2	15	c	d	a	b
4	2	20	c	d	a	b
10	2	10	c	d	a	b
10	2	15	c	d	a	b
10	2	20	c	d	a	b

For each simulation scenario (represented by a row in the table), prediction methods were assigned rankings a, b, c, etc. where a is the ranking assigned to the method with the lowest average MSPE. Again, methods sharing a letter ranking were not determined to be statistically significant in a pairwise comparison (that is, there is no significant difference between the methods). As we saw with the simulations for two-way interactions, the Quadratic AQMDR method performed better than ACAS ($c = .05$) and QMDR. However, there were no scenarios in which Quadratic AQMDR received a rank equivalent to that of the Oracle Method.

4.5 Simulation Results: Two-way and Three-way Interactions Combined

In chapter 3, we also examined the combination of two-way and three-way interactions through a simulation study. In each simulation, we consider ten SNPs ($M = 10$) and SNP states are generated using Hardy-Weinberg Equilibrium [13]. For example, consider a SNP, call it SNPA, with two alleles, each of which take states of A or a. We let the probability that an individual carries allele A on a single chromosome be $p = .5$, and the probability that a person carries allele a on a single chromosome

be $q = .5$. Then the SNP states AA and aa each have probability p^2 and q^2 of occurring. SNP state Aa occurs with probability $2pq$. In the two previous sections, p was fixed at $p = .5$. As we saw in chapter 3, for the study of two and three-way interactions combined, the value of p for each SNP is generated using a continuous uniform distribution on (.5, .9). Phenotypes for individuals were randomly generated by a $N(120, \sigma^2)$ distribution, where the mean of this normal distribution was increased by 30 with the presence of an interaction. For example, suppose that the two-way interaction of SNP1 and SNP2 was included in simulation. Then the mean of the random normal distribution used to generate the phenotype was increased in the event that an individual carried a genotypic combination such as (SNP1 = 2)(SNP2 = 2). The standard deviation of this normal distribution was fixed for each scenario in the study.

In table 4.5, the "One-way" column indicates which of the ten SNPs considered in simulation (SNPA, SNPB,..., SNPJ) are included as a one-way interaction (main effect). For example, if SNPJ is included as a main effect in simulation, then "J" appears in this column. The "Two-way" column indicates which two-way interactions were included. For example, if the interaction between SNPA and SNPB was included in simulation, then AB appears in this column. Similarly, the "Three-way" column indicates which three-way interactions were included. For example, if the interaction between SNPC, SNPD and SNPE was included in simulation, then "CDE" appears in this column. The magnitude of all one-way, two-way and three-way interactions included in simulation was 30. That is, the presence of an interaction in simulation increased the mean of the random normal distribution by 30. When calculating aggregated scores in the AQMDR method, the arbitrary cutoff aggregated score with a cutoff of $c = .05$ was used. Further, the Simultaneous Inclusion method (see chapter 3) was selected as the representative AQMDR method in this study. That is,

Simultaneous Inclusion was used for both the original AQMDR implementation and the Quadratic AQMDR implementation, respectively labeled "Sim." and "Sim-Quad" in table 4.5. Note that "Sim-Quad" data analysis model required a quadratic term for the two-way interaction aggregated score and a quadratic term for the three-way aggregated score. Each value in the "Sim.", "Sim-Quad", and "QMDR" columns in table 4.5 is the average MSPE calculated from 100 independent testing data sets. For discussion that follows, we have assigned a numerical value to each scenario in the first column of table 4.5. One can see that Quadratic AQMDR yielded smaller average MSPEs than the original AQMDR in all simulation scenarios (with the exception of scenarios 1 and 16 in which no interactions were present). Quadratic AQMDR also produced lower average MSPEs than QMDR in most simulation scenarios. However, in scenarios 1-6 and 16, QMDR performed better than Quadratic AQMDR.

Table 4.6 displays the results of post hoc multiple pairwise comparisons among the prediction methods, performed after single factor, within subjects ANOVA tests which all yielded significance at the $\alpha = .05$ significance level. The prediction methods were ranked according to the multiple pairwise comparisons (with Bonferroni correction, $\alpha = \frac{.05}{3}$), where a ranking of a is assigned to the method with the lowest MSPE. As in previous sections, methods sharing a letter ranking were not determined to be statistically different. Note that Quadratic AQMDR received a ranking of "a" in nearly all scenarios, except for 1-6 and 16 as mentioned above.

The example presented in this chapter implores us to recommend model selection as a general strategy for AQMDR. Although we saw promising empirical results in chapters 2 and 3 of this work, these results were improved with the implementation of the quadratic model in place of the simple linear model applied in previous chapters. Other models, such as a piece-wise linear model may have also yielded promising results.

Table 4.5: Average MSPEs for two-way and three-way interactions combined (with Quadratic AQMDR included).

No.	One-way	Two-way	Three-way	Sim.	Sim-Quad	QMDR
1	-	-	-	239.69	241.85	229.69
2	-	-	ABE	261.20	247.60	229.65
3	-	-	ABE & CDF	261.85	250.80	229.75
4	-	-	EFG	261.68	247.28	230.93
5	-	-	EFG & HIJ	258.02	242.99	229.04
6	-	AB	-	267.84	240.29	230.91
7	-	AB	ABE	301.97	279.18	311.47
8	-	AB	ABE & CDF	301.70	279.22	304.53
9	-	AB	EFG	292.28	280.61	358.02
10	-	AB	EFG & HIJ	288.42	277.01	358.65
11	-	AB & CD	-	303.04	280.28	355.12
12	-	AB & CD	ABE	339.34	318.10	465.36
13	-	AB & CD	ABE & CDF	338.14	320.09	481.86
14	-	AB & CD	EFG	327.28	313.96	511.12
15	-	AB & CD	EFG & HIJ	326.83	313.37	511.89
16	J	-	-	230.71	230.90	227.02
17	J	-	ABE	261.74	255.56	359.98
18	J	-	ABE & CDF	261.85	256.58	366.59
19	J	-	EFG	261.74	255.81	363.91
20	J	-	EFG & HIJ	258.51	252.98	352.67
21	J	AB	-	271.15	259.59	296.91
22	J	AB	ABE	305.91	290.42	488.95
23	J	AB	ABE & CDF	302.56	289.19	481.66
24	J	AB	EFG	295.72	286.20	482.74
25	J	AB	EFG & HIJ	292.92	283.20	489.47
26	J	AB & CD	-	305.49	290.70	451.17
27	J	AB & CD	ABE	341.44	328.28	635.78
28	J	AB & CD	ABE & CDF	340.91	327.75	651.35
29	J	AB & CD	EFG	330.26	319.47	646.87
30	J	AB & CD	EFG & HIJ	329.04	318.07	637.61

Table 4.6: Within-subjects ANOVA with a Bonferroni correction, $\alpha = \frac{.05}{3}$ for two-way and three-way interactions combined (with Quadratic AQMDR included).

No.	One-way	Two-way	Three-way	Sim.	Sim-Quad	QMDR
1	-	-	-	b	b	a
2	-	-	ABE	c	b	a
3	-	-	ABE and CDF	c	b	a
4	-	-	EFG	c	b	a
5	-	-	EFG and HIJ	c	b	a
6	-	AB	-	c	b	a
7	-	AB	ABE	b	a	c
8	-	AB	ABE and CDF	b	a	c
9	-	AB	EFG	a	a	b
10	-	AB	EFG and HIJ	b	a	c
11	-	AB and CD	-	b	a	c
12	-	AB and CD	ABE	b	a	c
13	-	AB and CD	ABE and CDF	b	a	c
14	-	AB and CD	EFG	b	a	c
15	-	AB and CD	EFG and HIJ	b	a	c
16	J	-	-	ab	b	a
17	J	-	ABE	a	a	b
18	J	-	ABE and CDF	a	a	b
19	J	-	EFG	a	a	b
20	J	-	EFG and HIJ	a	a	b
21	J	AB	-	b	a	c
22	J	AB	ABE	b	a	c
23	J	AB	ABE and CDF	b	a	c
24	J	AB	EFG	a	a	b
25	J	AB	EFG and HIJ	a	a	b
26	J	AB and CD	-	b	a	c
27	J	AB and CD	ABE	b	a	c
28	J	AB and CD	ABE and CDF	b	a	c
29	J	AB and CD	EFG	a	a	b
30	J	AB and CD	EFG and HIJ	a	a	b

Chapter 5

Theoretical Considerations

5.1 Theoretical Support of AQMDR with Arbitrary Cutoff Aggregated Score: Two-way Interactions

Suppose we have the following 2×2 table: Assume $\mu_{high} > \mu_{low}$. $Y_1, ..., Y_{4n}$

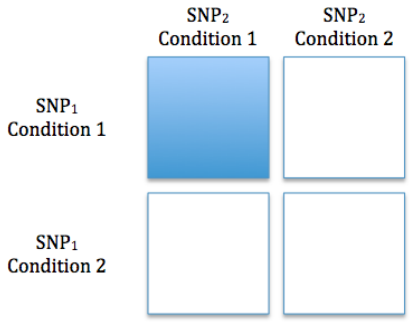
Table 5.1: Distribution of Y phenotypes among SNP condition combinations.

	SNP_2 Condition 1	SNP_2 Condition 2
SNP_1 Condition 1	$Y_1, ..., Y_n \overset{iid}{\sim} N(\mu_{high}, \sigma^2)$	$Y_{n+1}, ..., Y_{2n} \overset{iid}{\sim} N(\mu_{low}, \sigma^2)$
SNP_1 Condition 2	$Y_{2n+1}, ..., Y_{3n} \overset{iid}{\sim} N(\mu_{low}, \sigma^2)$	$Y_{3n+1}, ..., Y_{4n} \overset{iid}{\sim} N(\mu_{low}, \sigma^2)$

represent the values of some quantitative phenotype for each subject, $i = 1, 2, ..., 4n$. For interpretative purposes, for SNPs with three states (AA, Aa, and aa), Condition 1 might be the combination of SNP states AA and Aa, while Condition 2 is the SNP state aa. Alternatively, Condition 1 may be the SNP state AA, while condition 2 is the combination of states Aa and aa. By relabeling condition 1 and condition 2, the configuration in table 5.1 covers any situation in which the phenotype mean of one quadrant is higher than that of the other three quadrants (under the assumption that the other three quadrants all have the same mean).

Let p_1 be the high/low classification from the AQMDR method of a 2×2 table in which cell 1 (the cell in the top left) is correctly identified as "high" and the three remaining cells are classified as "low". See figure 5.1, in which the shaded cell is the one categorized as "high", and the light cells are categorized as "low". In what follows, we assume that AQMDR is applied to two-way interactions in which there are four multifactor classes, rather than the nine multifactor classes we observed for each two-way interaction in previous chapters.

Figure 5.1: The p_1 "high"/"low" configuration.



Our goal is to prove that the under the parameter state in table 5.1, the probability that the gene-to-gene interaction of SNP_1 and SNP_2 is included in the AQMDR Arbitrary Cutoff aggregated score (ACAS with fixed $c > 0$) and that the p_1 high/low configuration of this interaction is correctly assigned in AQMDR tends to 1 as n goes to infinity. To simplify notation, all probability calculations in this proof are based on the parameter state shown in table 5.1. To clarify, suppose we have the following events:

1. Event A: the event that the gene-to-gene interaction of SNP_1 and SNP_2 is included in the AQMDR arbitrary cutoff aggregated score. That is, the interaction receives a weight of 1.

2. Event B: the event that the p_1 high/low configuration of this interaction is correctly assigned in AQMDR.

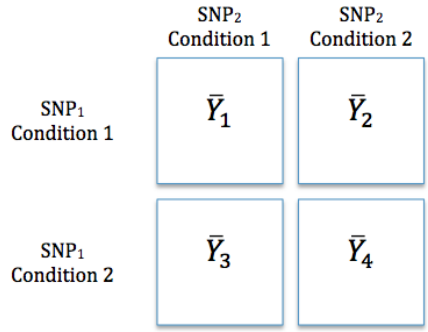
We want to show that $P(A \cap B) \to 1$ as $n \to \infty$.

The proof proceeds as follows:

Without loss of generality, let $c = .05$. First note that $P(B) = P((\bar{Y}_1 > \bar{\bar{Y}}) \cap (\bar{Y}_2 \leq \bar{\bar{Y}}) \cap (\bar{Y}_3 \leq \bar{\bar{Y}}) \cap (\bar{Y}_4 \leq \bar{\bar{Y}}))$, where $\bar{\bar{Y}}$ is the overall mean of of $Y_1, Y_2, ..., Y_{4n}$, and $\bar{Y}_j$

for $j \in \{1, 2, 3, 4\}$ is the mean within cell j. That is, $\bar{Y}_1$ is the mean of $Y_1, Y_2, ..., Y_n$, $\bar{Y}_2$ is the mean of $Y_{n+1}, Y_{n+2}, ..., Y_{2n}$, etc. See figure 5.2.

Figure 5.2: Individual cell means.



By DeMorgan's Law and Boole's inequality, we have

$$\begin{aligned} P(A \cap B) &= 1 - P(A^c \cup B^c) \\ &\geq 1 - P(A^c) - P(B^c) \\ &= 1 - P(A^c) - P((\bar{Y}_1 \leq \bar{\bar{Y}}) \cup (\bar{Y}_2 > \bar{\bar{Y}}) \cup (\bar{Y}_3 > \bar{\bar{Y}}) \cup (\bar{Y}_4 > \bar{\bar{Y}})) \\ &\geq 1 - P(A^c) - (P(\bar{Y}_1 \leq \bar{\bar{Y}}) + P(\bar{Y}_2 > \bar{\bar{Y}}) + P(\bar{Y}_3 > \bar{\bar{Y}}) + P(\bar{Y}_4 > \bar{\bar{Y}})) \end{aligned} \tag{5.1}$$

By the Weak Law of Large Numbers and Slutsky's Theorem, we have the following:

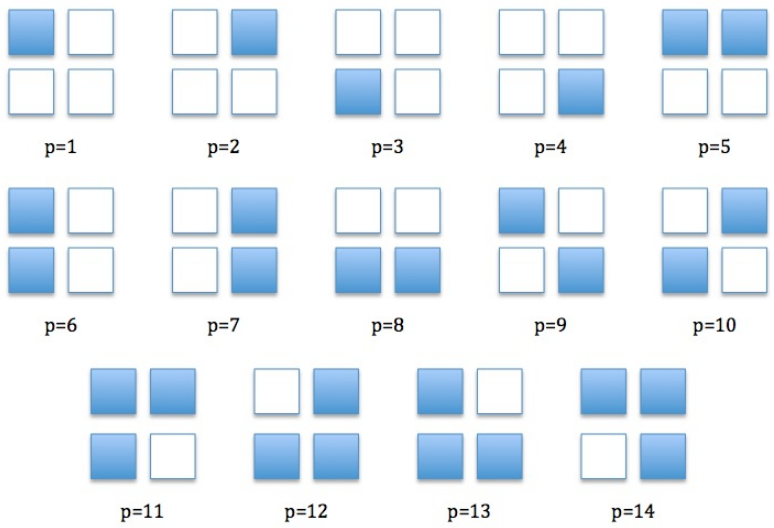
1. $\bar{Y}_1 \xrightarrow{p} \mu_{high}$, as $n \to \infty$.
2. $\bar{Y}_2 \xrightarrow{p} \mu_{low}$, $\bar{Y}_3 \xrightarrow{p} \mu_{low}$, $\bar{Y}_4 \xrightarrow{p} \mu_{low}$, as $n \to \infty$.
3. $\bar{\bar{Y}} = \frac{n(\bar{Y}_1+\bar{Y}_2+\bar{Y}_3+\bar{Y}_4)}{4n} = \frac{(\bar{Y}_1+\bar{Y}_2+\bar{Y}_3+\bar{Y}_4)}{4} \xrightarrow{p} \frac{\mu_{high}+3\mu_{low}}{4} := \mu$, as $n \to \infty$.
4. $\mu < \mu_{high}$ and $\mu > \mu_{low}$.

Thus, as $n \to \infty$, $P(\bar{Y}_1 \leq \bar{\bar{Y}}) + P(\bar{Y}_2 > \bar{\bar{Y}}) + P(\bar{Y}_3 > \bar{\bar{Y}}) + P(\bar{Y}_4 > \bar{\bar{Y}}) \to 0$.

This leaves only the consideration of $P(A^c)$. Recall that in AQMDR with ACAS, the inclusion of a gene-to-gene interaction in the aggregated score is based on a p-value determined from permutation testing (see chapter 2). When $c = .05$, an interaction is assigned a weight of 1 in the aggregated score if the permutation p-value associated with that interaction is less than .05. Our goal is to show that, with probability approaching 1 as $n \to \infty$, the p-value achieved by permutation for the T-statistic corresponding to this interaction is less than .05. That is, the probability that the interaction is included in the aggregated score tends to 1 as $n \to \infty$. By establishing this, we can then determine that $P(A^c) \to 0$ as $n \to \infty$. We will prove this with a series of corollaries. Then, by the expression in (5.1), we will have shown that $P(A \cap B) \to 1$ as $n \to \infty$.

First note that (disregarding 0 probability events) there are 14 possible "high"/"low" configurations (labeling of quadrants as "high" or "low") which can be assigned to a 2×2 table with AQMDR. Let them be indexed by $p = 1, 2, ..., 14$. See figure 5.3, in which shaded cells are categorized as "high" and the light cells are categorized as "low".

Figure 5.3: 14 possible "high"/"low" configurations.



