Robust and Efficient Modeling for Gene-Environment and Gene-Gene Interactions in Longitudinal Cohort Studies

by

Yi-An Ko

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2014

Doctoral Committee:

Professor Bhramar Mukherjee, Chair
Professor Thomas M. Braun
Assistant Professor Sung Kyun Park
Professor Naisyin Wang

ACKNOWLEDGEMENTS

First of all, I am grateful for God's provision of blessings, challenges, and grace for growth. He provides everything I need to accomplish this dissertation. I can do all things through Christ who gives me strength (Philippians 4:13).

Second, I would like to express my sincere gratitude to my advisor, Dr. Bhramar Mukherjee, for her tremendous mentorship, patient guidance, and constant encouragement throughout my PhD training. She is not only my mentor, but my friend, whom I can discuss life with. Also, I would like to thank my committee members, Drs. Thomas Braun, Sung Kyun Park, and Naisyin Wang for their contributions and thought-provoking advice. I especially appreciate Dr. Park for his input and suggestions for the analysis of the Normative Aging Study data. Moreover, I would like to thank Ana Diez Roux, Jennifer Smith, Sharon Kardia, Xu (Steven) Wang, and Kari Moore for their help in obtaining and analyzing the data from the Multi-Ethnic Study of Atherosclerosis.

Additionally, I thank my friends in the department, who provided me with assistance in my research and made my day-to-day office life more enjoyable. I also want to thank my brothers and sisters in Christ at Ann Arbor Chinese Christian Church for an amazing period of growing and walking in the Lord. I really do appreciate their support, perspective, and encouragement. Last but not least, I would like to thank my husband, Liang-Ching Tsai and my parents, Kuang-Jung Ko and Hsu-Chin Lin, for their endless love and support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Robust and Efficient Modeling for Gene-Environment and Gene-Gene Interactions in Longitudinal Cohort Studies

by

Yi-An Ko

Chair: Bhramar Mukherjee

While there have been extensive statistical methods on gene-environment interaction (GEI) in case-control studies, little attention has been given to robust and efficient modeling of GEI in longitudinal cohort studies. In a two-way table for GEI with row and column as categorical variables, a conventional saturated model involves estimation of distinct interaction effect for each cell. However, the degrees of freedom (df) for testing interaction can grow quickly with increasing number of categories, resulting in decreased efficiency and reduced power for detecting interaction. This dissertation considers the problem of modeling GEI with repeated measures data on a quantitative trait using parsimonious models for non-additivity proposed in the classical Analysis of Variance literature. We first provide an overview of these classical models and explore the interaction structures by simply reducing repeated measurements to summary level cell means. In the first project, we modify the cell-mean method and propose a parametric bootstrap approach using these interaction models. Both methods account for the unbalanced and longitudinal nature of the data. In the second project, we propose a shrinkage estimator that combines estimates from a saturated

interaction model and Tukey's single df model for non-additivity. It is useful for conducting multiple GEI tests where distinct interaction patterns could occur in different genetic markers. The proposed estimator is robust to various interaction structures and the corresponding test is valid based on simulation results. In the third project, we focus on additive main effects and multiplicative interaction (AMMI) models. We develop an alternating maximum likelihood estimation procedure for AMMI models and approximate the null distribution of the likelihood ratio test statistic by a chi-square with fractional df. The proposed methods are illustrated using data from the Normative Aging Study and the Multi-Ethnic Study of Atherosclerosis. Both datasets come from longitudinal cohort studies involving rich genetic data and several environmental exposure factors that could be time varying or time invariant. Overall, this dissertation contributes to adaptation of classical interaction models to longitudinal studies, with the goal of understanding the dynamic interplay between genes and environment over time.

CHAPTER I

Introduction

1.1 Gene-Environment Interaction and Gene-Gene Interaction

Genome-wide association studies (GWAS), in which millions of single nucleotide polymorphisms (SNPs) are measured on thousands of samples, have identified many genetic variants that are associated with complex diseases and disorders. In the post-GWAS era, there is growing recognition that additive effects of genetic variants alone may not explain all the variation in complex disease traits. The development of complex diseases may be related to interactions between genes (i.e., epistasis used in population genetics) and/or interactions between genetic variants and environmental exposures and health behavioral factors (i.e., gene-environment interactions).

The presence of gene-environment interaction (GEI), or gene-gene interaction (GGI), indicates that the association between an outcome and an environmental exposure (or gene) depends on genotype. In other words, the effect of an environmental exposure (or a gene) on a disease outcome may be enhanced or reduced in a particular genotype group. GEI/GGI is important in genetic studies because if a true interaction remains unidentified, it can mask the detection of a genetic effect and lead to inconsistencies in the findings of genetic associations with disease. Given as such, the search for GEI and GGI has been receiving considerable attention in recent years (Khoury and Wacholder, 2009). Specifically, interaction models have been of interest in genetic association studies since appropriate modeling may lead to increased sta-

tistical power (Brem et al., 2005; Marchini et al., 2005; Evans et al., 2006) and help discover the underlying biological mechanisms of GEI/GGI. A better understanding of GEI will assist in developing practical strategies for disease prevention by modifying behavioral factors and/or avoiding harmful exposure to genetically susceptible sub-groups to attenuate or modify the effects of deleterious genes by avoiding the harmful environmental exposure.