We will define $c_i(p)$ for each $p \in \{1, 2, ..., 14\}$, as shown in table 5.2. Note that for all $p \in \{1, 2, ..., 14\}$, $\sum_{i=1}^{4n} c_i(p) = 0$. Define $T(p)$ for each $p \in \{1, ..., 14\}$ as follows:

Table 5.2: $c_i(p)$ for each $p \in \{1, 2, ..., 14\}$.

p	$i \in \{1, ..., n\}$	$i \in \{n+1, ..., 2n\}$	$i \in \{2n+1, ..., 3n\}$	$i \in \{3n+1, ..., 4n\}$
1	3	-1	-1	-1
2	-1	3	-1	-1
3	-1	-1	3	-1
4	-1	-1	-1	3
5	1	1	-1	-1
6	1	-1	1	-1
7	-1	1	-1	1
8	-1	-1	1	1
9	1	-1	-1	1
10	-1	1	1	-1
11	1	1	1	-3
12	-3	1	1	1
13	1	-3	1	1
14	1	1	-3	1

$$T(p) := \frac{\sum_{i=1}^{4n} c_i(p) Y_i}{\sqrt{\sum_{i=1}^{4n} c_i(p)^2}}.$$

Note that for any p with one cell categorized as "high" and three cells categorized as "low", or one cell categorized as "low" and three cells categorized as "high",

$$\sum_{i=1}^{4n} c_i(p)^2 = \sum_{i=1}^{n} (3)^2 + \sum_{i=1}^{3n} (-1)^2 = 12n.$$

This includes $p \in \{1, 2, 3, 4, 11, 12, 13, 14\}$. For all other values of p,

$$\sum_{i=1}^{4n} c_i(p)^2 = \sum_{i=1}^{2n} (1)^2 + \sum_{i=1}^{2n} (-1)^2 = 4n.$$

Corollary 1: $T(p) \sim N(k_n(p)(\mu_{high} - \mu_{low}), \sigma^2)$, for a constant $k_n(p)$.

Proof: There are six cases we will consider for the distribution of $T(p)$.

Case 1: One cell is classified as "high", and the "high" cell is cell 1. That is, $p = 1$.

Consider T(1).

$$\begin{aligned} T(1) =& \frac{\sum_{i=1}^{4n} c_i(1) Y_i}{\sqrt{12n}} \\ =& \frac{\sum_{i=1}^{n} (3) Y_i + \sum_{i=n+1}^{4n} (-1) Y_i}{\sqrt{12n}} \\ =& \frac{3n\bar{Y}_{high} - 3n\bar{Y}_{low}}{\sqrt{12n}} \\ =& \frac{3n(\bar{Y}_{high} - \bar{Y}_{low})}{\sqrt{12n}} \end{aligned}$$

where $\bar{Y}_{high}$ is a sample mean of realized values of Y_i from a cell with mean μ_{high} (cell 1), and $\bar{Y}_{low}$ is a sample mean of realized values of Y_i from cells with mean μ_{low} (cells 2, 3, and 4). As $\sum_{i=1}^{4n} c_i(1) = 0$,

$$T(1) \sim N\left(k_n(1)(\mu_{high} - \mu_{low}), \sigma^2\right)$$

where $k_n(1) = \frac{3n}{\sqrt{12n}} = \left(\frac{3n}{4}\right)^{1/2}$.

Case 2: One cell is classified as "high", and the "high" cell is not cell 1. That is, $p \in \{2, 3, 4\}$.

Consider T(2).

$$\begin{aligned} T(2) =& \frac{\sum_{i=1}^{4n} c_i(2) Y_i}{\sqrt{12n}} \\ =& \frac{\sum_{i=1}^{n}(-1)Y_i + \sum_{i=n+1}^{2n}(3)Y_i + \sum_{i=2n+1}^{4n}(-1)Y_i}{\sqrt{12n}} \\ =& \frac{-n\bar{Y}_{high} + 3n\bar{Y}_{low} - 2n\bar{Y}_{low}}{\sqrt{12n}} \end{aligned}$$

As $\sum_{i=1}^{4n} c_i(2) = 0$,

$$T(2) \sim N\left(\frac{(-n\mu_{high} + 3n\mu_{low} - 2n\mu_{low})}{\sqrt{12n}}, \sigma^2\right) = N\left(k_n(2)\left(\mu_{high} - \mu_{low}\right), \sigma^2\right)$$

where $k_n(2) = \frac{-n}{\sqrt{12n}} = -\left(\frac{n}{12}\right)^{1/2}$. The same is true for $T(3)$ and $T(4)$.

Case 3: Two cells are classified as "high", and one of the "high" cells is cell 1. That is, $p \in \{5, 6, 9\}$.

Consider T(5).

$$\begin{aligned} T(5) =& \frac{\sum_{i=1}^{4n} c_i(5) Y_i}{\sqrt{4n}} \\ =& \frac{\sum_{i=1}^{2n} (1) Y_i + \sum_{i=n+1}^{2n} (1) Y_i + \sum_{i=2n+1}^{4n} (-1) Y_i}{\sqrt{4n}} \\ =& \frac{n\bar{Y}_{high} + n\bar{Y}_{low} - 2n\bar{Y}_{low}}{\sqrt{4n}} \end{aligned}$$

As $\sum_{i=1}^{4n} c_i(5) = 0$,

$$T(5) \sim N\left(\frac{(n\mu_{high} + n\mu_{low} - 2n\mu_{low})}{\sqrt{4n}}, \sigma^2\right) = N\left(k_n(5)\left(\mu_{high} - \mu_{low}\right), \sigma^2\right)$$

where $k_n(5) = \frac{n}{\sqrt{4n}} = \left(\frac{n}{4}\right)^{1/2}$. The same is true for $T(6)$ and $T(9)$.

Case 4: Two cells are classified as "high", but cell 1 is classified as "low". That is, $p \in \{7, 8, 10\}$.

Consider T(7).

$$\begin{aligned} T(7) =& \frac{\sum_{i=1}^{4n} c_i(7) Y_i}{\sqrt{4n}} \\ =& \frac{\sum_{i=1}^{n} (-1) Y_i + \sum_{i=n+1}^{2n} (1) Y_i + \sum_{i=2n+1}^{3n} (-1) Y_i + \sum_{i=3n+1}^{4n} (1) Y_i}{\sqrt{4n}} \\ =& \frac{-n\bar{Y}_{high} + n\bar{Y}_{low} - n\bar{Y}_{low} + n\bar{Y}_{low}}{\sqrt{4n}} \end{aligned}$$

As $\sum_{i=1}^{4n} c_i(7) = 0$,

$$T(7) \sim N\left(\frac{(-n\mu_{high} + n\mu_{low} - n\mu_{low} + n\mu_{low})}{\sqrt{4n}}, \sigma^2\right) = N\left(k_n(7)\left(\mu_{high} - \mu_{low}\right), \sigma^2\right)$$

where $k_n(7) = \frac{-n}{\sqrt{4n}} = -\left(\frac{n}{4}\right)^{1/2}$. The same is true for $T(8)$ and $T(10)$.

Case 5: Three cells are classified as "high", one of which is cell 1. That is, $p \in \{11, 13, 14\}$.

Consider T(11).

$$\begin{aligned} T(11) =& \frac{\sum_{i=1}^{4n} c_i(11) Y_i}{\sqrt{12n}} \\ =& \frac{\sum_{i=1}^{n}(1)Y_i + \sum_{i=n+1}^{2n}(1)Y_i + \sum_{i=2n+1}^{3n}(1)Y_i + \sum_{i=3n+1}^{4n}(-3)Y_i}{\sqrt{12n}} \\ =& \frac{n\bar{Y}_{high} + n\bar{Y}_{low} + n\bar{Y}_{low} - 3n\bar{Y}_{low}}{\sqrt{12n}} \end{aligned}$$

As $\sum_{i=1}^{4n} c_i(11) = 0$,

$$T(11) \sim N\left(\frac{(n\mu_{high} + n\mu_{low} + n\mu_{low} - 3n\mu_{low})}{\sqrt{12n}}, \sigma^2\right) = N\left(k_n(11)\left(\mu_{high} - \mu_{low}\right), \sigma^2\right)$$

where $k_n(11) = \frac{n}{\sqrt{12n}} = \left(\frac{n}{12}\right)^{1/2}$. The same is true for $T(13)$ and $T(14)$.

Case 6: Three cells are classified as "high", but cell 1 is classified as "low". That is, $p = 12$.

Consider T(12).

$$\begin{aligned} T(12) =& \frac{\sum_{i=1}^{4n} c_i(12) Y_i}{\sqrt{12n}} \\ =& \frac{\sum_{i=1}^{n} (-3) Y_i + \sum_{i=n+1}^{4n} (1) Y_i}{\sqrt{12n}} \\ =& \frac{-3n\bar{Y}_{high} + 3n\bar{Y}_{low}}{\sqrt{12n}} \end{aligned}$$

As $\sum_{i=1}^{4n} c_i(12) = 0$,

$$T(12) \sim N\left(\frac{(-3n\mu_{high} + 3n\mu_{low})}{\sqrt{12n}}, \sigma^2\right) = N\left(k_n(12)\left(\mu_{high} - \mu_{low}\right), \sigma^2\right)$$

where $k_n(12) = \frac{-3n}{\sqrt{12n}} = -\left(\frac{3n}{4}\right)^{1/2}$.

In summary, $T(p) \sim N(k_n(p)(\mu_{high} - \mu_{low}), \sigma^2)$, for a constant $k_n(p)$. The values of $k_n(p)$ for each p are displayed in table 5.3. Note that $k_n(1) = max_{1 \le p \le 14} k_n(p)$. Thus

Table 5.3: $k_n(p)$ for each $p \in \{1, 2, ..., 14\}$.

p	$k_n(p)$	p	$k_n(p)$
1	$\left(\frac{3n}{4}\right)^{1/2}$	8	$-\left(\frac{n}{4}\right)^{1/2}$
2	$-\left(\frac{n}{12}\right)^{1/2}$	9	$\left(\frac{n}{4}\right)^{1/2}$
3	$-\left(\frac{n}{12}\right)^{1/2}$	10	$-\left(\frac{n}{4}\right)^{1/2}$
4	$-\left(\frac{n}{12}\right)^{1/2}$	11	$\left(\frac{n}{12}\right)^{1/2}$
5	$\left(\frac{n}{4}\right)^{1/2}$	12	$-\left(\frac{3n}{4}\right)^{1/2}$
6	$\left(\frac{n}{4}\right)^{1/2}$	13	$\left(\frac{n}{12}\right)^{1/2}$
7	$-\left(\frac{n}{4}\right)^{1/2}$	14	$\left(\frac{n}{12}\right)^{1/2}$

we have proved Corollary 1. Now, define a T-statistic as follows:

$$T := max_{1 \leq p \leq 14} T(p)$$

This T-statistic is not quite identical to the one used in AQMDR (see equation 2.1) because sample variances are not incorporated. We forgo the estimation of the variances because we assume that $Y_1, ..., Y_{4n}$ all have the same underlying variance. Let δ be an arbitrary small positive number. Assume n is large enough such that

$$(k_n(1) - max_{2 \leq p \leq 14} k_n(p))(\mu_{high} - \mu_{low}) \geq \frac{2\sigma}{\delta} \tag{5.2}$$

That is, $[(\frac{3n}{4})^{1/2} - (\frac{n}{4})^{1/2}](\mu_{high} - \mu_{low}) \geq \frac{2\sigma}{\delta}$.

Corollary 2: $P\left(T = T(1)\right) \geq 1 - 13\delta^2$.

Proof: Put $\mu_{diff} := \mu_{high} - \mu_{low}$. Note that $E[T(p) - T(1)] = (k_n(p) - k_n(1))\mu_{diff}$ and $Var[T(p) - T(1)] \leq 4\sigma^2$. By Chebychev's Inequality, for all $p \in \{1, 2, ..., 14\}$ we have

$$P\left(|T(p) - T(1) + (k_n(1) - k_n(p))\mu_{diff}| \geq \frac{2\sigma}{\delta}\right) \leq \delta^2 \tag{5.3}$$

For any $p \in \{2, 3, ..., 14\}$,

$$\begin{aligned} P\left(T(p) > T(1)\right) &= P\left(T(p) - T(1) \geq 0\right) \\ &= P\left(T(p) - T(1) + \left(k_n(1) - k_n(p)\right)\left(\mu_{diff}\right) \geq \left(k_n(1) - k_n(p)\right)\mu_{diff}\right) \\ &\leq P\left(T(p) - T(1) + \left(k_n(1) - k_n(p)\right)\left(\mu_{diff}\right) \geq \frac{2\sigma}{\delta}\right) \text{ by (5.2),} \\ &\leq P\left(\left|T(p) - T(1) + \left(k_n(1) - k_n(p)\right)\left(\mu_{diff}\right)\right| \geq \frac{2\sigma}{\delta}\right) \\ &\leq \delta^2 \text{ by the inequality expressed in (5.3).} \end{aligned}$$

Now, consider $P(T \neq T(1))$. By Boole's inequality, we have

$$\begin{aligned} P\left(T \neq T(1)\right) &= P\left(T(1) \neq max_{1 \leq p \leq 14} T(p)\right) \\ &= P\left(\left(T(2) > T(1)\right) \cup \left(T(3) > T(1)\right) \cup ... \cup \left(T(14) > T(1)\right)\right) \\ &\leq P\left(T(2) > T(1)\right) + P(T(3) > T(1)) + ... + P(T(14) > T(1)) \\ &\leq 13\delta^2 \end{aligned}$$

Hence,

$$P\left(T = T(1)\right) \geq 1 - 13\delta^2 \tag{5.4}$$

and we have proved Corollary 2. Note that since $T(p)$ is normally distributed, this bound is not sharp.

Consider a fixed permutation $\pi : \{1, 2, ..., 4n\} \rightarrow \{1, 2, ..., 4n\}$. Define $T_\pi(p)$ as follows:

$$T_\pi(p) := \frac{\sum_{i=1}^{4n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{4n} c_{\pi(i)}(p)^2}} = \frac{\sum_{i=1}^{4n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{4n} c_i(p)^2}}$$

Define the test statistic for permutation π as follows:

$$T_\pi := max_{1 \leq p \leq 14} T_\pi(p)$$

Corollary 3: For all $p \in \{1, 2, ..., 14\}$:

$$T_\pi(p) \sim N(k_n^\pi(p)(\mu_{diff}), \sigma^2)$$

where $k_n^\pi(p)$ depends upon π only through X_p, where X_p is the number of phenotypes which were originally in cell 1 that remained in a "high" cell after permutation.

Proof: Let $C = \{1, 2, ..., n\}$. We will consider 3 cases for the distribution of $T_\pi(p)$.

Case 1: $p \in \{1, 2, 3, 4\}$

For p configurations in which one cell is categorized as "high", and three are categorized as "low" ($p \in \{1, 2, 3, 4\}$), we define D to be the set of $\pi(i)$'s corresponding to the cell categorized as "high" in the p configuration. For example, if $p = 1$, $D = \{\pi(1), ..., \pi(n)\}$, and if $p = 2$, $D = \{\pi(n + 1), ..., \pi(2n)\}$. Let $X_p = |C \cap D|$. That is, X_p is the number of phenotypes which were originally in cell 1, that remained in a cell classified as "high" after permutation. For all $p \in \{1, 2, 3, 4\}$, we have the following:

1. $|C \cap D^c| = n - X_p$
2. $|C^c \cap D| = n - X_p$
3. $|C^c \cap D^c| = 4n - X_p - (n - X_p) - (n - X_p) = 2n + X_p$

Consider $T_\pi(p)$:

$$\begin{aligned}
T_\pi(p) &= \frac{\sum_{i=1}^{4n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{4n} c_i(p)^2}} \\
&= \frac{\sum_{i=1}^{4n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{12n}} \\
&= \frac{\sum_{C\cap D}(3)Y_{\pi(i)} + \sum_{C^c\cap D}(3)Y_{\pi(i)} + \sum_{C\cap D^c}(-1)Y_{\pi(i)} + \sum_{C^c\cap D^c}(-1)Y_{\pi(i)}}{\sqrt{12n}} \\
&= \frac{3X_p\bar{Y}_{high} + 3(n-X_p)\bar{Y}_{low} - (n-X_p)\bar{Y}_{high} - (2n+X_p)\bar{Y}_{low}}{\sqrt{12n}}
\end{aligned}$$

Thus,

$$\begin{aligned}
T_\pi(p) &\sim N\left(\frac{3X_p\mu_{high} + (3n-3X_p)\,\mu_{low} - (n-X_p)\,\mu_{high} - (2n+X_p)\,\mu_{low}}{\sqrt{12n}}, \sigma^2\right) \\
&= N\left(\frac{(4X_p - n)\,(\mu_{diff})}{\sqrt{12n}}, \sigma^2\right)
\end{aligned}$$

That is, $T_\pi(p) \sim N(k_n^\pi(p)(\mu_{diff}), \sigma^2)$ where $k_n^\pi(p) = \frac{4X_p - n}{\sqrt{12n}}$ depends on π only through X_p.

Case 2: $p \in \{5, 6, 7, 8, 9, 10\}$

For p configurations in which two cells are categorized as "high", and two are categorized as "low" ($p \in \{5, 6, 7, 8, 9, 10\}$), we define D to be the set of $\pi(i)$'s corresponding to cells categorized as "high" in the p configuration. For example, if $p = 5$, $D = \{\pi(1), ..., \pi(2n)\}$, and if $p = 6$, $D = \{\pi(1), ..., \pi(n)\} \cup \{\pi(2n+1), ..., \pi(3n)\}$. Let $X_p = |C \cap D|$. For all $p \in \{5, 6, 7, 8, 9, 10\}$, we have the following:

1. $|C \cap D^c| = n - X_p$
2. $|C^c \cap D| = 2n - X_p$

3. $|C^c \cap D^c| = 4n - X_p - (2n - X_p) - (n - X_p) = n + X_p$

For $p \in \{5, 6, 7, 8, 9, 10\}$, we have

$$\begin{aligned} T_\pi(p) &= \frac{\sum_{i=1}^{4n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{4n} c_i(p)^2}} \\ &= \frac{\sum_{i=1}^{4n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{4n}} \\ &= \frac{\sum_{C \cap D} (1) Y_{\pi(i)} + \sum_{C^c \cap D} (1) Y_{\pi(i)} + \sum_{C \cap D^c} (-1) Y_{\pi(i)} + \sum_{C^c \cap D^c} (-1) Y_{\pi(i)}}{\sqrt{4n}} \\ &= \frac{X_p \bar{Y}_{high} + (2n - X_p) \bar{Y}_{low} - (n - X_p) \bar{Y}_{high} - (n + X_p) \bar{Y}_{low}}{\sqrt{4n}} \end{aligned}$$

Thus,

$$\begin{aligned} T_\pi(p) &\sim N\left(\frac{X_p \mu_{high} + (2n - X_p)\, \mu_{low} - (n - X_p)\, \mu_{high} - (n + X_p)\, \mu_{low}}{\sqrt{4n}}, \sigma^2 \right) \\ &= N\left(\frac{(2X_p - n)\, (\mu_{diff})}{\sqrt{4n}}, \sigma^2 \right) \end{aligned}$$

That is, $T_\pi(p) \sim N(k_n^\pi(p)(\mu_{diff}), \sigma^2)$ where $k_n^\pi(p) = \frac{2X_p - n}{\sqrt{4n}}$ depends on π only through X_p.