Given that many of the GWAS consortia are based on case-control studies, there has been extensive statistical research for testing GGI and GEI in case-control studies. On the other hand, prospective longitudinal cohort studies have gained interest over the years because they not only are a natural choice for assessing causal relationship (not subject to recall bias), but also provide time dependent information of exposure history for detecting potential GEI. Furthermore, prospective longitudinal studies allow for the study of GEI effects on quantitative traits that are linked to chronic diseases, rather than a binary disease occurrence outcome as in case-control studies. Several large-scale longitudinal environmental epidemiology studies with characterization of exposure history, such as the UK Biobank (www.ukbiobank.ac.uk) and the Multi-Ethnic Study of Atherosclerosis (www.mesa-nhlbi.org), have been collecting genetic data to study GEI. Therefore, it is warranted to develop powerful and valid statistical methods for interactions in longitudinal cohort studies.

1.2 Motivation

For the analysis of GGI or GEI in prospective cohort studies, current statistical strategies typically attempt to model the interaction effect by fitting a regression model to the conditional mean structure of an outcome Y with main effects of G, E and G×E terms while adjusting for confounding factors. A product term for interaction reflects that the effect of the row and the column variables may not be additive. However, this simple regression model may prevent the identification of the actual

interaction structure with time varying exposure. Moreover, a product form (i.e., saturated interaction form) for GEI may not yield efficient estimates when both G and E are categorical variables. Alternatively, one can try to model the interaction term using the generalized additive mixed model framework (Lin and Zhang, 1999), but tests for such non-parametric, smoothed interaction terms will also result in reduced power for studies with moderate sample size. Therefore, robust and efficient modeling of GGI and GEI should be considered when constructing powerful tests for longitudinal data.

In this dissertation, we borrow several parsimonious models for non-additivity proposed in the classical analysis of variance (ANOVA) literature to address the issue of efficient modeling of GEI/GGI in longitudinal cohort studies. Genetic and environmental exposure factors are considered as categorical variables. Any main effect and interaction effect are considered as fixed effects (i.e., population-average effects as opposed to subject-specific effects) although extensions to random-effects models have been studied (Oman, 1991; Piepho, 1997, 1998; Smith et al., 2001). We restrict ourselves to the models in which the error terms and the random effects are normally distributed. Missingness in longitudinal data is assumed to be missing completely at random or missing at random.

1.3 Organization of This Dissertation

Chapter II provides an overview of classical ANOVA models for non-additivity in a two-way table context, including Tukey's one degree of freedom (df) model (Tukey, 1949), Mandel's row or column regression model (Mandel, 1961), Tukey's row-column regression model (Tukey, 1962), and additive main effects and multiplicative interaction (AMMI) models (Gollob, 1968; Mandel, 1971; Gauch Jr., 1992). We explore interaction structures using these models in a two-way classification array for longitudinal cohort studies by simply reducing data to cell means. In Chapter III, we

modify the cell-mean approach to account for within-subject correlation and propose a parametric bootstrap resampling procedure to test interaction effects using these classical interaction models. In addition, we describe a visual and diagnostic tool for characterizing subject-specific and time-specific contributions to the interaction factors. In Chapter IV, we propose a shrinkage estimator that combines the estimates from Tukey's one df model and a saturated interaction model. This shrinkage estimator is found to be robust to misspecification of interaction structure and is very useful when searching for GEI across multiple SNPs. In Chapter V, we specifically focus on AMMI models and develop an alternating maximum likelihood estimation algorithm and propose a likelihood-based test for AMMI models. We further explore the possibility of time-varying GEI effects. Throughout the chapters, we illustrate the proposed methods using data from the Normative Aging Study and the Multi-Ethnic Study of Atherosclerosis. Overall, the dissertation contributes to the adaptation of classical interaction models to longitudinal cohort studies, with the goal of understanding the dynamic interplay between genes and environment over time.

Remark: Part of the work presented in this dissertation has been published in peer-reviewed journals. Chapter II is extracted from Mukherjee et al. (2012). For more details, please refer to Mukherjee, B., Ko, Y.A., VanderWeele, T., Roy, A., Park, S.K., Chen, J. 2012. Principal interactions analysis for repeated measures data: application to gene–gene and gene–environment interactions. *Statistics in Medicine* 31(22): 2531-2551. Chapter III has been published in *Genetic Epidemiology* in 2013 (Ko, Y.A., Saha-Chaudhuri. P., Park, S.K., Vokonas, P.S., Mukherjee, B. 2013. Novel likelihood ratio tests for screening gene-gene and gene-environment interactions with unbalanced repeated-measures data. *Genetic Epidemiology* 37(6): 581-591). Chapter IV has been accepted for publication in *Statistics in Medicine*.