Case 3: $p \in \{11, 12, 13, 14\}$

For p configurations where 3 cells are categorized as "high", and 1 is categorized as "low" ($p \in \{11, 12, 13, 14\}$), we define D to be the set of $\pi(i)$'s corresponding to cells categorized as "high" in the p configuration. For example, if $p = 11$, $D = \{\pi(1), ..., \pi(3n)\}$. Let $X_p = |C \cap D|$. For $p \in \{11, 12, 13, 14\}$, we have the following:

1. $|C \cap D^c| = n - X_p$
2. $|C^c \cap D| = 3n - X_p$

3. $|C^c \cap D^c| = 4n - X_p - (3n - X_p) - (n - X_p) = X_p$

For $p \in \{11, 12, 13, 14\}$, we have

$$\begin{aligned} T_\pi(p) &= \frac{\sum_{i=1}^{4n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{4n} c_i(p)^2}} \\ &= \frac{\sum_{i=1}^{4n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{12n}} \\ &= \frac{\sum_{C \cap D} (1) Y_{\pi(i)} + \sum_{C^c \cap D} (1) Y_{\pi(i)} + \sum_{C \cap D^c} (-3) Y_{\pi(i)} + \sum_{C^c \cap D^c} (-3) Y_{\pi(i)}}{\sqrt{12n}} \\ &= \frac{X_p \bar{Y}_{high} + (3n - X_p) \bar{Y}_{low} + (-3n + 3X_p) \bar{Y}_{high} - 3(X_p) \bar{Y}_{low}}{\sqrt{12n}} \end{aligned}$$

Thus,

$$\begin{aligned} T_\pi(p) &\sim N\left(\frac{X_p \mu_{high} + (3n - X_p)\, \mu_{low} + (-3n - 3X_p)\, \mu_{high} - (3X_p)\, \mu_{low}}{\sqrt{12n}}, \sigma^2 \right) \\ &= N\left(\frac{(4X_p - 3n)\, (\mu_{diff})}{\sqrt{12n}}, \sigma^2 \right) \end{aligned}$$

That is, $T_\pi(p) \sim N(k_n^\pi(p)(\mu_{diff}), \sigma^2)$ where $k_n^\pi(p) = \frac{4X_p - 3n}{\sqrt{12n}}$ depends on π only through X_p.

Thus, we have the following general result for all $p \in \{1, 2, ..., 14\}$:

$$T_\pi(p) \sim N(k_n^\pi(p)(\mu_{diff}), \sigma^2)$$

where $k_n^\pi(p)$ depends upon π only through X_p. Thus, we have proved Corollary 3. By Chebychev's inequality,

$$P\left(T_\pi(p) - k_n^\pi(p)(\mu_{diff}) \geq \frac{\sigma}{\delta} \right) \leq P\left(|T_\pi(p) - k_n^\pi(p)(\mu_{diff})| \geq \frac{\sigma}{\delta} \right) \leq \delta^2 \qquad (5.5)$$

Assume $(k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \geq \frac{2\sigma}{\delta}$. Then we have

$$\begin{aligned}
P(T(1) < T_{\pi}(p)) &= P(T_{\pi}(p) - T(1) \geq 0) \\
&= P(T_{\pi}(p) - T(1) + (k_n(1) - k_n^{\pi}(p))\mu_{diff} \geq (k_n(1) - k_n^{\pi}(p))\mu_{diff}) \\
&\leq P(|T_{\pi}(p) - T(1) + (k_n(1) - k_n^{\pi}(p))\mu_{diff}| \geq (k_n(1) - k_n^{\pi}(p))\mu_{diff}) \\
&\leq P\left(|T_{\pi}(p) - T(1) + (k_n(1) - k_n^{\pi}(p))\mu_{diff}| \geq \frac{2\sigma}{\delta}\right) \\
&\leq P(|T_{\pi}(p) - k_n^{\pi}(p)\mu_{diff}| + |T(1) - k_n(1)\mu_{diff}| \geq \frac{2\sigma}{\delta}) \\
&\qquad \text{by the triangle inequality.}
\end{aligned}$$

Note that $\{|T_{\pi}(p) - k_n^{\pi}(p)\mu_{diff}| + |T(1) - k_n(1)\mu_{diff}| \geq \frac{2\sigma}{\delta}\} \subseteq \{(|T_{\pi}(p) - k_n^{\pi}(p)\mu_{diff}| \geq \frac{\sigma}{\delta}) \cup (|T(1) - k_n(1)\mu_{diff}| \geq \frac{\sigma}{\delta})\}$. Thus,

$$\begin{aligned}
P(T(1) < T_{\pi}(p)) &\leq P(|T_{\pi}(p) - k_n^{\pi}(p)\mu_{diff}| + |T(1) - k_n(1)\mu_{diff}| \geq \frac{2\sigma}{\delta}) \\
&\leq P(|T_{\pi}(p) - k_n^{\pi}(p)\mu_{diff}| \geq \frac{\sigma}{\delta}) + P(|T(1) - k_n(1)\mu_{diff}| \geq \frac{\sigma}{\delta}) \\
&\leq 2\delta^2
\end{aligned}$$

Thus, by Boole's inequality we have for any fixed permutation π such that $(k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \geq \frac{2\sigma}{\delta}$, for $1 \leq p \leq 14$,

$$\begin{aligned}
P(T(1) < T_{\pi}) &= P((T(1) < T_{\pi}(1)) \cup (T(1) < T_{\pi}(2)) \cup ... \cup (T(1) < T_{\pi}(14))) \\
&\leq P(T(1) < T_{\pi}(1)) + P(T(1) < T_{\pi}(2)) + ... + P(T(1) < T_{\pi}(14)) \\
&\leq 2\delta^2 + 2\delta^2 + ... + 2\delta^2 \\
&= 28\delta^2
\end{aligned} \tag{5.6}$$

Hence, we have

$$P(T < T_\pi | T = T(1)) \leq \frac{P(T(1) < T_\pi)}{P(T = T(1))}$$
$$\leq \frac{28\delta^2}{1 - 13\delta^2} \text{ by (5.4) and (5.6)}$$

This gives us

$$P(T \geq T_\pi | T = T(1)) \geq 1 - \frac{28\delta^2}{1 - 13\delta^2}$$
$$= \frac{1 - 41\delta^2}{1 - 13\delta^2} \tag{5.7}$$
$$\geq 1 - 41\delta^2$$

Now we impose a discrete uniform distribution on all permutations. For all $p \in \{1, 2, ..., 14\}$, this imposes a hypergeometric distribution on X_p. This follows because we can think of X_p as the number of "successes" (that is, the number of phenotypes which were originally in cell 1 that remained in a "high" cell in configuration p after permutation) in K draws from a finite population of size $4n$ which contains n possible "successes". In particular, K is determined by the number of cells classified as "high" in the p configuration after permutation. For example, if $p = 1$, $K = n$, but if $p = 5$, $K = 2n$. Let A_p be a subset of the support of X_p, $Supp(X_p)$ such that $x \in A_p$ implies that $(k_n(1) - k_n^\pi(p))(\mu_{diff}) \geq \frac{2\sigma}{\delta}$.

Corollary 4: $P(X_p \in A_p) \geq 1 - \delta^2$ for all p.

Proof: Again, we will consider three cases.

Case 1: $p \in \{1, 2, 3, 4\}$

In this case, $K = n$. By known formulas for moments of a hypergeometric distribution, $E[X_p] = \frac{n}{4}$ and $Var[X_p] = \frac{9n^2}{16(4n-1)}$.

Thus, by Chebychev's Inequality we have

$$\begin{aligned}
P(X_p \notin A_p) &= P\left((k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \leq \frac{2\sigma}{\delta}\right) \\
&= P\left(\left(\frac{3n}{4}\right)^{1/2} - \frac{4X_p - n}{\sqrt{12n}} \leq \frac{2\sigma}{\delta\mu_{diff}}\right) \\
&= P\left(\frac{4X_p - n}{\sqrt{12n}} \geq -\frac{2\sigma}{\delta\mu_{diff}} + \left(\frac{3n}{4}\right)^{1/2}\right) \\
&\leq \left(\frac{9n^2}{12n(4n-1)}\right)\left(\frac{12n\mu_{diff}^2}{(-2\sigma\sqrt{12n} + 3n\delta\mu_{diff})^2}\right)\delta^2 \\
&\leq \delta^2 \text{ for sufficiently large } n
\end{aligned}$$

Thus, $P(X_p \in A_p) \geq 1 - \delta^2$ for sufficiently large n.

Case 2: $p \in \{5, 6, 7, 8, 9, 10\}$

In this case, $K = 2n$. By known formulas for moments of a hypergeometric distribution, $E[X_p] = \frac{n}{2}$ and $Var[X_p] = \frac{3n^2}{4(4n-1)}$. Thus, by Chebychev's Inequality we have

$$\begin{aligned}
P(X_p \notin A_p) &= P\left((k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \leq \frac{2\sigma}{\delta}\right) \\
&= P\left(\left(\frac{3n}{4}\right)^{1/2} - \frac{2X_p - n}{\sqrt{4n}} \leq \frac{2\sigma}{\delta\mu_{diff}}\right) \\
&= P\left(\frac{2X_p - n}{\sqrt{4n}} \geq -\frac{2\sigma}{\delta\mu_{diff}} + \left(\frac{3n}{4}\right)^{1/2}\right) \\
&\leq \left(\frac{3n^2}{4n(4n-1)}\right)\left(\frac{12n\mu_{diff}^2}{(-2\sigma\sqrt{12n} + 3n\delta\mu_{diff})^2}\right)\delta^2 \\
&\leq \delta^2 \text{ for sufficiently large } n
\end{aligned}$$

Thus, $P(X_p \in A_p) \geq 1 - \delta^2$ for sufficiently large n.

Case 3: $p \in \{11, 12, 13, 14\}$

In this case, $K = 3n$. By known formulas for moments of a hypergeometric distribution, $E[X_p] = \frac{3n}{4}$ and $Var[X_p] = \frac{9n^2}{16(4n-1)}$.
Thus, by Chebychev's Inequality we have

$$\begin{aligned}
P(X_p \notin A_p) &= P\left((k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \leq \frac{2\sigma}{\delta}\right) \\
&= P\left(\left(\frac{3n}{4}\right)^{1/2} - \frac{4X_p - 3n}{\sqrt{12n}} \leq \frac{2\sigma}{\delta\mu_{diff}}\right) \\
&= P\left(\frac{4X_p - 3n}{\sqrt{12n}} \geq -\frac{2\sigma}{\delta\mu_{diff}} + \left(\frac{3n}{4}\right)^{1/2}\right) \\
&\leq \left(\frac{9n^2}{12n(4n-1)}\right)\left(\frac{12n\mu_{diff}^2}{(-2\sigma\sqrt{12n} + 3n\delta\mu_{diff})^2}\right)\delta^2 \\
&\leq \delta^2 \text{ for sufficiently large } n
\end{aligned}$$

Thus, $P(X_p \in A_p) \geq 1 - \delta^2$ for sufficiently large n.

This yields the following general result: Let A_p be a subset of $Supp(X_p)$ such that $x \in A_p$ implies $(k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \geq \frac{2\sigma}{\delta}$. For sufficiently large n, we have $P(X_p \in A_p) \geq 1 - \delta^2$ for all $1 \leq p \leq 14$, and we have proved Corollary 4. By Boole's inequality we have the following:

$$\begin{aligned}
P\left(X_p \in A_p \forall p \in \{1, 2, ..., 14\}\right) &= P\left((X_1 \in A_1) \cap (X_2 \in A_2) \cap ... \cap (X_{14} \in A_{14})\right) \\
&= 1 - P\left((X_1 \notin A_1) \cup (X_2 \notin A_2) \cup ... \cup (X_{14} \notin A_{14})\right) \\
&\geq 1 - P\left(X_1 \notin A_1\right) - P\left(X_2 \notin A_2\right) ... - P\left(X_{14} \notin A_{14}\right) \\
&\geq 1 - 14\delta^2
\end{aligned} \tag{5.8}$$

By independence we have $P(X_p \in A_p \forall p | T = T(1)) = P(X_p \in A_p \forall p)$. Thus, by the

law of total probability we have

$$\begin{aligned}
P(T > T_\pi | T = T(1)) &= P(T > T_\pi | X_p \in A_p \forall p, T = T(1)) P(X_p \in A_p \forall p | T = T(1)) \\
&\quad + P(T > T_\pi | X_p \notin A_p \forall p, T = T(1)) P(X_p \notin A_p \forall p | T = T(1)) \\
&\geq P(T > T_\pi | X_p \in A_p \forall p, T = T(1)) P(X_p \in A_p \forall p | T = T(1)) \\
&= P(T > T_\pi | X_p \in A_p \forall p, T = T(1)) P(X_p \in A_p \forall p) \\
&\geq (1 - 41\delta^2)(1 - 14\delta^2) \text{ by (5.7) and (5.8)} \\
&= 1 - 55\delta^2 + (14)(41)\delta^4 \\
&\geq 1 - 55\delta^2
\end{aligned} \tag{5.9}$$

As δ can be chosen arbitrarily, this yields $P(T > T_\pi | T = T(1)) \to 1$ as $n \to \infty$. Thus, we have

$$\begin{aligned}
P(T > T_\pi) &= P(T > T_\pi | T = T(1)) P(T = T(1)) + P(T > T_\pi | T \neq T(1)) P(T \neq T(1)) \\
&\geq P(T > T_\pi | T = T(1)) P(T = T(1)) \\
&\geq (1 - 55\delta^2)(1 - 13\delta^2) \text{ by (5.4) and (5.9)} \\
&= 1 - 68\delta^2 + (55)(13)\delta^4 \\
&\geq 1 - 68\delta^2
\end{aligned} \tag{5.10}$$

As δ can be chosen arbitrarily, this yields $P(T > T_\pi) \to 1$ as $n \to \infty$. Thus, for 1000 randomly selected permutations, $\{\pi_1, \pi_2 ..., \pi_{1000}\}$, the probability that $T > T_{\pi_i}$ for at

least 951 of the π_i's (i.e., that the permutation p-value is less than .05) is

$$\begin{aligned} P(T > T_{\pi_i} \text{ for at least 951 of the } \pi_i\text{'s}) &\geq P(T > T_{\pi_i} \text{ for 1000 of the } \pi_i\text{'s}) \\ &\geq (1 - 68\delta^2)^{1000} \text{ by (5.10)} \\ &= 1 - \epsilon \end{aligned}$$

where for every $\epsilon > 0$, δ can be chosen such that

$$\delta = \sqrt{\frac{1 - (1-\epsilon)^{\frac{1}{1000}}}{68}}$$

Thus, the probability that the permutation p-value is less than .05 must also tend to 1 as $n \to \infty$. Returning to the beginning of the proof, we have shown that $P(A) \to 1$ as $n \to \infty$, implying that $P(A^c)$ tends to 0 as n tends to infinity. Thus, by the expression in (5.1), we have shown $P(A \cap B) \to 1$ as $n \to \infty$. Note that this result still holds for any arbitrary value chosen for c, instead of $c = .05$.

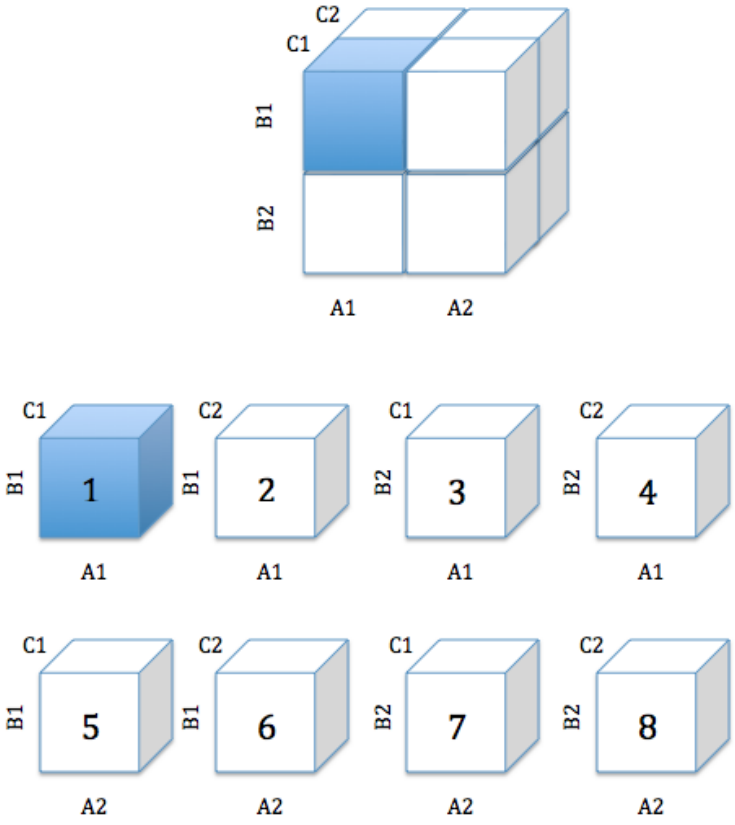
5.2 Analogous Theoretical Results for Three-way Interactions

Consider the three-way interaction among three SNPs: SNP_A, SNP_B and SNP_C. As in the previous section, suppose each of these SNPs carry two states, condition 1 and condition 2. For illustration, assign labels to each of the SNP states as follows:

- A1 := SNP_A, condition 1
- A2 := SNP_A, condition 2
- B1 := SNP_B, condition 1
- B2 := SNP_B, condition 2
- C1 := SNP_C, condition 1
- C2 := SNP_C, condition 2

This three-way interaction is represented by the cube at the top of figure 5.4. Each multifactor class (cell) is represented by one of the eight smaller cubes, illustrated at the bottom of the figure. We assign labels to these multifactor classes (1-8) as shown in figure 5.4. Suppose $i = 1, 2, ..., n$ correspond to subjects in cell 1, $i =$

Figure 5.4: Illustration of the three-way interaction between SNP_A, SNP_B, SNP_C



$n + 1, n + 2, ..., 2n$ correspond to subjects in cell 2, and so forth. Let $Y_1, ..., Y_{8n}$ represent the values of some quantitative phenotype for each subject, $i = 1, 2, ..., 8n$. Let $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu_{high}, \sigma^2)$ and $Y_{n+1}, ..., Y_{8n} \overset{iid}{\sim} N(\mu_{low}, \sigma^2)$, where $\mu_{high} > \mu_{low}$.