CHAPTER II

Overview of Classical Interaction Models and AMMI Models

2.1 Introduction

This chapter gives an overview of several parsimonious models for the structure of interaction (or non-additivity) in a two-way table context in the classical ANOVA literature. We explore interaction structures using these models for longitudinal cohort studies by considering the average of repeated measures as a single observation per subject and then examining the cell-mean structure corresponding to the $G = g, E = e$ in a two-way genotype $\times$ environment classification array ($G_1 = g_1, G_2 = g_2$ for a two-way gene $\times$ gene array). Due to the two-way ANOVA formulation, the methods presented are applicable to genotype data on single nucleotide polymorphisms (SNP) and *categorical* environmental exposures. Though we study the methods in the context of GGI or GEI, they are applicable to exploring interactions in any two-way classification array.

In Section 2.2, we describe four classical models, Tukey's (Tukey, 1949), Mandel's row and column regression (Mandel, 1961), and Tukey's row-column models (Tukey, 1962). In Section 2.2.2, we introduce the "principal interactions analysis (PIA)" via the additive main effects and multiplicative interaction effects (AMMI) model (Gollob, 1968; Mandel, 1971). In Section 2.3, we present the analysis results of the Normative Aging Study (NAS) data using the five models described in Section 2.2. In Section 2.4, we conduct simulation studies using cell-mean based approaches to

examine the robustness of the classical models for a general $I \times J$ table and under common epistasis models.

2.2 Model

Since the models are generic to any two-way table, instead of using G and E for the two factors, we use R to denote the row variable with I levels, and C to denote the column variable with J levels. Let y_{ijkh} be the h-th observation corresponding to the k-th subject in the (i, j)-th cell of this $I \times J$ array. We consider a simple model:

$$y_{ijkh} = \mu + R_i + C_j + \gamma_{ij} + b_{ijk} + e_{ijkh}, \tag{2.1}$$
$$h = 1, \ldots, n_{ijk}, k = 1, \ldots, N_{ij}, i = 1, \ldots, I, j = 1, \ldots, J.$$

Here μ describes the overall mean, R_i and C_j are the row and column main effects and γ_{ij} describes the interaction between the row and column factors. The standard constraints, $\sum_i R_i = \sum_j C_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$, are placed on the fixed effects parameters. Let $N = \sum_i \sum_j N_{ij}$ denote the total number of subjects. We assume that the possible subject-specific random effect $b_{ijk} \sim \mathcal{N}(0, \sigma_b^2)$ and the random errors $e_{ijkh} \sim \mathcal{N}(0, \sigma_e^2)$.

We create a two-way cell-mean array, first averaging all observations corresponding to the k-th subject in the (i, j)-th cell, namely $\overline{y}_{ijk.}$, and then averaging $\overline{y}_{ijk.}$ over all subjects in the (i, j)-th cell, to obtain $\{\overline{y}_{ij.}\}$. These cell means will have different variability, depending on the random effects structure and the number of observations per subject as well as number of subjects per cell. We abuse our notations slightly by dropping the $\{\cdot\}$ suffixes and describe the models in terms of the $I \times J$ array of cell means $\overline{y}_{ij} = \overline{y}_{ij.}$.

The implied mean model by (2.1) of a general saturated model for interaction for

the two-way table in terms of cell means $\overline{y}_{ij}$ is given by

$$\overline{y}_{ij} = \mu + R_i + C_j + \gamma_{ij} + \overline{e}_{ij}, i = 1, \ldots, I, j = 1, \ldots, J. \tag{2.2}$$

where $\overline{e}_{ij}$ is the mean of the errors of e_{ijkh} in (2.1). In the following, we denote $\overline{y}_{ij}$ by y_{ij} and $\overline{e}_{ij}$ by e_{ij}, pretending that it represents a single observation corresponding to the (i, j)-th cell (Barhdadi and Dubé, 2010). We assume that $e_{ij} \sim N(0, \tau^2)$. This assumption does not recognize the non-constant variance in the cell-means due to unbalanced nature of the data. The maximum likelihood estimates (MLEs) of main effects and interaction parameters are given by

$$\hat{\mu} = y_{..},\ \hat{R}_i = y_{i.} - y_{..},\ \hat{C}_j = y_{.j} - y_{..}. \tag{2.3}$$

Define the estimated residual contrast after fitting the additive terms as $z_{ij} = y_{ij} - \hat{\mu} - \hat{R}_i - \hat{C}_j = y_{ij} - y_{i.} - y_{.j} + y_{..}$. The df attributed to testing interaction in a saturated model is $(I-1)(J-1)$, and, in that case $\hat{\gamma}_{ij} = z_{ij}$. With more than one replication per cell, one can test for interaction in a saturated model; however, with a single observation per cell, one exhausts the df for a saturated interaction model, with no df left for errors. Thus, a test of interaction cannot be carried out. In such situation, several reduced df tests have been proposed by imposing special structures on the interaction parameters. These structures can be used for testing interactions in general regression models for a more powerful test with reduced df (Chatterjee et al., 2006; Maity et al., 2009).

2.2.1 Classical Models for Interaction

One Degree of Freedom Test for Non-Additivity

The essential idea behind this model (Tukey, 1949) is to think of interaction as $\gamma_{ij} = \theta R_i C_j + \xi_{ij}$, namely, a leading term and some residual noise ξ_{ij} that can be absorbed with the error term e_{ij}. Thus, of the $(I-1)(J-1)$ df attributed to the

interaction term, only 1 is used to test $H_0 : \theta = 0$ and the rest is attributed to the residual error term, making it possible to test for non-additivity with no replication. Tukey's model is given by:

$$y_{ij} = \mu + R_i + C_j + \theta R_i C_j + e_{ij}, \tag{2.4}$$

where θ is the coefficient for the linear by linear interaction effect. The least square estimate of θ, denoted as $\hat{\theta}$, is given by

$$\hat{\theta} = \frac{\sum_i \sum_j z_{ij} \hat{R}_i \hat{C}_j}{\sum_i \sum_j \hat{R}_i^2 \hat{C}_j^2} = \frac{\sum_i \sum_j y_{ij} \hat{R}_i \hat{C}_j}{\sum_i \sum_j \hat{R}_i^2 \hat{C}_j^2}.$$

Where $z_{ij} = y_{ij} - y_{i.} - y_{.j} + y_{..}$ is again the estimated residual contrast after removing additive main effects. This essentially reduces to regressing the cell residuals after fitting the additive terms on the product of estimated row and column main effects (Tukey, 1962). The model is not identifiable if there are no main effects present as any value of θ yields the same likelihood. Tukey's single df test for non-additivity is obtained by using the test statistic $F = MS_{ll}/MSE$ as presented in Table 2.3 in Appendix that has an F distribution with 1 and $(I-1)(J-1)-1$ degrees of freedom under $H_0 : \theta = 0$.

Column (Row) Regression Model

Mandel (1961) proposed the column regression model and row regression model for testing interactions. In the column-regression model, the interaction effect is a linear function of the column main effects, i.e.,

$$y_{ij} = \mu + R_i + C_j + \lambda_i C_j + e_{ij}, \tag{2.5}$$

where λ_i is the coefficient corresponding to the i-th row, and $\sum_i \lambda_i = 0$. The maximum likelihood estimate of λ_i, denoted as $\hat{\lambda}_i$, is

$$\hat{\lambda}_i = \frac{\sum_j z_{ij} \hat{C}_j}{\sum_j \hat{C}_j^2}.$$

The MLE of μ and R_i remain unchanged. A test of non-additivity is obtained by constructing an F-statistic for the hypothesis $H_0 : \lambda_i = 0, i = 1, \ldots, I$. Under the null hypothesis and normality, this test statistic as described in Appendix, has an F distribution with $(I-1)$ and $(I-1)(J-1)-(I-1)$ degrees of freedom. Table 2.4 in the Appendix presents the ANOVA table for this model. By replacing the columns with the rows, one can equivalently posit a row regression model of the following form:

$$y_{ij} = \mu + R_i + C_j + R_i\eta_j + e_{ij}, \tag{2.6}$$

with $\sum_j \eta_j = 0$ and testing $H_0 : \eta_j = 0, j = 1, \ldots, J$, with the resultant F-statistic having df $\{J-1, (I-1)(J-1)-(J-1)\}$.

Row-Column Regression Model

Tukey (1962) extended Mandel's column- or row-regression model in his seminal paper where he introduced the vacuum cleaner strategy for analyzing two-way arrays where a row regression was followed up with a column regression (or vice versa).

$$y_{ij} = \mu + R_i + C_j + \theta R_i C_j + \lambda_i C_j + R_i\eta_j + e_{ij}, \tag{2.7}$$

where λ_i and η_j are the row- and column-specific coefficients, with additional constraints $\sum_i \lambda_i = \sum_j \eta_j = 0$ and $\sum_i \lambda_i R_i = \sum_j \eta_j C_j = 0$. The MLEs of μ, R_i, C_j remain unchanged. The maximum likelihood estimates of θ, λ_i, and η_j are obtained as:

$$\hat{\theta} = \frac{\sum_i \sum_j z_{ij}\hat{R}_i\hat{C}_j}{\sum_i \sum_j \hat{R}_i^2\hat{C}_j^2}, \qquad \hat{\lambda}_i = \frac{\sum_j z_{ij}\hat{C}_j}{\sum_j \hat{C}_j^2} - \hat{\theta}\hat{R}_i, \qquad \hat{\eta_j} = \frac{\sum_i z_{ij}\hat{R}_i}{\sum_i \hat{R}_i^2} - \hat{\theta}\hat{C}_j. \tag{2.8}$$

Table 2.5 in Appendix presents the ANOVA Table corresponding to this model. The F-statistic for testing $H_0 : \theta = \lambda_i = \eta_j = 0, \ \forall \ i, j$ under the above constraints have numerator df $1 + (I-2) + (J-2) = (I+J-3)$. The denominator df is thus $(I-1)(J-1) - (I+J-3)$. For a 3×3 table for studying GGI, Tukey's row-column model has 3 df for the interaction term, offering little power gain over the

saturated model with 4 df. Thus, we refrain from presenting results for this model in our simulation studies in Section 2.4.