Let p_1 be the high/low classification from the AQMDR method of a three-way interaction in which cell 1 is correctly identified as "high" and the seven remaining cells are classified as "low". See figure 5.5, in which the shaded cell is the one categorized

as "high", and the light cells are categorized as "low". In what follows, we assume that AQMDR is applied to three-way interactions in which there are eight multifactor classes, rather than the 27 multifactor classes we observed for each three-way interaction in previous chapters.

Figure 5.5: The p_1 "high"/"low" configuration.

Once again, our goal is to prove that the probability that the gene-to-gene interaction of SNP_A, SNP_B and SNP_C is included in the AQMDR Arbitrary Cutoff aggregated score (ACAS with fixed $c > 0$) and that the p_1 high/low configuration of this interaction is correctly assigned in AQMDR tends to 1 as n goes to infinity. We will do so with an argument that is analogous to that of the previous section. We have the following events:

1. Event A: the event that the three-way gene-to-gene interaction of SNP_A, SNP_B and SNP_C is included in the AQMDR arbitrary cutoff aggregated score. That is, the interaction receives a weight of 1.

2. Event B: the event that the p_1 high/low configuration of this interaction is correctly assigned in AQMDR.

We want to show that $P(A \cap B) \to 1$ as $n \to \infty$.

The proof proceeds as follows:

Without loss of generality, let $c = .05$. First note that $P(B) = P((\bar{Y}_1 > \bar{\bar{Y}}) \cap (\bar{Y}_2 \leq \bar{\bar{Y}}) \cap (\bar{Y}_3 \leq \bar{\bar{Y}}) \cap (\bar{Y}_4 \leq \bar{\bar{Y}}) \cap (\bar{Y}_5 \leq \bar{\bar{Y}}) \cap (\bar{Y}_6 \leq \bar{\bar{Y}}) \cap (\bar{Y}_7 \leq \bar{\bar{Y}}) \cap (\bar{Y}_8 \leq \bar{\bar{Y}}))$, where $\bar{\bar{Y}}$ is the overall mean of $Y_1, Y_2, ..., Y_{8n}$, and $\bar{Y}_j$ for $j \in \{1, ..., 8\}$ is the mean within cell j. That is, $\bar{Y}_1$ is the mean of $Y_1, Y_2, ..., Y_n$, $\bar{Y}_2$ is the mean of $Y_{n+1}, Y_{n+2}, ..., Y_{2n}$, etc. By DeMorgan's Law and Boole's inequality, we have

$$\begin{aligned}
P(A \cap B) &= 1 - P(A^c \cup B^c) \\
&\geq 1 - P(A^c) - P(B^c) \\
&= 1 - P(A^c) - P((\bar{Y}_1 \leq \bar{\bar{Y}}) \cup (\bar{Y}_2 > \bar{\bar{Y}}) \cup (\bar{Y}_3 > \bar{\bar{Y}}) \cup (\bar{Y}_4 > \bar{\bar{Y}}) \cup (\bar{Y}_5 > \bar{\bar{Y}}) \\
&\cup (\bar{Y}_6 > \bar{\bar{Y}}) \cup (\bar{Y}_7 > \bar{\bar{Y}}) \cup (\bar{Y}_8 > \bar{\bar{Y}})) \\
&\geq 1 - P(A^c) - (P(\bar{Y}_1 \leq \bar{\bar{Y}}) + P(\bar{Y}_2 > \bar{\bar{Y}}) + P(\bar{Y}_3 > \bar{\bar{Y}}) + P(\bar{Y}_4 > \bar{\bar{Y}}) \\
&+ (\bar{Y}_5 > \bar{\bar{Y}}) + (\bar{Y}_6 > \bar{\bar{Y}}) + (\bar{Y}_7 > \bar{\bar{Y}}) + (\bar{Y}_8 > \bar{\bar{Y}}))
\end{aligned} \tag{5.11}$$

By the Weak Law of Large Numbers and Slutsky's Theorem, we have the following:

1. $\bar{Y}_1 \xrightarrow{p} \mu_{high}$, as $n \to \infty$.
2. $\bar{Y}_2, \bar{Y}_3, \bar{Y}_4, \bar{Y}_5, \bar{Y}_6, \bar{Y}_7, \bar{Y}_8 \xrightarrow{p} \mu_{low}$, as $n \to \infty$.
3. $\bar{\bar{Y}} = \frac{n(\bar{Y}_1+\bar{Y}_2+\bar{Y}_3+\bar{Y}_4+\bar{Y}_5+\bar{Y}_6+\bar{Y}_7+\bar{Y}_8)}{8n} = \frac{\bar{Y}_1+\bar{Y}_2+\bar{Y}_3+\bar{Y}_4+\bar{Y}_5+\bar{Y}_6+\bar{Y}_7+\bar{Y}_8}{8} \xrightarrow{p} \frac{\mu_{high}+7\mu_{low}}{8} := \mu$, as $n \to \infty$.
4. $\mu < \mu_{high}$ and $\mu > \mu_{low}$.

Thus, as $n \to \infty$, $P(\bar{Y}_1 \leq \bar{\bar{Y}}) + P(\bar{Y}_2 > \bar{\bar{Y}}) + P(\bar{Y}_3 > \bar{\bar{Y}}) + P(\bar{Y}_4 > \bar{\bar{Y}}) + P(\bar{Y}_5 > \bar{\bar{Y}}) + P(\bar{Y}_6 > \bar{\bar{Y}}) + P(\bar{Y}_7 > \bar{\bar{Y}}) + P(\bar{Y}_8 > \bar{\bar{Y}}) \to 0$.

This leaves only the consideration of $P(A^c)$. Recall that in AQMDR with ACAS, the inclusion of a gene-to-gene interaction in the aggregated score is based on a p-value

determined from permutation testing (see chapter 2). When $c = .05$, an interaction is assigned a weight of 1 in the aggregated score if the permutation p-value associated with that interaction is less than .05. Our goal is to show that, with probability approaching 1 as $n \to \infty$, the p-value achieved by permutation for the T-statistic corresponding to this interaction is less than .05. That is, the probability that the interaction is included in the aggregated score tends to 1 as $n \to \infty$. By establishing this, we can then determine that $P(A^c) \to 0$ as $n \to \infty$. We will prove this with a series of corollaries. Then, by the expression in (5.11), we will have shown that $P(A \cap B) \to 1$ as $n \to \infty$.

First note that (disregarding 0 probability events) there are 254 possible "high"/"low" configurations (labeling of cells as "high" or "low") which can be assigned to a three-way interaction with AQMDR. Denote these configurations as $p = 1, ..., 254$. All possible configurations can be categorized into one of 14 configuration cases. Let P_j represent the group of p's in configuration case j for $j \in \{1, ..., 14\}$. These configuration cases are described in the following table.

Table 5.4: Description of configuration cases.

j	**Case Description**
1	One cell is labeled "high" and that cell is cell 1.
2	Two cells are labeled "high", one of which is cell 1.
3	Three cells are labeled "high", one of which is cell 1.
4	Four cells are labeled "high", one of which is cell 1.
5	Five cells are labeled "high", one of which is cell 1.
6	Six cells are labeled "high", one of which is cell 1.
7	Seven cells are labeled "high", one of which is cell 1.
8	One cell is labeled "high" and that cell is not cell 1.
9	Two cells are labeled "high", neither of which are cell 1.
10	Three cells are labeled "high", none of which are cell 1.
11	Four cells are labeled "high", none of which are cell 1.
12	Five cells are labeled "high", none of which are cell 1.
13	Six cells are labeled "high", none of which are cell 1.
14	Seven cells are labeled "high", none of which are cell 1.

We will define $c_i(p)$ for each $p \in \{1, 2, ..., 254\}$ such that $\sum_{i=1}^{8n} c_i(p) = 0$. The $c_i(p)$ definitions for representative examples of p's within each configuration case P_j for $j \in \{1, 2, ..., 14\}$ are shown in table 5.5.

Table 5.5: $c_i(p)$ for representative examples of p's within each configuration case P_j for $j \in \{1, 2, ..., 14\}$.

j	cell 1	cell 2	cell 3	cell 4	cell 5	cell 6	cell 7	cell 8
1	7	-1	-1	-1	-1	-1	-1	-1
2	3	3	-1	-1	-1	-1	-1	-1
3	5	5	5	-3	-3	-3	-3	-3
4	1	1	1	1	-1	-1	-1	-1
5	3	3	3	3	3	-5	-5	-5
6	1	1	1	1	1	1	-3	-3
7	1	1	1	1	1	1	1	-7
8	-1	7	-1	-1	-1	-1	-1	-1
9	-1	3	3	-1	-1	-1	-1	-1
10	-3	5	5	5	-3	-3	-3	-3
11	-1	1	1	1	1	-1	-1	-1
12	-5	3	3	3	3	3	-5	-5
13	-3	1	1	1	1	1	1	-3
14	-7	1	1	1	1	1	1	1

Define $T(p)$ for each $p = 1, ..., 254$ as follows:

$$T(p) := \frac{\sum_{i=1}^{8n} c_i(p) Y_i}{\sqrt{\sum_{i=1}^{8n} c_i(p)^2}}.$$

Corollary 1: $T(p) \sim N(k_n(p)(\mu_{high} - \mu_{low}), \sigma^2)$, for a constant $k_n(p)$.

Proof: There are 14 cases we will consider for the distribution of $T(p)$ (one for each configuration case P_j).

Case 1: $p \in P_1$

There is only one configuration in P_1, so let's call this configuration $p = 1$.

$$\begin{aligned} T(1) =& \frac{\sum_{i=1}^{8n} c_i(1) Y_i}{\sqrt{56n}} \\ =& \frac{\sum_{i=1}^{n} (7) Y_i + \sum_{i=n+1}^{8n} (-1) Y_i}{\sqrt{56n}} \\ =& \frac{7n\bar{Y}_{high} - 7n\bar{Y}_{low}}{\sqrt{56n}} \\ =& \frac{7n(\bar{Y}_{high} - \bar{Y}_{low})}{\sqrt{56n}} \end{aligned}$$

where $\bar{Y}_{high}$ is a sample mean of realized values of Y_i from a cell with mean μ_{high} (cell 1), and $\bar{Y}_{low}$ is a sample mean of realized values of Y_i from cells with mean μ_{low} (cells 2, 3,...,8). As $\sum_{i=1}^{8n} c_i(1) = 0$,

$$T(1) \sim N\left(k_n(1)(\mu_{high} - \mu_{low}), \sigma^2\right)$$

where $k_n(1) = \frac{7n}{\sqrt{56n}} = \left(\frac{7n}{8}\right)^{1/2}$.

Case 2: $p \in P_2$.

$$\begin{aligned} T(p) =& \frac{\sum_{i=1}^{8n} c_i(p) Y_i}{\sqrt{24n}} \\ =& \frac{\sum_{i=1}^{n} (3) Y_i + \sum_{i=n+1}^{2n} (3) Y_i + \sum_{i=2n+1}^{8n} (-1) Y_i}{\sqrt{24n}} \\ =& \frac{3n\bar{Y}_{high} + 3n\bar{Y}_{low} - 6n\bar{Y}_{low}}{\sqrt{24n}} \end{aligned}$$

As $\sum_{i=1}^{8n} c_i(2) = 0$,

$$T(p) \sim N\left(\frac{(3n\mu_{high} + 3n\mu_{low} - 6n\mu_{low})}{\sqrt{24n}}, \sigma^2\right) = N\left(k_n(p)\left(\mu_{high} - \mu_{low}\right), \sigma^2\right)$$

where $k_n(p) = \frac{3n}{\sqrt{24n}} = \left(\frac{3n}{8}\right)^{1/2}$ for $p \in P_2$.

The 12 remaining cases can be considered using similar calculations. In summary, $T(p) \sim N(k_n(p)(\mu_{high} - \mu_{low}), \sigma^2)$, for a constant $k_n(p)$. Thus, we have proved Corollary 1. The values of $k_n(p)$ for each configuration case are displayed in table 5.6. Note that $k_n(1) = max_{1 \le p \le 254} k_n(p)$. Now, define a T-statistic as follows:

Table 5.6: $k_n(p)$

p	$k_n(p)$	p	$k_n(p)$
$p = 1$	$\left(\frac{7n}{8}\right)^{1/2}$	$p \in P_8$	$-\left(\frac{n}{56}\right)^{1/2}$
$p \in P_2$	$\left(\frac{3n}{8}\right)^{1/2}$	$p \in P_9$	$-\left(\frac{n}{24}\right)^{1/2}$
$p \in P_3$	$\left(\frac{5n}{24}\right)^{1/2}$	$p \in P_{10}$	$-\left(\frac{3n}{40}\right)^{1/2}$
$p \in P_4$	$\left(\frac{n}{8}\right)^{1/2}$	$p \in P_{11}$	$-\left(\frac{n}{8}\right)^{1/2}$
$p \in P_5$	$\left(\frac{3n}{40}\right)^{1/2}$	$p \in P_{12}$	$-\left(\frac{5n}{24}\right)^{1/2}$
$p \in P_6$	$\left(\frac{n}{24}\right)^{1/2}$	$p \in P_{13}$	$-\left(\frac{3n}{8}\right)^{1/2}$
$p \in P_7$	$\left(\frac{n}{56}\right)^{1/2}$	$p \in P_{14}$	$-\left(\frac{7n}{8}\right)^{1/2}$

$$T := max_{1 \le p \le 254} T(p).$$

This T-statistic is not quite identical to the one used in AQMDR (see equation 2.1) because sample variances are not incorporated. We forgo the estimation of the variances because we assume that $Y_1, ..., Y_{8n}$ all have the same underlying variance. Let δ be an arbitrary small positive number. Assume n is large enough such that

$$(k_n(1) - max_{2\leq p\leq 254} k_n(p))(\mu_{high} - \mu_{low}) \geq \frac{2\sigma}{\delta} \tag{5.12}$$

That is, $[(\frac{7n}{8})^{1/2} - (\frac{3n}{8})^{1/2}](\mu_{high} - \mu_{low}) \geq \frac{2\sigma}{\delta}$.

Corollary 2: $P(T = T(1)) \geq 1 - 253\delta^2$.

Proof: Put $\mu_{diff} := \mu_{high} - \mu_{low}$. Note that $E[T(p) - T(1)] = (k_n(p) - k_n(1))\mu_{diff}$ and $Var[T(p) - T(1)] \leq 4\sigma^2$. By Chebychev's Inequality, for all $p \in \{1, 2, ..., 254\}$ we have

$$P\left(|T(p) - T(1) + (k_n(1) - k_n(p))\mu_{diff}| \geq \frac{2\sigma}{\delta}\right) \leq \delta^2 \tag{5.13}$$

For any $p \in \{2, 3, ..., 254\}$,

$$\begin{aligned}
P(T(p) > T(1)) &= P(T(p) - T(1) \geq 0) \\
&= P(T(p) - T(1) + (k_n(1) - k_n(p))\,\mu_{diff} \geq (k_n(1) - k_n(p))\,\mu_{diff}) \\
&\leq P\left(T(p) - T(1) + (k_n(1) - k_n(p))\,\mu_{diff} \geq \frac{2\sigma}{\delta}\right) \text{ by (5.12)}, \\
&\leq P\left(|T(p) - T(1) + (k_n(1) - k_n(p))\,\mu_{diff}| \geq \frac{2\sigma}{\delta}\right) \\
&\leq \delta^2 \text{ by (5.13)}.
\end{aligned}$$

Now, consider $P(T \neq T(1))$. By Boole's inequality, we have

$$\begin{aligned} P\left(T \neq T(1)\right) &= P\left(T(1) \neq max_{1 \leq p \leq 254} T(p)\right) \\ &= P\left((T(2) > T(1)) \cup (T(3) > T(1)) \cup ... \cup (T(254) > T(1))\right) \\ &\leq P\left(T(2) > T(1)\right) + P(T(3) > T(1)) + ... + P(T(254) > T(1)) \\ &\leq 253\delta^2 \end{aligned}$$

Hence,

$$P\left(T = T(1)\right) \geq 1 - 253\delta^2 \tag{5.14}$$

and we have proved Corollary 2.

Consider a fixed permutation $\pi : \{1, 2, ..., 8n\} \rightarrow \{1, 2, ..., 8n\}$. Define $T_\pi(p)$ as follows:

$$T_\pi(p) := \frac{\sum_{i=1}^{8n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{8n} c_{\pi(i)}(p)^2}} = \frac{\sum_{i=1}^{8n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{8n} c_i(p)^2}}.$$

Define the test statistic for permutation π as follows:

$$T_\pi := max_{1 \leq p \leq 254} T_\pi(p).$$

Corollary 3: For all p:

$$T_\pi(p) \sim N(k_n^\pi(p)(\mu_{diff}), \sigma^2)$$

where $k_n^\pi(p)$ depends upon π only through X_p, where X_p is the number of phenotypes which were originally in a cell 1 that remained in a "high" cell after permutation..

Proof: Let $C = \{1, 2, ..., n\}$. We will consider 7 cases for the distribution of $T_\pi(p)$.

Case 1: Cases in which one cell is labeled "high", including $p = 1$ and $p \in P_8$.