Models (2.4)-(2.7) are hierarchically built in an increasing order of complexity in the interaction structure. They provide different degrees of efficiency gain and model robustness. For a large two-way array, say a 9×5 array, the interaction tests will have 1 (Tukey's 1-df), 8 (Mandel's column), 4 (Mandel's row) for the numerator of the F-statistic and 31, 24, and, 28 df for the denominator, respectively. thus providing different degrees of efficiency gain and model robustness. The F-test statistics and the corresponding df for testing interaction effects of the classical models as well as the MLEs for interaction parameters are provided in Appendix (Section 2.5).

Given that the interaction structures in these classical models are functions of main effects, the models would encounter problem with likelihood identifiability when main effects do not exist. Even in presence of main effects, under any form of misspecification of this specific structure, all of the above tests lose tremendous power.

2.2.2 Principal Interactions Analysis via the AMMI Model

The principle of AMMI is to first fit additive main effects and then to apply singular value decomposition (SVD) to the matrix of residuals that remain after the fitting of main effects. The general class of AMMI models (Gollob, 1968; Mandel, 1971) is given by

$$y_{ij} = \mu + R_i + C_j + \sum_{m=1}^{M} d_m \alpha_{im} \beta_{jm} + e_{ij}. \tag{2.9}$$

In AMMI models, the $I \times J$ interaction matrix $\boldsymbol{\Gamma} = ((\gamma_{ij}))$ is expressed by the representation: $\boldsymbol{\Gamma} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{B}^\top$. Here $\boldsymbol{A} = ((\alpha_{im}))$ and $\boldsymbol{B} = ((\beta_{km}))$ are $I \times M$ and $J \times M$ orthonormal matrices ($\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{I}$) and $\boldsymbol{D}$ is a $M \times M$ diagonal matrix with elements $d_1 \geq d_2 \cdots \geq d_M$, where $M \leq \min(I-1, J-1)$. Eckart and Young (1936) showed that for a fixed M, the least square estimates of $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D})$ can be found by expressing the estimated matrix $\hat{\boldsymbol{\Gamma}}$ of interaction parameters with

entries $\hat{\gamma}_{ij} = y_{ij} - y_{i.} - y_{.j} + y_{..}$ in terms of a SVD as specified by the factor model, $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{A}}\hat{\boldsymbol{D}}\hat{\boldsymbol{B}}^{\top}$.

An alternative interpretation is that the interaction parameter is expressed as a sum of several successive multiplicative contrasts $\Psi_{Fm} = \sum_i \sum_j \alpha_{im}\beta_{jm}\gamma_{ij}$ such that each contrast is orthogonal to all previous contrasts and accounts for a maximum of the remaining variance. Let $\hat{\Psi}_{Fm}$ denote the estimated normalized orthogonal multiplicative contrast among the interaction parameters $\{\gamma_{ij}\}$ and SS_{Fm} denote the sum of squares due to the m-th interaction factor. Then from the classical contrast theory, $SS_{Fm} = \hat{\Psi}_{Fm}^2$, where $\hat{\Psi}_{Fm}$ can be obtained by

$$\hat{\Psi}_{Fm} = \sum_i \sum_j \hat{\alpha}_{im}\hat{\beta}_{jm}\hat{\gamma}_{ij} = \sum_i \sum_j \hat{\alpha}_{im}\hat{\beta}_{jm}y_{ij}.$$

Because $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{A}}\hat{\boldsymbol{D}}\hat{\boldsymbol{B}}^{\top}$, we have $\hat{\boldsymbol{D}} = \hat{\boldsymbol{A}}^{\top}\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{B}}$, implying $\hat{d}_m = \sum_i \sum_j \hat{\alpha}_{im}\hat{\beta}_{jm}\hat{\gamma}_{ij}$. So, $\hat{\Psi}_{Fm}$ and $\hat{d}_m$ are equivalent. They both are $\sum_i \sum_j \hat{\alpha}_{im}\hat{\beta}_{jm}\hat{\gamma}_{ij}$. Hence, $SS_{Fm} = \hat{\Psi}_{Fm}^2 = \hat{d}_m^2$. Let SS_{RC} denote the total sum of squares due to row-column interaction. The sum of squares corresponding to the residual interaction after M successive interaction factors being extracted from $\{\gamma_{ij}\}$ is therefore

$$SS_{Fres} = SS_{RC} - \sum_{m=1}^{M} SS_{Fm} = SS_{RC} - \sum_{m=1}^{M} \hat{d}_m^2 = \sum_{m=M+1}^{I-1} \hat{d}_m^2,$$

$$SS_{RC} = \sum_i \sum_j (y_{ij} - y_{i.} - y_{.j} + y_{..})^2 = \sum_i \sum_j \hat{\gamma}_{ij}^2 = \hat{\boldsymbol{\Gamma}}^{\top}\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{D}}^{\top}\hat{\boldsymbol{D}} = \sum_{m=1}^{I-1} \hat{d}_m^2.$$

This method integrates ANOVA and principal component analysis (PCA). The AMMI models also target towards a sparse representation of interaction terms, but not through main effects. By considering a reduced rank approximation (rank one approximation if only the first component is retained) to the interaction matrix, one is able to save df and enhance efficiency when compared to the saturated interaction model. To this end, we call this method "Principal Interactions Analysis" (PIA) due to its similarity with PCA.

A special case of (2.9) is of particular interest when $M = 1$. Namely,

$$y_{ij} = \mu + R_i + C_j + d_1\alpha_i\beta_j + e_{ij}, \sum_i \alpha_i = \sum_j \beta_j = 0, \quad \sum_i \alpha_i^2 = \sum_j \beta_j^2 = 1. \quad (2.10)$$

The test of no interaction is equivalent to testing $H_0 : d_1 = 0$. Johnson and Graybill (1972a) derived the distributional properties for the likelihood ratio test (LRT) of $H_0 : d_1 = 0$. They showed that the MLE of d_1, $\hat{d}_1$ say, is given by the square-root of the largest characteristic root of $\hat{\mathbf{\Gamma}}^\top\hat{\mathbf{\Gamma}}$, say l_1. The maximum likelihood is attained when $\{\alpha_i\}$ and $\{\beta_j\}$ are given by the normalized characteristic vector corresponding to l_1 in $\hat{\mathbf{\Gamma}}^\top\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}^\top$ respectively. Consequently, the LRT for $H_0 : d_1 = 0$ vs. $H_a : d_1 \neq 0$ (denoted by AMMI-LRT) is given by

$$\Lambda = \Big(\frac{\sum_i \sum_j \hat{\gamma}_{ij}^2 - l_1}{\sum_i \sum_j \hat{\gamma}_{ij}^2}\Big)^{IJ/2}, \quad (2.11)$$

where $l_1 = \hat{d}_1^2$ again is the maximum non-zero (characteristic) root of $\hat{\mathbf{\Gamma}}^\top\hat{\mathbf{\Gamma}}$. That is, l_1 is the maximum value of $(\sum_i \sum_j \alpha_i\beta_j y_{ij})^2$ with respect to α_i and β_j subject to the restriction that $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\sum_i \alpha_i^2 = \sum_j \beta_j^2 = 1$. The critical region for $H_0 : d_1 = 0$ can equivalently be expressed as,

$$\Lambda^* = \frac{l_1}{\sum_{m=1}^{I-1} l_m} = \frac{d_1^2}{\sum_{m=1}^{I-1} d_m^2} > \text{Constant}.$$

Critical points of Λ^* for several choices of I and J are provided previously (Johnson and Graybill, 1972b; Hanumara and Thompson Jr, 1968), which are based on deriving asymptotic property of the ratio of the largest characteristic root to the trace of a Wishart matrix.

In general, the number of components M should be chosen in such a way that the residual ϕ_{ij} represents noise and can again be absorbed with e_{ij} leading to a more powerful test with reduced df. Several studies have investigated cross-validation and significance testing approaches for determining M, the appropriate number of multiplicative interaction terms to be retained (Gauch Jr, 1988; Gauch and Zobel,

1988; Piepho, 1994). When the above model is saturated, $M = I - 1$. Here we focus on AMMI models with $M = 1$ (AMMI1) and do not address the issue of data-adaptive selection of M. A primary reason for this strategy is that analytical power assessments become intractable if another layer of such a data-adaptive selection procedure is implemented. In our NAS data example $M = 1$ component was sufficient.

2.2.3 Biplot

In this section, we discuss the best rank-two approximation to an interaction matrix as presented by a biplot (Gabriel, 1971). The biplot is a graphical planar display of the elements, rows and columns of a matrix. Any matrix of rank two can be displayed as a biplot which is defined through a vector for each row and a vector for each column, such that the inner product represents each matrix element. For a matrix with higher rank, one may use the biplot corresponding to the best rank-two approximation to the original matrix. With the factor analytic representation $\hat{\Gamma} = \hat{A}\hat{D}\hat{B}^{\top}$, each entry of the estimated interaction matrix can be approximated by the first two terms of the corresponding factor representation by

$$\hat{\gamma}_{ij} = \hat{d}_1\hat{\alpha}_{i1}\hat{\beta}_{j1} + \hat{d}_2\hat{\alpha}_{i2}\hat{\beta}_{j2}.$$

For GGI interaction, for example, the matrix of interest is a $(I = 3) \times (J = 3)$ matrix with maximal rank $I - 1 = 2$ and this representation is exact. There are several choices of defining the vectors, we define the points a $P_i = (\hat{d}_1^{1/2}\hat{\alpha}_{i1}, \hat{d}_2^{1/2}\hat{\alpha}_{i2})$ representing row i and the points $Q_j = (\hat{d}_1^{1/2}\hat{\beta}_{j1}, \hat{d}_2^{1/2}\hat{\beta}_{j2})$ describes column j.