We define D to be the set of $\pi(i)$'s corresponding to the cell categorized as "high" in the p configuration. For example, if $p = 1$, $D = \{\pi(1), ..., \pi(n)\}$. Let $X_p = |C \cap D|$. That is, X_p is the number of phenotypes which were originally in cell 1, that remained in a cell classified as "high" after permutation. For all $p \in (\{1\} \cup P_8)$, we have the following:

1. $|C \cap D^c| = n - X_p$
2. $|C^c \cap D| = n - X_p$
3. $|C^c \cap D^c| = 8n - X_p - (n - X_p) - (n - X_p) = 6n + X_p$

Consider $T_\pi(p)$:

$$\begin{aligned} T_\pi(p) &= \frac{\sum_{i=1}^{8n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{8n} c_i(p)^2}} \\ &= \frac{\sum_{i=1}^{8n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{56n}} \\ &= \frac{\sum_{C \cap D} (7) Y_{\pi(i)} + \sum_{C^c \cap D} (7) Y_{\pi(i)} + \sum_{C \cap D^c} (-1) Y_{\pi(i)} + \sum_{C^c \cap D^c} (-1) Y_{\pi(i)}}{\sqrt{56n}} \\ &= \frac{7X_p \bar{Y}_{high} + 7(n - X_p)\bar{Y}_{low} - (n - X_p)\bar{Y}_{high} - (6n + X_p)\bar{Y}_{low}}{\sqrt{56n}} \end{aligned}$$

Thus,

$$\begin{aligned} T_\pi(p) &\sim N\left(\frac{7X_p\mu_{high} + (7n - 7X_p)\,\mu_{low} - (n - X_p)\,\mu_{high} - (6n + X_p)\,\mu_{low}}{\sqrt{56n}}, \sigma^2\right) \\ &= N\left(\frac{(8X_p - n)\,(\mu_{diff})}{\sqrt{56n}}, \sigma^2\right) \end{aligned}$$

That is, $T_\pi(p) \sim N(k_n^\pi(p)(\mu_{diff}), \sigma^2)$ where $k_n^\pi(p) = \frac{8X_p - n}{\sqrt{56n}}$ depends on π only through

X_p.

Case 2: Cases in which two cells are labeled "high", including $p \in (P_2 \cup P_9)$.

We define D to be the set of $\pi(i)$'s corresponding to the cell categorized as "high" in the p configuration. Let $X_p = |C \cap D|$. We have the following:

1. $|C \cap D^c| = n - X_p$
2. $|C^c \cap D| = 2n - X_p$
3. $|C^c \cap D^c| = 8n - X_p - (2n - X_p) - (n - X_p) = 5n + X_p$

For $p \in (P_2 \cup P_9)$, we have

$$\begin{aligned}
T_\pi(p) &= \frac{\sum_{i=1}^{8n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{8n} c_i(p)^2}} \\
&= \frac{\sum_{i=1}^{8n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{24n}} \\
&= \frac{\sum_{C \cap D} (3) Y_{\pi(i)} + \sum_{C^c \cap D} (3) Y_{\pi(i)} + \sum_{C \cap D^c} (-1) Y_{\pi(i)} + \sum_{C^c \cap D^c} (-1) Y_{\pi(i)}}{\sqrt{24n}} \\
&= \frac{3X_p \bar{Y}_{high} + 3(2n - X_p)\bar{Y}_{low} - (n - X_p)\bar{Y}_{high} - (5n + X_p)\bar{Y}_{low}}{\sqrt{24n}}
\end{aligned}$$

Thus,

$$\begin{aligned}
T_\pi(p) &\sim N\left(\frac{3X_p \mu_{high} + (6n - 3X_p)\,\mu_{low} - (n - X_p)\,\mu_{high} - (5n + X_p)\,\mu_{low}}{\sqrt{24n}}, \sigma^2 \right) \\
&= N\left(\frac{(4X_p - n)\,(\mu_{diff})}{\sqrt{24n}}, \sigma^2 \right)
\end{aligned}$$

That is, $T_\pi(p) \sim N(k_n^\pi(p)(\mu_{diff}), \sigma^2)$ where $k_n^\pi(p) = \frac{4X_p - n}{\sqrt{24n}}$ depends on π only through X_p.

The remaining cases can be considered using similar calculations. We have the following general result for all $p \in \{1, 2, ..., 254\}$:

$$T_\pi(p) \sim N(k_n^\pi(p)(\mu_{diff}), \sigma^2)$$

where $k_n^\pi(p)$ depends upon π only through X_p. Thus we have proved Corollary 3. The $k_n^\pi(p)$'s for each of the 7 cases are displayed in the following table.

Table 5.7: $k_n^\pi(p)$

p	$k_n^\pi(p)$
$p \in (\{1\} \cup P_8)$	$\frac{8X_p - n}{\sqrt{56n}}$
$p \in (P_2 \cup P_9)$	$\frac{4X_p - n}{\sqrt{24n}}$
$p \in (P_3 \cup P_{10})$	$\frac{8X_p - 3n}{\sqrt{120n}}$
$p \in (P_4 \cup P_{11})$	$\frac{2X_p - n}{\sqrt{8n}}$
$p \in (P_5 \cup P_{12})$	$\frac{8X_p - 5n}{\sqrt{120n}}$
$p \in (P_6 \cup P_{13})$	$\frac{4X_p - 3n}{\sqrt{24n}}$
$p \in (P_7 \cup P_{14})$	$\frac{8X_p - 7n}{\sqrt{56n}}$

By Chebychev's inequality,

$$P\left(T_\pi(p) - k_n^\pi(p)\mu_{diff} \geq \frac{\sigma}{\delta}\right) \leq P\left(|T_\pi(p) - k_n^\pi(p)\mu_{diff}| \geq \frac{\sigma}{\delta}\right) \leq \delta^2 \qquad (5.15)$$

Assume $(k_n(1) - k_n^\pi(p))\,\mu_{diff} \geq \frac{2\sigma}{\delta}$. Then we have

$$\begin{aligned}
P\left(T(1) < T_\pi(p)\right) &= P\left(T_\pi(p) - T(1) \geq 0\right) \\
&= P\left(T_\pi(p) - T(1) + (k_n(1) - k_n^\pi(p))\,\mu_{diff} \geq (k_n(1) - k_n^\pi(p))\,\mu_{diff}\right) \\
&\leq P\left(|T_\pi(p) - T(1) + (k_n(1) - k_n^\pi(p))\,\mu_{diff}| \geq (k_n(1) - k_n^\pi(p))\,\mu_{diff}\right) \\
&\leq P\left(|T_\pi(p) - T(1) + (k_n(1) - k_n^\pi(p))\,\mu_{diff}| \geq \frac{2\sigma}{\delta}\right) \\
&\leq P(|T_\pi(p) - k_n^\pi(p)\mu_{diff}| + |T(1) - k_n(1)\mu_{diff}| \geq \frac{2\sigma}{\delta}) \\
&\qquad\qquad \text{by the triangle inequality.}
\end{aligned}$$

Note that $\{|T_\pi(p) - k_n^\pi(p)\mu_{diff}| + |T(1) - k_n(1)\mu_{diff}| \geq \frac{2\sigma}{\delta}\} \subseteq \{(|T_\pi(p) - k_n^\pi(p)\mu_{diff}| \geq \frac{\sigma}{\delta}) \cup (|T(1) - k_n(1)\mu_{diff}| \geq \frac{\sigma}{\delta})\}$. Thus,

$$\begin{aligned}
P\left(T(1) < T_\pi(p)\right) &\leq P(|T_\pi(p) - k_n^\pi(p)\mu_{diff}| + |T(1) - k_n(1)\mu_{diff}| \geq \frac{2\sigma}{\delta}) \\
&\leq P(|T_\pi(p) - k_n^\pi(p)\mu_{diff}| \geq \frac{\sigma}{\delta}) + P(|T(1) - k_n(1)\mu_{diff}| \geq \frac{\sigma}{\delta}) \\
&\leq 2\delta^2
\end{aligned}$$

Thus, by Boole's inequality we have for any fixed permutation π such that $(k_n(1) - k_n^\pi(p))(\mu_{diff}) \geq \frac{2\sigma}{\delta}$, for $1 \leq p \leq 254$,

$$\begin{aligned}
P\left(T(1) < T_\pi\right) &= P\left((T(1) < T_\pi(1)) \cup (T(1) < T_\pi(2)) \cup ... \cup (T(1) < T_\pi(254))\right) \\
&\leq P\left(T(1) < T_\pi(1)\right) + P\left(T(1) < T_\pi(2)\right) + ... + P\left(T(1) < T_\pi(254)\right) \\
&\leq 2\delta^2 + 2\delta^2 + ... + 2\delta^2 \\
&= 508\delta^2.
\end{aligned} \tag{5.16}$$

Hence, we have

$$P(T < T_\pi | T = T(1)) \leq \frac{P(T(1) < T_\pi)}{P(T = T(1))}$$
$$\leq \frac{508\delta^2}{1 - 253\delta^2} \text{ by (5.14) and (5.16).}$$

For any fixed permutation π such that $(k_n(1) - k_n^\pi(p))(\mu_{diff}) \geq \frac{2\sigma}{\delta}$, for $1 \leq p \leq 254$,

$$P(T \geq T_\pi | T = T(1)) \geq 1 - \frac{508\delta^2}{1 - 253\delta^2}$$
$$= \frac{1 - 761\delta^2}{1 - 253\delta^2} \tag{5.17}$$
$$\geq 1 - 761\delta^2.$$

Now we impose a discrete uniform distribution on all permutations. For all $p \in \{1, 2, ..., 254\}$, this imposes a hypergeometric distribution on X_p. This follows because we can think of X_p as the number of "successes" (that is, the number of phenotypes which were originally in cell 1 that remained in a "high" cell in configuration p after permutation) in K draws from a finite population of size $8n$ which contains n possible "successes". In particular, K is determined by the number of cells classified as "high" in the p configuration after permutation. For example, if $p = 1$, $K = n$, but if $p \in P_2$, $K = 2n$. Let A_p be a subset of the support of X_p, $Supp(X_p)$ such that $x \in A_p$ implies that $(k_n(1) - k_n^\pi(p))(\mu_{diff}) \geq \frac{2\sigma}{\delta}$.

Corollary 4: $P(X_p \notin A_p) \geq 1 - \delta^2$ for all p.

Proof: Again, we will consider seven cases.

Case 1: $p \in (\{1\} \cup P_8)$

In this case, $K = n$. By known formulas for moments of a hypergeometric distribution, $E[X_p] = \frac{n}{8}$ and $Var[X_p] = \frac{49n^2}{64(8n-1)}$.

Thus, by Chebychev's Inequality we have

$$\begin{aligned}
P(X_p \notin A_p) &= P\left((k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \leq \frac{2\sigma}{\delta}\right) \\
&= P\left(\left(\frac{7n}{8}\right)^{1/2} - \frac{8X_p - n}{\sqrt{56n}} \leq \frac{2\sigma}{\delta\mu_{diff}}\right) \\
&= P\left(\frac{8X_p - n}{\sqrt{56n}} \geq -\frac{2\sigma}{\delta\mu_{diff}} + \left(\frac{7n}{8}\right)^{1/2}\right) \\
&\leq \left(\frac{49n^2}{56n(8n-1)}\right)\left(\frac{56n\mu_{diff}^2}{(-2\sigma\sqrt{56n} + 7n\delta\mu_{diff})^2}\right)\delta^2 \\
&\leq \delta^2 \text{ for sufficiently large } n
\end{aligned}$$

Thus, $P(X_p \in A_p) \geq 1 - \delta^2$ for sufficiently large n.

Case 2: $p \in (P_2 \cup P_9)$

In this case, $K = 2n$. By known formulas for moments of a hypergeometric distribution, $E[X_p] = \frac{n}{4}$ and $Var[X_p] = \frac{21n^2}{16(8n-1)}$. Thus, by Chebychev's Inequality we have

$$\begin{aligned}
P(X_p \notin A_p) &= P\left((k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \leq \frac{2\sigma}{\delta}\right) \\
&= P\left(\left(\frac{7n}{8}\right)^{1/2} - \frac{4X_p - n}{\sqrt{24n}} \leq \frac{2\sigma}{\delta\mu_{diff}}\right) \\
&= P\left(\frac{4X_p - n}{\sqrt{24n}} \geq -\frac{2\sigma}{\delta\mu_{diff}} + \left(\frac{7n}{8}\right)^{1/2}\right) \\
&\leq \left(\frac{21n^2}{24n(8n-1)}\right)\left(\frac{56n\mu_{diff}^2}{(-2\sigma\sqrt{56n} + 7n\delta\mu_{diff})^2}\right)\delta^2 \\
&\leq \delta^2 \text{ for sufficiently large } n
\end{aligned}$$

Thus, $P(X_p \in A_p) \geq 1 - \delta^2$ for sufficiently large n.

Case 3: $p \in (P_3 \cup P_{10})$

In this case, $K = 3n$. By known formulas for moments of a hypergeometric distribution, $E[X_p] = \frac{3n}{8}$ and $Var[X_p] = \frac{105n^2}{64(8n-1)}$. Thus, by Chebychev's Inequality we have

$$\begin{aligned}
P(X_p \notin A_p) &= P\left((k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \leq \frac{2\sigma}{\delta}\right) \\
&= P\left(\left(\frac{7n}{8}\right)^{1/2} - \frac{8X_p - 3n}{\sqrt{120n}} \leq \frac{2\sigma}{\delta\mu_{diff}}\right) \\
&= P\left(\frac{8X_p - 3n}{\sqrt{120n}} \geq -\frac{2\sigma}{\delta\mu_{diff}} + \left(\frac{7n}{8}\right)^{1/2}\right) \\
&\leq \left(\frac{105n^2}{120n(8n-1)}\right)\left(\frac{56n\mu_{diff}^2}{(-2\sigma\sqrt{56n} + 7n\delta\mu_{diff})^2}\right)\delta^2 \\
&\leq \delta^2 \text{ for sufficiently large } n
\end{aligned}$$

Thus, $P(X_p \in A_p) \geq 1 - \delta^2$ for sufficiently large n.

Case 4: $p \in (P_4 \cup P_{11})$

In this case, $K = 4n$. By known formulas for moments of a hypergeometric distribution, $E[X_p] = \frac{n}{2}$ and $Var[X_p] = \frac{7n^2}{4(8n-1)}$. Thus, by Chebychev's Inequality we

have

$$\begin{aligned}
P(X_p \notin A_p) &= P\left((k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \leq \frac{2\sigma}{\delta}\right) \\
&= P\left(\left(\frac{7n}{8}\right)^{1/2} - \frac{2X_p - n}{\sqrt{8n}} \leq \frac{2\sigma}{\delta\mu_{diff}}\right) \\
&= P\left(\frac{2X_p - n}{\sqrt{8n}} \geq -\frac{2\sigma}{\delta\mu_{diff}} + \left(\frac{7n}{8}\right)^{1/2}\right) \\
&\leq \left(\frac{7n^2}{8n(8n-1)}\right)\left(\frac{56n\mu_{diff}^2}{(-2\sigma\sqrt{56n} + 7n\delta\mu_{diff})^2}\right)\delta^2 \\
&\leq \delta^2 \text{ for sufficiently large } n
\end{aligned}$$

Thus, $P(X_p \in A_p) \geq 1 - \delta^2$ for sufficiently large n.

Case 5: $p \in (P_5 \cup P_{12})$

In this case, $K = 5n$. By known formulas for moments of a hypergeometric distribution, $E[X_p] = \frac{5n}{8}$ and $Var[X_p] = \frac{105n^2}{64(8n-1)}$. Thus, by Chebychev's Inequality we have

$$\begin{aligned}
P(X_p \notin A_p) &= P\left((k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \leq \frac{2\sigma}{\delta}\right) \\
&= P\left(\left(\frac{7n}{8}\right)^{1/2} - \frac{8X_p - 5n}{\sqrt{120n}} \leq \frac{2\sigma}{\delta\mu_{diff}}\right) \\
&= P\left(\frac{8X_p - 5n}{\sqrt{120n}} \geq -\frac{2\sigma}{\delta\mu_{diff}} + \left(\frac{7n}{8}\right)^{1/2}\right) \\
&\leq \left(\frac{105n^2}{120n(8n-1)}\right)\left(\frac{56n\mu_{diff}^2}{(-2\sigma\sqrt{56n} + 7n\delta\mu_{diff})^2}\right)\delta^2 \\
&\leq \delta^2 \text{ for sufficiently large } n
\end{aligned}$$

Thus, $P(X_p \in A_p) \geq 1 - \delta^2$ for sufficiently large n.

Case 6: $p \in (P_6 \cup P_{13})$

In this case, $K = 6n$. By known formulas for moments of a hypergeometric distribution, $E[X_p] = \frac{3n}{4}$ and $Var[X_p] = \frac{21n^2}{16(8n-1)}$. Thus, by Chebychev's Inequality we have

$$\begin{aligned}
P(X_p \notin A_p) &= P\left((k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \leq \frac{2\sigma}{\delta}\right) \\
&= P\left(\left(\frac{7n}{8}\right)^{1/2} - \frac{4X_p - 3n}{\sqrt{24n}} \leq \frac{2\sigma}{\delta\mu_{diff}}\right) \\
&= P\left(\frac{4X_p - 3n}{\sqrt{24n}} \geq -\frac{2\sigma}{\delta\mu_{diff}} + \left(\frac{7n}{8}\right)^{1/2}\right) \\
&\leq \left(\frac{21n^2}{24n(8n-1)}\right)\left(\frac{56n\mu_{diff}^2}{(-2\sigma\sqrt{56n} + 7n\delta\mu_{diff})^2}\right)\delta^2 \\
&\leq \delta^2 \text{ for sufficiently large } n
\end{aligned}$$

Thus, $P(X_p \in A_p) \geq 1 - \delta^2$ for sufficiently large n.

Case 7: $p \in (P_7 \cup P_{14})$

In this case, $K = 7n$. By known formulas for moments of a hypergeometric distribution, $E[X_p] = \frac{7n}{8}$ and $Var[X_p] = \frac{49n^2}{64(8n-1)}$. Thus, by Chebychev's Inequality we

have

$$\begin{aligned}
P(X_p \notin A_p) &= P\left((k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \leq \frac{2\sigma}{\delta}\right) \\
&= P\left(\left(\frac{7n}{8}\right)^{1/2} - \frac{8X_p - 7n}{\sqrt{56n}} \leq \frac{2\sigma}{\delta\mu_{diff}}\right) \\
&= P\left(\frac{8X_p - 7n}{\sqrt{56n}} \geq -\frac{2\sigma}{\delta\mu_{diff}} + \left(\frac{7n}{8}\right)^{1/2}\right) \\
&\leq \left(\frac{49n^2}{56n(8n-1)}\right)\left(\frac{56n\mu_{diff}^2}{(-2\sigma\sqrt{56n} + 7n\delta\mu_{diff})^2}\right)\delta^2 \\
&\leq \delta^2 \text{ for sufficiently large } n
\end{aligned}$$

Thus, $P(X_p \in A_p) \geq 1 - \delta^2$ for sufficiently large n.

This yields the following general result: Let A_p be a subset of $Supp(X_p)$ such that $x \in A_p$ implies $(k_n(1) - k_n^{\pi}(p))(\mu_{diff}) \geq \frac{2\sigma}{\delta}$. For sufficiently large n, we have $P(X_p \in A_p) \geq 1 - \delta^2$ for all $1 \leq p \leq 254$. Thus we have proved Corollary 4. By Boole's inequality we have the following:

$$\begin{aligned}
P\left(X_p \in A_p \forall p \in \{1, 2, ..., 254\}\right) &= P\left((X_1 \in A_1) \cap (X_2 \in A_2) \cap ... \cap (X_{254} \in A_{254})\right) \\
&= 1 - P\left((X_1 \notin A_1) \cup (X_2 \notin A_2) \cup ... \cup (X_{254} \notin A_{254})\right) \\
&\geq 1 - P\left(X_1 \notin A_1\right) - P\left(X_2 \notin A_2\right) ... - P\left(X_{254} \notin A_{254}\right) \\
&\geq 1 - 254\delta^2
\end{aligned} \tag{5.18}$$

By independence we have $P(X_p \in A_p \forall p | T = T(1)) = P(X_p \in A_p \forall p)$. Thus, by the

law of total probability we have

$$\begin{aligned}
P(T > T_\pi | T = T(1)) &= P((T > T_\pi | X_p \in A_p \forall p, T = T(1)) P(X_p \in A_p \forall p | T = T(1)) \\
&\quad + P((T > T_\pi | X_p \notin A_p \forall p, T = T(1)) P(X_p \notin A_p \forall p | T = T(1)) \\
&\geq P((T > T_\pi | X_p \in A_p \forall p, T = T(1)) P(X_p \in A_p \forall p | T = T(1)) \\
&= P((T > T_\pi | X_p \in A_p \forall p, T = T(1)) P(X_p \in A_p \forall p) \\
&\geq (1 - 761\delta^2)(1 - 254\delta^2) \text{ by (5.17) and (5.18)} \\
&= 1 - 1015\delta^2 + (761)(254)\delta^4 \\
&\geq 1 - 1015\delta^2
\end{aligned} \tag{5.19}$$

As δ can be chosen arbitrarily, this yields $P(T > T_\pi | T = T(1)) \to 1$ as $n \to \infty$. Thus, we have

$$\begin{aligned}
P(T > T_\pi) &= P(T > T_\pi | T = T(1)) P(T = T(1)) + P(T > T_\pi | T \neq T(1)) P(T \neq T(1)) \\
&\geq P(T > T_\pi | T = T(1)) P(T = T(1)) \\
&\geq (1 - 1015\delta^2)(1 - 253\delta^2) \text{ by (5.14) and (5.19)} \\
&= 1 - 1268\delta^2 + (1268)(253)\delta^4 \\
&\geq 1 - 1268\delta^2
\end{aligned} \tag{5.20}$$

As δ can be chosen arbitrarily, this yields $P(T > T_\pi) \to 1$ as $n \to \infty$. Thus, for 1000 randomly selected permutations, $\{\pi_1, \pi_2 ..., \pi_{1000}\}$, the probability that $T > T_{\pi_i}$ for at

least 951 of the π_i's (i.e., that the permutation p-value is less than .05) is

$$\begin{aligned} P(T > T_{\pi_i} \text{ for at least 951 of the } \pi_i\text{'s}) &\geq P(T > T_{\pi_i} \text{ for 1000 of the } \pi_i\text{'s}) \\ &\geq (1 - 1268\delta^2)^{1000} \text{ by (5.20)} \\ &= 1 - \epsilon \end{aligned}$$

where for every $\epsilon > 0$, δ can be chosen such that

$$\delta = \sqrt{\frac{1 - (1-\epsilon)^{\frac{1}{1000}}}{1268}}$$

Thus, the probability that the permutation p-value is less than .05 must also tend to 1 as $n \to \infty$. Returning to the beginning of the proof, we have shown that $P(A) \to 1$ as $n \to \infty$, implying that $P(A^c)$ tends to 0 as n tends to infinity. Thus, by the expression in (5.11), we have shown $P(A \cap B) \to 1$ as $n \to \infty$. Note that this result still holds for any arbitrary value chosen for c, instead of $c = .05$.

5.3 Theoretical Considerations for Present and Non-present Two-way and Three-way Interactions

Suppose we have ten SNPs, $SNP_A, SNP_B, ..., SNP_J$, and are interested in four independent interactions:

1. $SNP_A \times SNP_B$
2. $SNP_C \times SNP_D \times SNP_E$
3. $SNP_F \times SNP_G$
4. $SNP_H \times SNP_I \times SNP_J$

Suppose each of the ten SNPs carry two states, condition 1 and condition 2. For illustration, assign labels to each of the SNP states as shown in table 5.8. Suppose

Table 5.8: Labels for the SNP states.

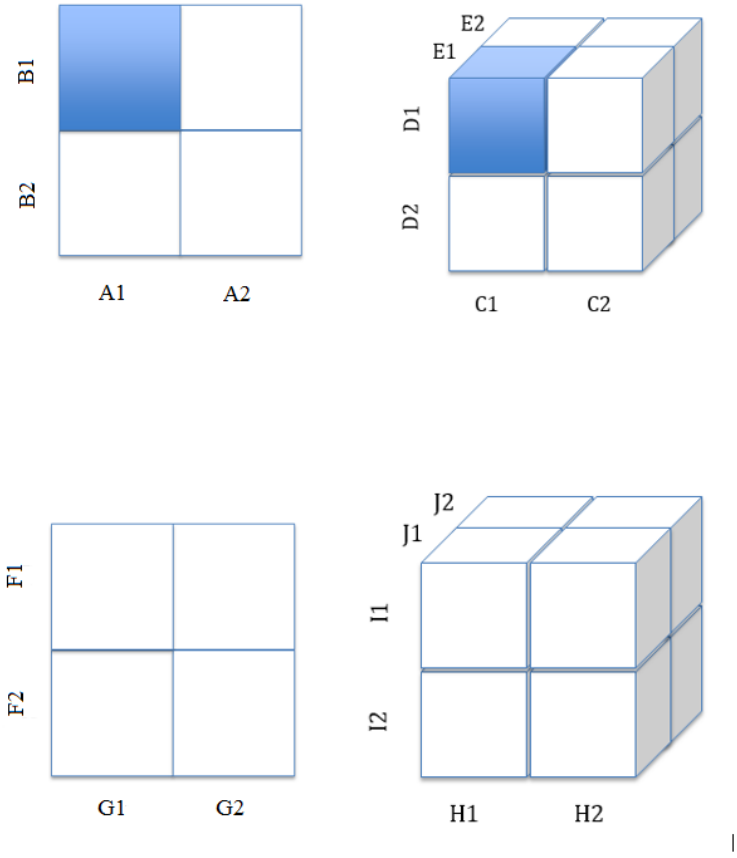
A1 := SNP_A, condition 1	A2 := SNP_A, condition 2
B1 := SNP_B, condition 1	B2 := SNP_B, condition 2
C1 := SNP_C, condition 1	C2 := SNP_C, condition 2
D1 := SNP_D, condition 1	D2 := SNP_D, condition 2
E1 := SNP_E, condition 1	E2 := SNP_E, condition 2
F1 := SNP_F, condition 1	F2 := SNP_F, condition 2
G1 := SNP_G, condition 1	G2 := SNP_G, condition 2
H1 := SNP_H, condition 1	H2 := SNP_H, condition 2
I1 := SNP_I, condition 1	I2 := SNP_I, condition 2
J1 := SNP_J, condition 1	J2 := SNP_J, condition 2

we have $8n$ subjects, and let $Y_1, ..., Y_{8n}$ represent the values of some quantitative phenotype for each subject, $i = 1, 2, ..., 8n$. Assume that subjects are distributed evenly among multifactor classes within interactions. That is, each multifactor cell within a two-way interaction contains $2n$ subjects, and each multifactor cell within a three-way interaction contains n subjects. Further, suppose that the $SNP_A \times SNP_B$ interaction and the $SNP_C \times SNP_D \times SNP_E$ interaction are present and that the remaining two interactions are not. That is,

$$Y_i = \mu_0 + \mu_1(A1_i)(B1_i) + \mu_2(C1_i)(D1_i)(E1_i) + \epsilon_i \tag{5.21}$$

where $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$ and $\mu_1, \mu_2 > 0$. Note that some loss of generality occurs with the $\mu_1, \mu_2 > 0$ assumption. Figure 5.6 illustrates the four interactions. Here, the shaded

Figure 5.6: Illustration of the four interactions to be considered.



multifactor cells represent genotypic combinations for which the phenotype mean is increased by the presence of an interaction.

Regarding the arbitrary cutoff aggregated score (ACAS) with $c = .05$, we will show the following propositions:

1. $P(SNP_A \times SNP_B$ is included in the aggregated score$) \to 1$, $n \to \infty$.
2. $P(SNP_C \times SNP_D \times SNP_E$ is included in the aggregated score$) \to 1$, $n \to \infty$.
3. $P(SNP_F \times SNP_G$ is included in the aggregated score$) \leq .05$
4. $P(SNP_H \times SNP_I \times SNP_J$ is included in the aggregated score$) \leq .05$

The proofs proceed as follows:

Proposition 1: $P(SNP_A \times SNP_B$ is included in the aggregated score$) \to 1$, $n \to \infty$.

To show this, we refer back to section 5.1. If we let $\mu_{high} = \mu_0 + \mu_1 + \frac{1}{8}\mu_2$ and adjust for a doubled sample size, the proof for proposition follows from the sequence in which we proved that $P(A) \to 1, n \to \infty$ in section 4.1.

Proposition 2: $P(SNP_C \times SNP_D \times SNP_E$ is included in the aggregated score$)$ $\to 1$, $n \to \infty$.

To show this, we refer back to section 5.2. If we let $\mu_{high} = \mu_0 + \mu_2 + \frac{1}{4}\mu_1$, the proof of this proposition follows from the sequence in which we proved that $P(A) \to 1, n \to \infty$ in section 4.2.

Proposition 3: $P(SNP_F \times SNP_G$ is included in the aggregated score$) \leq .05$

As illustrated in section 5.1, there are there are 14 possible "high"/"low" configurations (labeling of quadrants as "high" or "low") which can be assigned to a 2×2 table with AQMDR. Let them be indexed by $p = 1, 2, ..., 14$. See figure 5.3. Again, we define $c_i(p)$ for each $p = \{1, 2, ..., 14\}$, as shown in table 5.2. Define $T(p)$ for each $p \in \{1, ..., 14\}$ as follows:

$$T(p) := \frac{\sum_{i=1}^{8n} c_i(p) Y_i}{\sqrt{\sum_{i=1}^{8n} c_i(p)^2}}$$

Consider T(1).

$$\begin{aligned} T(1) =& \frac{\sum_{i=1}^{8n} c_i(1) Y_i}{\sqrt{24n}} \\ =& \frac{\sum_{i=1}^{2n} (3) Y_i + \sum_{i=2n+1}^{8n} (-1) Y_i}{\sqrt{24n}} \\ =& \frac{6n\bar{Y}_1 - 6n\bar{Y}_2}{\sqrt{24n}} \\ =& \frac{6n(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{24n}} \end{aligned}$$

where $\bar{Y}_1$ is a sample mean of realized values of Y_i for $i \in \{1, ..., 2n\}$, and $\bar{Y}_2$ is a sample mean of Y_i for $i \in \{2n+1, ..., 8n\}$ (all of which have mean $\mu_0 + \frac{1}{4}\mu_1 + \frac{1}{8}\mu_2$). As $\sum_{i=1}^{8n} c_i(1) = 0$,

$$T(1) \sim N\left(0, \sigma^2\right).$$

A series of similar calculations yield the following general result for all $p \in \{1, ..., 14\}$:

$$T(p) \sim N\left(0, \sigma^2\right).$$

Define a T-statistic as follows:

$$T := max_{1 \leq p \leq 14} T(p).$$

Consider a fixed permutation $\pi : \{1, 2, ..., 8n\} \to \{1, 2, ..., 8n\}$. Define $T_\pi(p)$ as follows:

$$T_\pi(p) := \frac{\sum_{i=1}^{8n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{8n} c_{\pi(i)}(p)^2}} = \frac{\sum_{i=1}^{8n} c_{\pi(i)}(p) Y_{\pi(i)}}{\sqrt{\sum_{i=1}^{8n} c_i(p)^2}}.$$

Define the test statistic for permutation π as follows:

$$T_\pi := max_{1 \leq p \leq 14} T_\pi(p).$$

As T and T_π are functions of the Y_i's, let's call them $T(\mathbf{Y})$ and $T_\pi(\mathbf{Y})$ where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_{8n} \end{bmatrix}.$$

Now, suppose we have 1000 permutations $\pi_1, \pi_2, ..., \pi_{1000}$, and these correspond to $T_{\pi_1}(\mathbf{Y}), T_{\pi_2}(\mathbf{Y})$, ..., $T_{\pi_{1000}}(\mathbf{Y})$. It is obvious that if we were to sort the set $T(\mathbf{Y})$, $T_{\pi_1}(\mathbf{Y})$,..., $T_{\pi_{1000}}(\mathbf{Y})$ from lowest to highest, the ordering would be random. That is, $T(\mathbf{Y})$ is equally likely to be between any two of the $T_{\pi_i}(\mathbf{Y})$'s for $i \in \{1, ..., 1000\}$.

Let A_i be $I[T(\mathbf{Y}) < T_{\pi_i}(\mathbf{Y})]$. Then $\sum_{j=1}^{1000} A_j$ is equal to the number of $T_{\pi_i}(\mathbf{Y})$'s on the right side of $T(\mathbf{Y})$ in the ordering mentioned above. Because this ordering is random, $\sum_{j=1}^{1000} A_j \sim Unif(0, 1000)$. In order for this interaction to be included in the arbitrary cutoff aggregated score with $c = .05$, less than 50 of 1000 independent permutations of the Y_i's should produce T_π's that are greater than T. That is, $\sum_{j=1}^{1000} A_j \leq 50$. Based on the uniform distribution of $\sum_{j=1}^{1000} A_j$, P(the interaction is included in the aggregated score) $= P(\sum_{j=1}^{1000} A_j < 50) < .05$.

Proposition 4: $P(SNP_H \times SNP_I \times SNP_J$ is included in the aggregated score) $\leq .05$ The proof of this proposition is analogous to that of Proposition 3.

Chapter 6

Application: Exploring Interactions Between APOE and Known Alzheimer's Disease Associated SNPs

6.1 Introduction

Alzheimer's disease (AD) is a devastating neurodegenerative disorder that is characterized by impairment in memory and decreased cognition [14]. AD is progressive in nature, typically beginning with mild loss of memory but often leading to difficulties with communication, loss of ability to respond to surroundings and loss of autonomy [15]. AD is the most common form of dementia worldwide, affecting more than 24 million people [14]. Forecasts suggest that by 2050, Alzheimer's disease will affect 1 out 85 people. In the United States alone, an estimated 5 million people are currently diagnosed with AD, and this number is expected to increase to about 14 million by 2050 [16].

The criteria for diagnosis of AD established by the National Institute on Aging and Alzheimer's Association are dispersed between two categories. The first category includes core clinical criteria, which can be assessed by physicians without the need for advanced imaging equipment or cerebrospinal fluid analysis. The second category includes research criteria, which involve the use of biomarkers based on imaging and measures from cerebrospinal fluid [17]. For this research we focus on biomarkers that directly reflect the pathology of AD by providing evidence of the presence of proteins deposited in the brain during the course of AD, such as the amyloid-beta protein (Aβ) and tau. Research suggests that build up of these two proteins in the brain may be a mark of AD. A low measure of Aβ in cerebrospinal fluid (CSF) serves as a marker of Aβ deposition in the brain. Similarly, a high level of tau or phosphorylated-tau (p-tau) in CSF serves as a marker of tau accumulation in the brain [17].

In addition to these CSF biomarkers, genetic influences on AD risk have also been identified. The most notable of these is the *apolipoprotein E* (*APOE*) gene [17]. In humans, the *APOE* gene exists in three polymorphic alleles: $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$. Genome-wide association studies (GWAS) suggest that $\epsilon 4$ is the strongest genetic risk factor for AD. In the overall population, $\epsilon 4$ occurs in 13.7% of people, but among AD patients this frequency is increased to 40% [18]. In fact, research has shown that the $\epsilon 4$ allele of the *APOE* gene (*APOE4*) carries a two-to-three fold increase in AD risk for individuals with one $\epsilon 4$ allele, and a 12-fold increase in risk for those with two $\epsilon 4$ alleles [19].