Bradu and Gabriel (1978) explained the use of biplots for interaction models. The patterns of the points indicate certain models: additivity (the case of two orthogonal lines), Mandel's row regression model when P_i are collinear and Q_j are scattered or column regression when Q_j are collinear and P_i are scattered. The AMMI model typically will give rise to a configuration where P_i, Q_j are both scattered. For the special

case of AMMI1 the points are not collinear, but co-planar on the three-dimensional plane. We use this representation for repeated measures data with the cell means residual as described before, to visualize the interaction structure.

2.3 Exploring GGI and GEI in the Normative Aging Study

The Normative Aging Study (NAS) is a multidisciplinary longitudinal study of aging in Eastern Massachusetts established by the Veterans Administration in 1963 (Bell et al., 1966). Data were collected every 3-5 years, including extensive physical examination, laboratory, anthropometric, and questionnaire data. The outcome we consider is hearing threshold as measured by pure tone average (PTA) of thresholds at frequencies of 0.5, 1, 2, and 4 kHz. Smaller threshold represents better hearing ability (Cruickshanks et al., 1998). The dataset contained a total of 662 individuals. Each individual had at least two measurements, and 62% of them had at least 4 measurements over time. Descriptive characteristics of the study population is provided in Table 2.1.

We considered two SNPs on genes related to oxidative stress pathway and one environmental exposure, namely, occupational noise. The two genetic markers were rs2071746 (T/A) on it *HMOX-1* (heme-oxygenase 1), a stress response protein which may offer protection against oxidative stress, and rs1001179 (C/T) on *CAT* (catalase), a gene that decomposes hydrogen peroxide. Both of these SNPs have been studied in NAS as an effect modifier in a recent study of black carbon on blood pressure (Mordukhovich et al., 2009). However, the role of these genetic markers related to oxidative stress defense has not been studied for hearing threshold outcomes. An ordinal measure for lifetime exposure to noise with 5 levels (1 reflecting lowest noise exposure and 5 indicating highest) was created based on prior literature (Park et al., 2010). The estimated minor allele frequencies (MAF) for the SNPs considered on *CAT* and *HMOX-1* were 0.19 and 0.46, respectively, and both SNPs were in Hardy-

Weinberg Equilibrium (HWE) ($p = 0.30, 0.67$ respectively). There can be a maximal number of $M = I - 1 = 3 - 1 = 2$ principal interaction factors here and the biplot representation is exact.

2.3.1 *CAT* × *HMOX-1*

The cell means corresponding to the GGI cross-classification, the matrix $\hat{\Gamma}$ and the corresponding SVD, along with the corresponding biplot is presented in the upper panel of Figure 2.1. The plot of cell means suggest evidence for interaction. In the biplot, the points representing the column array appear to be nearly collinear, suggesting possible evidence for Mandel's column regression model. Table 2.2 presents the results from the different fitted models along with a random intercept mixed model (under a compound symmetry covariance) with main effects of both SNPs and pairwise interaction. The interaction is marginally significant in only Mandel's column regression model where the interaction is assumed to be proportional to the main effect of rs2071746 on *HMOX-1* ($p = 0.06$) and not significant in any other model. There is evidence of main effect of *HMOX-1* as well in the column regression model ($p = 0.05$) and from the descriptive statistics.

The AMMI model using the LRT with $M = 1$ has a p-value between 0.1 and 0.2 for the leading principal factor, whereas the pseudo F-test (Mandel, 1971) and used in the AMMI Macro in SAS (Lee, Lee) has a much larger p-value of 0.61. The 5% upper critical value of AMMI-LRT (Johnson and Graybill, 1972a) for a 3×3 array is 0.9994 whereas our observed value is 0.9533. The leading characteristic root of $\hat{\Gamma}'\hat{\Gamma}$, namely, $l_1 = \hat{d}_1^2$ is 6.82 and $l_2 = \hat{d}_2^2 = 0.33$. Since the LRT statistic also represents the fraction of the total variability due to the interaction term explained by the first component, $(\text{LRT} = \hat{d}_1^2/(\hat{d}_1^2 + \hat{d}_2^2))$, we note that the first principal interaction component explains 95% of the interaction sum of squares and the second principal interaction component can be attributed to random noise.