Previous studies have identified SNPs with known main effect associations with Alzheimer's disease status. Lambert et al. [20] analyzed 7,055,881 SNPs as main effects, and found that 19 SNPs were associated with AD. Some of these susceptibility loci had previously been identified in similar studies (from the *ABCA7*, *APOE*, *BIN1*, *CLU*, *CR1*, *CD2AP*, *EPHA1*, *MS4A6A*, and *PICALM* genes), but several susceptibility loci were newly-identified in the study. These loci are found on the following genes: *HLA*, *PTK2B*, *SORL1*, *SLC24A4*, *DSG2*, *INPP5D*, *MEF2C*, *NME8*, *ZCWPW1*, *CELF1*, *FERMT2*, and *CASS4* [20]. Although *CD33* was not identified by the Lambert et al. study as an AD susceptibility gene, Griciuc et al. produced results which suggest that *CD33* may play a role in reducing Alzheimer's risk by preventing diminishing Aβ in cerebrospinal fluid [21]. Similarly, *MAPT* (which encodes for CSF tau) has been identified as a possible susceptibility gene for AD [22], as well as *Presenilin E318G* [23] and *TREM2* [24]. Based on these studies, a set of 23 known AD associated SNPs (in addition to *APOE4* count) were selected for this study, and are displayed in table 6.1 [20, 23, 24, 21]. Note that these SNPs each take values of 0, 1, or 2.

Table 6.1: SNPs considered in this study.

Chromosome	SNP	Position	Gene
1	rs6656401	207518704	*CR1*
2	rs6733839	127135234	*BIN1*
2	rs35349669	233159830	*INPP5D*
5	rs190982	88927603	*MEF2C*
6	rs75932628	41161514	*TREM2*
6	rs10948363	47520026	*CD2AP*
7	rs2718058	37801932	*NME8*
7	rs1476679	100406823	*ZCWPW1*
7	rs11771145	143413669	*EPHA1*
8	rs28834970	27337604	*PTK2B*
8	rs9331896	27610169	*CLU*
11	rs10838725	47536319	*CELF1*
11	rs983392	60156035	*MS4A6A*
11	rs10792832	86156833	*PICALM*
11	rs11218343	121564878	*SORL1*
14	rs17125944	52933911	*FERMT2*
14	rs17125721	73206470	*PSEN1*
14	rs10498633	92460608	*SLC24A4*
17	rs8070723	46003698	*MAPT*
18	rs8093731	31508995	*DSG2*
19	rs4147929	1063444	*ABCA7*
19	rs3865444	51224706	*CD33*
20	rs7274581	56443204	*CASS4*

In this study, we aimed to identify gene-to-gene interactions that demonstrated association with counts of tau and Aβ found in cerebrospinal fluid, while restricting ourselves to only those two-way and three-way interactions including *APOE4*. We applied the Aggregated Quantitative Multifactor Dimensionality Reduction (AQMDR) method proposed in this work to exhaustively search for significant interactions and aggregate them into two predictors (one aggregated score for two-way interactions, and one for three-way interactions). We also applied the Quantitative Multifactor Dimensionality Reduction (QMDR) technique, which resulted in candidate two-way and three-way interactions, and selected an optimal interaction between the two candidates. We then evaluated the resulting AQMDR and QMDR models by providing a training data R^2 for each method.

6.2 Methods

Data for this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The ADNI study is a longitudinal study conducted in multiple North American sites in which participants are assessed clinically, and genetic and biomarker information is recorded for these individuals throughout the aging process [25]. The sample data set for this application study included observations from 419 individuals for whom baseline measures of CSF tau and CSF Aβ were available. 253 of these subjects were female and 166 subjects were male. The subjects ranged in age from 54 to 89, with an average age of 74.51. About 28% of the subjects were classified as normal controls (N) regarding disease status, about 48% were diagnosed with mild cognitive impairment (MCI), and 24% were diagnosed with mild Alzheimer's disease (AD).

We restricted attention to gene-to-gene interactions which included *APOE4*. That is, we considered 23 two-way interactions (*APOE4* with each of the SNPs listed in table 6.1) and $\binom{23}{2}$ three-way interactions. These interactions were evaluated for association with two quantitative outcome variables, CSF $A\beta$ and CSF tau. Each of these outcome variables were considered separately using the following analysis steps:

1. Subjects were stratified based on disease status (N, MCI or AD), and a separate analysis was performed for each stratification group. The rationale for this stratification stems from the fact that the three disease strata represent three vastly different populations.

2. To control for common AD covariates [26], the quantitative outcome variables were regressed on *gender*, *age*, and *education* and the residuals were used as the new quantitative outcome variable of interest.

3. We applied the Aggregated Quantitative Multifactor Dimensionality Reduction

(AQMDR) method to these subsets of interactions and recorded the permutation p-values obtained for each interaction. Note that the sample data had missing values, which were excluded in this initial application study. That is, for each interaction, only subjects with observations for all SNP factors involved in the interaction were included in the analysis. For each interaction, the main effects for involved SNPs were incorporated in a regression model for the outcome variable, and then residuals of this model were used as the outcome variable for the AQMDR analysis.

4. The permutation p-values were then used to develop an aggregated score for two-way interactions and an aggregated score for three-way interactions. These aggregated scores were developed using the Arbitrary Cutoff Aggregated Score (ACAS) with $c = .10$. Note that lower values for c were initially considered, but a higher cutoff was chosen in order to have at least one interaction included in an aggregated score for the quantitative outcomes. Again, subjects with missing values for SNPs which were included in the aggregated score were excluded from the analysis.

5. We combined the aggregated scores (the two-way aggregated score and the three-way aggregated score) in a linear regression model for the outcome using the Simultaneous Inclusion method described in chapter 3.

6. The regression model was evaluated with a training R^2.

7. To compare the proposed AQMDR method with current methodology, for each of the outcome variables we used Quantitative Multifactor Dimensionality Reduction (QMDR) to select a candidate two-way interaction and a candidate three-way interaction, and to identify an optimal interaction between the two candidates (see chapter 1 of this work for details of the QMDR method). The

optimal interaction was used as a predictor in a linear regression model and evaluated with a training R^2.

6.3 AQMDR Results

In the results that follow, an AQMDR permutation p-value that is less than .10 is denoted with (*), a p-value that is less than .05 is denoted with (**). Also note that although the interactions we consider in this study are SNP-by-SNP interactions, for ease of interpretation, we denote each SNP by its gene name. For example, in some of the following tables we write *CR1* instead of rs6656401.

Table 6.2 displays the permutation p-values obtained in the AQMDR implementation for two-way interactions regarding CSF tau. Each row in the table represents the two-way interaction between the given SNP and *APOE4*. Permutation p-values are given for each disease status stratum (N, MCI or AD). Note that for an arbitrary cutoff of $c = .10$, only one of the 23 two-way interactions is identified as significant, *APOE4* by *ABCA7*. Further, *APOE4* by *ABCA7* was only identified as significant within the normal control (N) stratum.

Table 6.3 summarizes the significant interactions for CSF tau obtained from the consideration of three-way interactions and the corresponding permutation p-values obtained for the interactions in the AQMDR analysis. Note that if an interaction was significant in any of the three disease strata, it is included in this table along with the permutation p-values for all three strata.

Table 6.4 displays the permutation p-values obtained in the AQMDR implementation for two-way interactions regarding CSF Aβ. Again, each row in the table represents the two-way interaction between the given SNP and *APOE4*. Permutation p-values are given for each disease status stratum (N, MCI or AD). For an arbitrary

Table 6.2: Two-way interaction permutation p-values for CSF tau.

SNP	Gene	N	MCI	AD
rs6656401	*CR1*	1.000	0.961	0.743
rs6733839	*BIN1*	0.995	0.642	0.937
rs35349669	*INPP5D*	0.589	0.996	0.982
rs190982	*MEF2C*	0.996	0.998	0.390
rs75932628	*TREM2*	1.000	1.000	1.000
rs10948363	*CD2AP*	0.921	0.964	0.593
rs2718058	*NME8*	1.000	1.000	0.896
rs1476679	*ZCWPW1*	1.000	0.993	0.691
rs11771145	*EPHA1*	0.950	1.000	0.641
rs28834970	*PTK2B*	0.337	0.940	0.895
rs9331896	*CLU*	0.924	0.988	0.462
rs10838725	*CELF1*	0.682	0.985	0.999
rs983392	*MS4A6A*	0.993	0.574	1.000
rs10792832	*PICALM*	0.402	0.759	0.872
rs11218343	*SORL1*	1.000	1.000	1.000
rs17125944	*FERMT2*	0.995	0.982	0.717
rs17125721	*PSEN1*	0.707	0.180	1.000
rs10498633	*SLC24A4*	1.000	0.896	1.000
rs8070723	*MAPT*	0.953	0.991	0.985
rs8093731	*DSG2*	1.000	0.388	0.981
rs4147929	*ABCA7*	0.084*	0.792	0.984
rs3865444	*CD33*	1.000	0.877	0.891
rs7274581	*CASS4*	0.883	1.000	0.636

cutoff of $c = .10$, only one interaction is identified as significant, *APOE4* by *PTK2B*. Further, *APOE4* by *PTK2B* was only significant in disease stratum MCI.

Table 6.5 summarizes the significant interactions for CSF Aβ obtained from the consideration of three-way interactions and the corresponding permutation p-values obtained for the interactions in the AQMDR analysis. Note that if an interaction was significant in any of the three disease strata, it is included in this table along with the permutation p-values for all three strata.

Table 6.3: Significant three-way interactions for CSF tau.

Interaction	N	MCI	AD
APOE4 x *CD2AP* x *PICALM*	0.086*	0.969	0.998
APOE4 x *ABCA7* x *NME8*	0.080*	0.656	0.984
APOE4 x *CR1* x *MS4A6A*	0.836	0.065*	0.987
APOE4 x *MEF2C* x *CD2AP*	0.988	0.998	0.043**
APOE4 x *TREM2* x *ZCWPW1*	0.981	1.000	0.084*
APOE4 x *MAPT* x *CELF1*	0.845	0.983	0.093*

6.4 QMDR Results

Table 6.6 displays the candidate two-way and three-way interactions for CSF tau selected by the QMDR analysis, and identifies which candidate interaction was selected as the optimal interaction. Interactions printed in bold face indicate those that were identified as significant (permutation p-value less than .10) in the AQMDR analysis. Note that for the quantitative outcome variable, CSF tau, all of the interactions identified as optimal by QMDR were also selected by AQMDR (with $c = .10$) for inclusion in the Arbitrary Cutoff Aggregated Score. For disease stratum N, *APOE4* x *INPP5D* x *PTK2B* was selected as the candidate three-way interaction in QMDR, but the permutation p-value obtained for that interaction in AQMDR was 0.327. As we saw in table 6.3, two three-way interactions were identified as significant by AQMDR, neither of which were *APOE4* x *INPP5D* x *PTK2B*. For disease stratum MCI, *APOE4* x *PSEN1* was selected as the candidate two-way interaction by QMDR, but the AQMDR permutation p-value was 0.180. AQMDR did not identify any significant two-way interactions for disease stratum MCI. For disease stratum AD, *APOE4* x *SLC24A4* (which had an AQMDR permutation p-value of 1.000) was identified as a candidate interaction in QMDR. AQMDR did not identify any significant two-way interactions in the AD stratum.

Table 6.7 displays the candidate two-way and three-way interactions for CSF Aβ selected by the QMDR analysis, and identifies which candidate interaction was se-

Table 6.4: Two-way interaction permutation p-values for CSF $A\beta$.

SNP	Gene	N	MCI	AD
rs6656401	*CR1*	0.527	0.824	0.776
rs6733839	*BIN1*	0.799	0.848	0.933
rs35349669	*INPP5D*	0.838	0.971	0.983
rs190982	*MEF2C*	0.998	0.883	0.403
rs75932628	*TREM2*	1.000	1.000	1.000
rs10948363	*CD2AP*	1.000	0.945	0.625
rs2718058	*NME8*	0.705	0.986	0.871
rs1476679	*ZCWPW1*	0.595	0.632	0.611
rs11771145	*EPHA1*	0.867	0.956	0.660
rs28834970	*PTK2B*	1.000	0.011**	0.896
rs9331896	*CLU*	1.000	0.780	0.487
rs10838725	*CELF1*	0.988	0.839	1.000
rs983392	*MS4A6A*	0.945	0.616	1.000
rs10792832	*PICALM*	0.706	0.973	0.878
rs11218343	*SORL1*	1.000	1.000	1.000
rs17125944	*FERMT2*	0.999	0.842	0.710
rs17125721	*PSEN1*	0.707	0.180	1.000
rs10498633	*SLC24A4*	0.634	1.000	1.000
rs8070723	*MAPT*	0.617	0.986	0.988
rs8093731	*DSG2*	1.000	0.806	0.984
rs4147929	*ABCA7*	0.995	0.977	0.984
rs3865444	*CD33*	0.961	0.971	0.872
rs7274581	*CASS4*	0.712	1.000	0.632

lected as the optimal interaction. Interactions printed in bold face indicate those that were identified as significant (permutation p-value less than .10) in the AQMDR analysis. In disease stratum MCI, both candidate interactions were also identified as significant by AQMDR. For disease stratum N, *APOE4* x *MAPT* was selected as the two-way candidate interaction by QMDR. However, this interaction yielded a permutation p-value of 0.617 in the AQMDR analysis. AQMDR did not identify any significant two-way interactions within the N disease stratum. QMDR identified *APOE4* x *BIN1* x *PICALM* as the candidate three-way interaction (as well as the optimal interaction) for the N stratum, yet this interaction yielded an AQMDR p-value of 0.238. AQMDR identified two significant three-way interactions for the N group, neither of which were *APOE4* x *BIN1* x *PICALM*. Within the AD group, *APOE4*

Table 6.5: Significant three-way interactions for CSF Aβ.

Interaction	N	MCI	AD
APOE4 x *CR1* x *SLC24A4*	0.067*	0.986	0.518
APOE4 x *BIN1* x *SLC24A4*	0.098*	0.539	0.997
APOE4 x *PTK2B* x *MAPT*	0.986	0.100*	0.989
APOE4 x *PTK2B* x *TREM2*	0.999	0.021**	0.936
APOE4 x *PTK2B* x *NME8*	0.698	0.074*	0.956
APOE4 x *PTK2B* x *ZCWPW1*	0.941	0.012*	0.642
APOE4 x *PTK2B* x *CELF1*	1.000	0.099*	0.553
APOE4 x *PTK2B* x *SORL1*	1.000	0.011**	0.697
APOE4 x *PTK2B* x *FERMT2*	0.983	0.016**	0.950
APOE4 x *PTK2B* x *PSEN1*	0.985	0.031**	0.980
APOE4 x *PTK2B* x *DSG2*	0.827	0.014**	0.584
APOE4 x *PTK2B* x *ABCA7*	0.999	0.085*	0.270
APOE4 x *PTK2B* x *CD33*	0.993	0.008**	0.970
APOE4 x *PTK2B* x *CASS4*	0.930	0.077*	0.688

Table 6.6: QMDR candidate interactions for CSF tau.

Strat.	Two-way Candidate	Three-way Candidate	Optimal
N	***APOE4* x *ABCA7***	*APOE4* x *INPP5D* x *PTK2B*	two-way
MCI	*APOE4* x *PSEN1*	***APOE4* x *CR1* x *MS4A6A***	three-way
AD	*APOE4* x *SLC24A4*	***APOE4* x *MAPT* x *CELF1***	three-way

x *TREM2* and *APOE4* x *CD2AP* x *CLU* were selected as candidate interactions by QMDR, and yielded respective AQMDR permutation p-values of 1.000 and 0.329. AQMDR did not identify any significant two-way or three-way interactions for the AD disease stratum.

Table 6.7: QMDR candidate interactions for CSF Aβ.

Strat.	Two-way Candidate	Three-way Candidate	Optimal
N	*APOE4* x *MAPT*	*APOE4* x *BIN1* x *PICALM*	three-way
MCI	***APOE4* x *PTK2B***	***APOE4* x *PTK2B* x *CD33***	two-way
AD	*APOE4* x *TREM2*	*APOE4* x *CD2AP* x *CLU*	two-way

6.5 Model Implementation

The results described in the previous sections for both CSF tau and CSF Aβ were implemented in linear models, and these models were evaluated using training R^2

values. To illustrate the model implementation process (steps 4-6 from section 6.2 above) for AQMDR, we will consider the quantitative outcome variable, CSF tau, and focus on the N disease group. Recall from tables 6.2 and 6.3 that *APOE4* x *ABCA7*, *APOE4* x *ABCA7* x *NME8*, and *APOE4* x *CD2AP* x *PICALM* yielded permutation p-values less than the arbitrary cutoff of $c = .10$. The two-way aggregated score will contain only the *APOE4* x *ABCA7* interaction and the three-way aggregated score will contain both *APOE4* x *ABCA7* x *NME8* and *APOE4* x *CD2AP* x *PICALM*. All other interactions will be ignored. As the two-way aggregated score is based off of only one interaction, the score for each subject in the data set will be the "high"/"low" classification (1 for "high" or "0" for low) assigned to the subject in the AQMDR classification. That is, the value of this aggregated score for each subject will be 0, or 1. The three-way aggregated score is based on two interactions, and for each subject will be the sum of the "high"/"low" classification value from *APOE4* x *ABCA7* x *NME8* and the "high"/"low" classification value from *APOE4* x *CD2AP* x *PICALM*. That is, the value of this aggregated score for each subject will be 0, 1, or 2.

Before implementing the AQMDR and QMDR models, we first regressed the outcome variable, CSF tau on the three covariates and the main effects of all SNPs which are included in the aggregated scores and the optimal QMDR interaction (*APOE4*, *ABCA7*, *NME8*, *CD2AP*, and *PICALM*). We include all SNPs included in the aggregated score and the QMDR optimal interaction in the interest of using the same set of residuals as the quantitative outcome for the AQMDR model implementation and the QMDR model implementation. Only those subjects in the N stratum with observations for all of these covariates and relevant SNPs were included in the analysis.

For the AQMDR model implementation, the residuals from the previously mentioned regression model were used as the response in a linear regression model with

the aggregated scores as predictors. Using the Simultaneous Inclusion method described in chapter 3, the two-way aggregated score and the three-way aggregated score were included in the linear model simultaneously. For the QMDR implementation, the same set of residuals used for the AQMDR implementation were used as the response in a linear regression model with the "high"/"low" classification of subjects based on the optimal interaction (*APOE4* x *ABCA7* in our example) as the only predictor.

This process was repeated for each disease stratum, and again for CSF Aβ in the three disease strata. Training R^2's for each of the QMDR and AQMDR linear models were recorded in tables 6.8 and 6.9. For CSF tau (table 6.8), the AQMDR implementation yields higher training R^2's for disease strata N and AD. Note that the two implementations were identical for stratum MCI because AQMDR identified only one significant interaction, which was the same interaction identified as optimal by QMDR.

Table 6.8: Training R^2's for CSF tau.

Stratum	AQMDR R^2	QMDR R^2
N	0.2342	0.0730
MCI	0.1435	0.1435
AD	0.2790	0.1884

For CSF Aβ, in disease stratum N, AQMDR yielded a slightly higher training R^2 than QMDR. Recall that in this stratum, neither candidate interaction from QMDR was identified as significant by AQMDR. As AQMDR still yielded a training R^2 value comparable to that of QMDR, perhaps AQMDR identified two important three-way interactions, yet failed to pick up on *APOE4* x *BIN1* x *PICALM* (which was selected as optimal by QMDR). This may have been addressed with a larger ACAS cutoff value, c. For disease stratum MCI, the AQMDR implementation yielded a higher

training R^2 than QMDR. Recall that in this case, AQMDR identified both of the QMDR candidate interactions as significant, and many others as well. This may be the type of situation for which AQMDR was designed. That is, there may be many gene-to-gene interactions present in nature. If this is the case, QMDR ignores all but one of these interactions due to the selection of a single optimal interaction. AQMDR would be advantageous in this case, as it incorporates the effects of multiple gene-to-gene interactions based on permutation p-values. For the AD stratum, recall that AQMDR did not identify any significant two-way or three-way interactions, thus there was no aggregated score to implement in a model. QMDR selected an optimal interaction by default, but when this interaction was implemented in a linear model, the model yielded a very low training R^2 value of 4.685 x 10^{-32}, likely due to the regression of the main effects.

Table 6.9: Training R^2's for CSF Aβ.

Stratum	AQMDR R^2	QMDR R^2
N	0.2717	0.2369
MCI	0.4373	0.1512
AD	-	4.685 x 10^{-32}

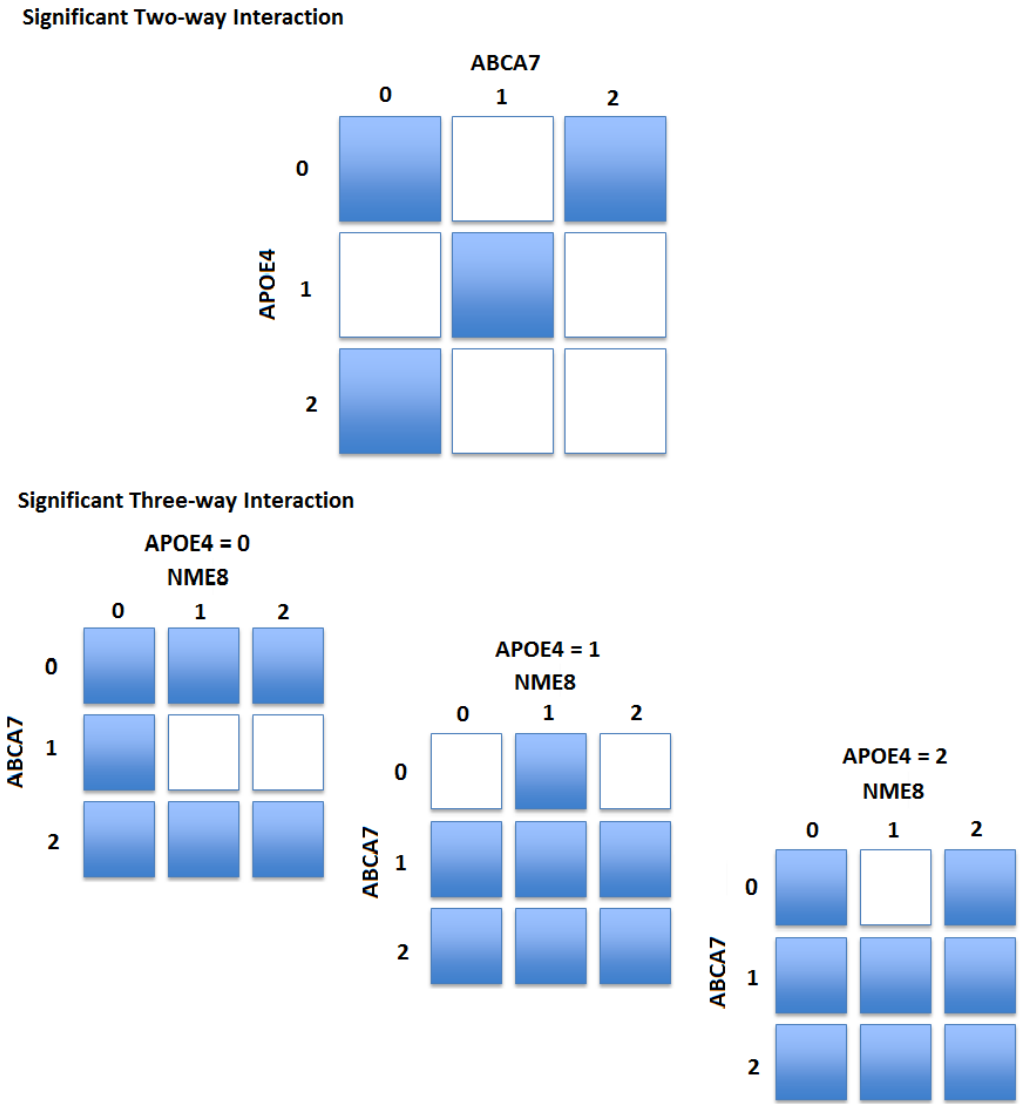
6.6 Confounding

One concern surrounding the AQMDR results is the possibility of confounding. For example, for CSF tau within disease stratum N, the two-way interaction between *APOE4* and *ABCA7* was identified as significant. The three-way interaction among *APOE4*, *ABCA7* and *NME8* was also identified as significant. It is possible that this three-way interaction was only significant due to confounding caused by *APOE4* x *ABCA7*. If this is the case, inclusion of both of these interactions in the AQMDR implementation would be redundant. In order to explore this possibility, we have displayed the "high"/"low" classifications assigned to the significant interactions by

AQMDR in figure 6.1. In this figure, the shaded multifactor cells are those classified as "high" (1), and the white cells are those classified as "low".

Based on the classification assigned to the two-way interaction, we would expect a three-way interaction induced by confounding to exhibit the same pattern of classification. That is, we expect the three-way interaction to exhibit "high" classifications in multifactor cells where $(APOE4 = 0 \cap ABCA7 = 1)$, $(APOE4 = 1 \cap ABCA7 = 0)$, $(APOE4 = 1 \cap ABCA7 = 2)$, or $(APOE4 = 2 \cap ABCA7 = 1)$. However, when we observe the classification pattern of the three-way interaction among *APOE4*, *ABCA7* and *NME8*, we see that the "high" multifactor cells extend beyond the *APOE4* and *ABCA7* combinations identified in the two-way interaction. Based on these patterns, we cannot conclude that the significance of the three-way interaction is merely a consequence of the two-way interaction.

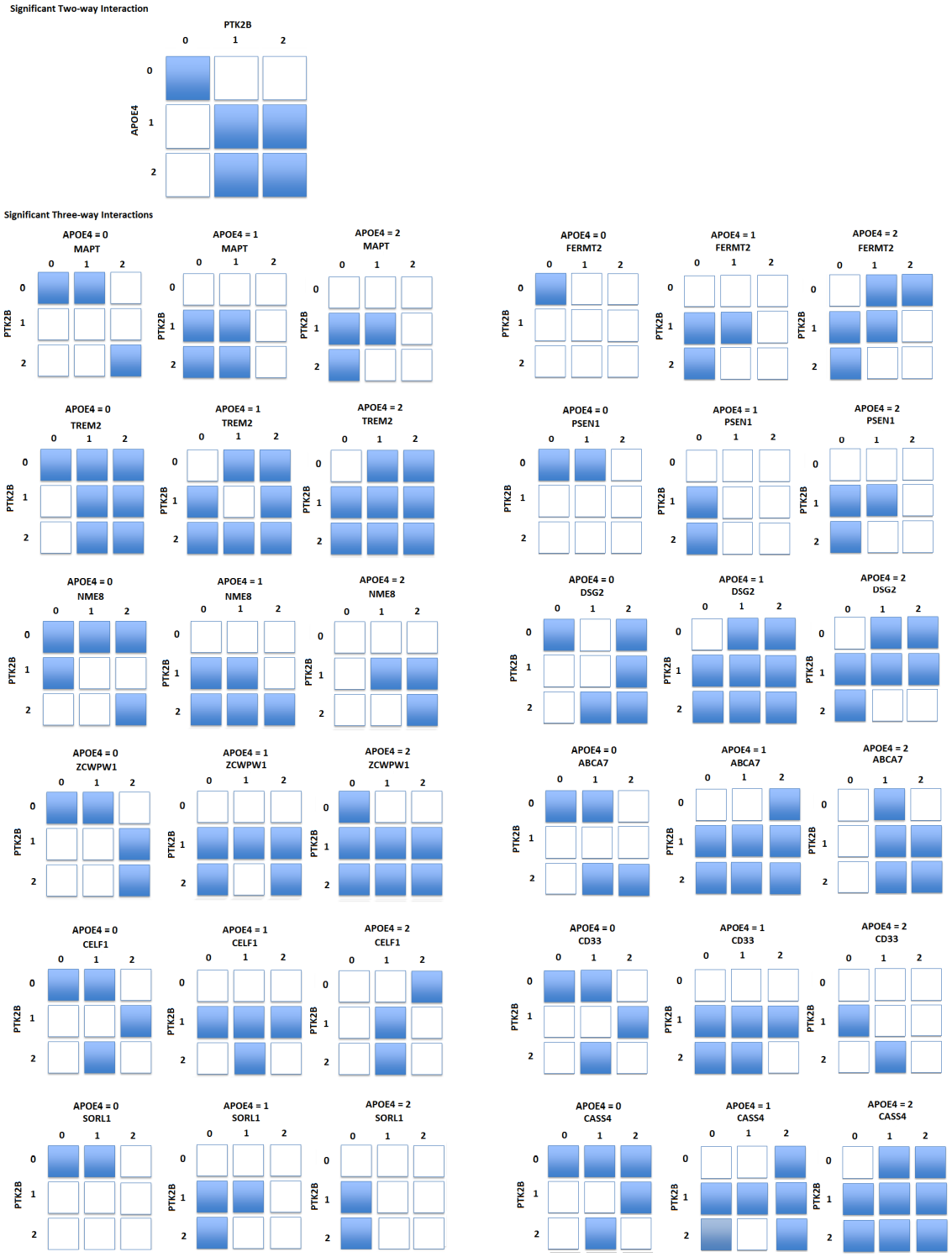
Figure 6.1: "High"/"low" classifications for significant AQMDR interactions for CSF tau in stratum N.



For CSF Aβ within the MCI stratum, the interaction between *APOE4* and *PTK2B* was identified as significant in AQMDR. Twelve three-way interactions were also identified, all of which contain *APOE4* and *PTK2B*. Figure 6.2 illustrates the "high"/"low" classifications assigned to the significant interactions by AQMDR. In this figure, the shaded multifactor cells are those classified as "high", and the white cells are those

classified as “low”. As we saw with the CSF tau interactions, the classification patterns extend beyond those we would expect based on confounding caused by the two-way interaction, and the three-way interactions all display distinct “high”/“low” classification patterns. Thus, we cannot conclude that the three-way interactions are significant due only to the significance of the two-way interaction.

Figure 6.2: "High"/"low" classifications for significant AQMDR interactions for CSF Aβ in stratum MCI.



6.7 Future Work

In this chapter, we have provided some interesting results regarding the analysis of ADNI data using AQMDR and QMDR. We have provided data-driven evidence that AQMDR may provide better predictions and identification of significant interactions than QMDR in situations where more than one interaction is influencing a quantitative trait. Future analysis of the sample data will be preceded by imputation of the data to eliminate missing observations. Data with fewer missing SNP values will allow for consistent sample sizes across all interaction considerations, and will also facilitate the implementation of the more complex aggregated scores (CWAS and HAS) discussed in chapter 2 of this work. A larger sample size for each interaction consideration will also provide higher power. It is also important to note that if a bonferroni correction were used to correct for multiple tests, none of the interactions identified as significant with a cutoff of $c = .10$ in AQMDR would have been identified as significant using the correction. In the continuation of this study with imputed data, we hope to explore the necessity for correction of multiple tests.

In addition to the continuation of this data application, there are also avenues to explore regarding the AQMDR method. Future work may include the consideration of outcome variables in vector form. For example, rather than performing separate analyses regarding CSF tau and $A\beta$, we may be able to use the AQMDR framework to consider the vectorized outcome (tau, $A\beta$). It may also be of interest to consider alternate approaches to the categorization of multifactor cells. For example, we may want to consider weighting each multifactor class within an interaction with some continuous score between 0 and 1. In addition, AQMDR is based on an extension of QMDR, but could easily be adapted to incorporate Generalized MDR (GMDR) [10]. As the GMDR approach yields a "high"/"low" classification of multifactor cells within each interaction as well as a p-value associated with each k-way interaction,

the method easily lends itself to the aggregated scores discussed in this work.

References

[1] Dai H, Charnigo RJ, Becker ML, Leeder JS, and Motsinger-Reif AA. Risk score modeling of multiple gene to gene interactions using aggregated-multifactor dimensionality reduction. *Biodata Min*, 6(1):1, 2013.

[2] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, and Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 69(1): 138–147, 2001.

[3] Hosmer DW and Lemeshow S. *Applied logistic regression*. John Wiley and Sons, New York, 2000.

[4] Nelson M, Kardia SLR, Ferrell RE, and Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res*, 11(458-470), 2001.

[5] Hahn LW, Ritchie MD, and Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19(3):376–382, 2003.

[6] Motsinger AA and Ritchie MD. Multifactor dimensionality reduction: an analysis strategy for modeling and detecting gene-gene interactions in genetics and pharmacological studies. *Human Genomics*, 2(5):318–328, 2006.

[7] Bush WS, Edwards TL, Dudek SM, Mckinney BA, and Ritchie MD. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *BMC Bioinformatics*, 9:238, 2008.

[8] Winham SJ and Motsinger-Reif AA. An r package implementation of multifactor dimensionality reduction. *BioData Min*, 4(1):24, 2011.

[9] Moore JH, 2004-2015. URL `www.epistasis.org`.

[10] Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, and et al. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *The American Journal of Human Genetics*, 80:1125–1137, 2007.

[11] Calle ML, Urrea V, Malatsi Riera N, and Van Steen K. Mb-mdr: Model based multifactor dimensionality reduction for detecting interaction in high-dimensional genomic data. *Ann Hum Genet*, 75(1):78–89, Jan 2011.

[12] Gui J, Moore JH, Williams SM, Andrews P, Hillege HL, van der Harst P, Navis G, Van Gilst WH, Asselbergs FW, and Gilbert-Diamond D. A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. *PLoS ONE*, 8(6):e666545, 2013.

[13] Guo SW and Thompson EA. Performing the exact test of hardy-weinberg proportion for multiple alleles. *Biometrics*, 48(2):361–372, June 1992.

[14] Coon KD, Myers AJ, Webster JA, Pearson JV, Lince DH, and et. al. A high-density whole-genome association study reveals that *APOE* is the major susceptibility gene for sporadic late-onset alzheimer's disease. *J Clin Psychiatry*, 68: 613–618, 2013.

[15] Center for Disease Control and Prevention. Healthy aging, 2015. URL `http://www.cdc.gov/aging/aginginfo/alzheimers.html`.

[16] Brookmeyer R, Johnson E, Ziegler-Graham K, and Arrighi HM. Forecasting the global burden of alzheimer's disease. *Alzheimer's Dement*, 3(3):186–191, 2007.

[17] Albert MS, Dekosky ST, and Dickson et. al. D. The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's Dement*, 7(3):270–279, 2011.

[18] Liu CC, Kanekiyo T, Xu H, and Bu G. Apolipoprotein e and alzheimer's disease: risk, mechanisms, and therapy. *Nat Rev Neurol*, 9(2):106–118, 2013.

[19] Saykin AJ et. al. Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. *Alzheimer's Dement*, 69(1):265–273, 2010.

[20] Lambert JC et. al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nat Genet.*, 45(12):1452–1458, Dec 2013.

[21] Griciuc A et al. Alzheimer's disease risk gene *CD33* inhibits microglial uptake of amyloid β. *Neuron*, 78:631–643, 2013.

[22] Allen M et al. Association of *MAPT* haplotypes with alzheimer's disease risk and *MAPT* brain gene expression levels. *Alzheimer's Research and Therapy*, 6: 39, 2014.

[23] Hippen AA et al. *Presenilin E318G* variant and alzheimer's disease risk: the cache county study. *BMC Genomics*, 17:438, June 2016.

[24] Jonsson T et al. Variant of *TREM2* associated with the risk of alzheimer's disease. *N. Engl. J. Med.*, 368:117–127, 2013.

[25] URL `http://adni.loni.usc.edu/study-design/`.

[26] Sharp ES and Gatz M. The relationship between education and dementia: An updated systematic review. *Alzheimer Dis Assoc Discord*, 25(4):289–304, 2011.

Vita

Rebecca Elaine Crouch

Birth Place: La Grange, Kentucky

Education:
University of Kentucky, Lexington, KY
M.S. in Statistics, 2014
Western Kentucky University, Bowling Green, KY
B.A. in Mathematics and Economics, 2012

Employment
Teaching assistant, August 2012 - December 2016
Department of Statistics, University of Kentucky
Summer Program for Operations Research Technology Intern, May 2016-August 2016
National Security Agency, Ft. Meade, Maryland

Scholastic and Professional Honors
R.L. Anderson Research Award
Department of Statistics, University of Kentucky
R.L. Anderson Teaching Award
Department of Statistics, University of Kentucky