2.3.2 *HMOX-1* × Occupational Noise Exposure

We carried out the same analysis for GEI model with a 3×5 table for *HMOX-1* and occupational noise exposure. The maximal number of interaction factors is still $3 - 1 = 2$. The lower panel of Figure 2.1 displays the cell means corresponding to the GEI cross-classification, the matrix $\hat{\Gamma}$, and the corresponding biplot. No obvious pattern was observed in the cell means plot and biplot. The results of fitting different models for the interaction between *HMOX-1* and noise exposure and fitting a mixed model with random intercepts are also shown in Table 2.2. No main effects of gene, exposure or GEI was detected in any of the models. The AMMI model using the LRT with $M = 1$ has a p-value greater than 0.4 for the leading principal factor, whereas the pseudo F-test used in the AMMI Macro in SAS has a larger p-value of 0.50. The 5% upper critical value of AMMI-LRT from Johnson and Graybill for a 3×5 array is 0.9648 whereas our observed value is 0.8476. The leading characteristic root of $\hat{\Gamma}'\hat{\Gamma}$, equivalently, $l_1 = \hat{d}_1^2$ is 4.14 and $l_2 = \hat{d}_2^2 = 0.74$. Thus only the first principal interaction component explains 85% of the interaction sum of squares and the second principal interaction component explains the remaining noise. A LRT based on fitting the two nested models also supports the same conclusion.

2.4 Simulation Study

We carried out a simulation study to assess the power and Type I error properties of the four tests for interaction (Tukey's one df, Mandel's row and column, and AMMI1). We also considered common epistasis models beyond these four models. In each simulation, we generated individual level data on outcome Y with four repeated measures on each subject for a total of N subjects. The general description of the model is given by (2.1). The structure of γ_{ij} was changed according to the different simulation models. Cell means were first generated and then the vector of observations per individual with given mean and covariance structure were generated from a

multivariate normal distribution. A total of 1000 simulations were used under each setting.

2.4.1 General Two-Way Interaction Models

Design and Parameter Setting

Under each model, for a 3×3 table, the interaction terms were scaled in such a way that they contributed to 15% of the total effect while the remainder is attributed to row and column main effects. We mimicked the simulation as if we had two unlinked causal loci with minor allele frequency (MAF) 0.3 and 0.4 to generate all 3×3 tables. For the larger 9×5 table, we considered combinations of the two loci with MAF 0.3 and 0.4 respectively, along with an environmental exposure with five categories with prevalence 0.2 in each category. For the 9×5 table, interaction terms were scaled to contribute to 20% of the total effect while the rest was attributed to main effects. To assess the Type I error, we generated data with only additive main effects for $N = 1800, 3600$.

Main Results

Figure 2.2 shows the simulation results corresponding to four tests for a 3×3 and a 9×5 table. When the true model is Tukey's one df, Mandel-row and Mandel-column models can capture the interaction structure. AMMI1 is the worst in this setting, but still can detect some interactions. Under Mandel's row model, Tukey's one df and Mandel's column model can not detect any interaction. Again AMMI1 is less powerful. Similar features hold for simulations under Mandel-column model. Under AMMI1, all other alternatives fail to capture the interaction except the true model.

Figure 2.3 presents the percentage of false rejections at 5% significance level. All Type I error rates are inflated. This is due to use of the cell-mean based model and ignoring the unbalanced nature of the data.

In summary, AMMI1 model follows the "mediocrity" principle of not being the

best, but perform reasonably across a spectrum of general interaction models, a robustness feature that is desirable in agnostic search for interaction.

2.4.2 Common Epistasis Models

Design and Parameter Setting

Data were simulated according to 10 common epistasis models (Barhdadi and Dubé, 2010) The left panel in Figure 2.4 gives a visual representation of the interaction pattern with true cell means overlayed. The general model (10) has an arbitrary interaction pattern which was simulated without main effects. MAF for the two loci are still set at 0.3 and 0.4 respectively.

Main Results

The right panel of Figure 2.4 shows that Tukey's model and Mandel's row and column models perform well for epistasis models with main effects (1)-(8). When main effects do not exist (models 9 and 10), AMMI1 is the only model that can detect interaction. Thus to conclude, AMMI model does not appear to be a desirable choice for common epistasis structures compared to the classical models except for the case when there is no main effect of either loci but epistasis is present.

2.5 Discussion

We have made an initial attempt to explore PIA for repeated measures data on quantitative traits. We compared PIA and alternative reduced df tests for interaction and established robustness properties of the AMMI-LRT for repeated measures data via simulation studies across a spectrum of interaction models. Our simulation study indicates that the AMMI test may not be very powerful for common epistasis models unless epistasis occurs in the absence of main effects.

We have primarily concentrated on the AMMI model with $M = 1$ and used the LRT (Johnson and Graybill, 1972a). The F-tests proposed by Gollob (1968) have

significantly inflated Type I error rates compared to LRT based on our simulations and thus have not been presented in numerical results. The use of pseudo F-tests needs to be investigated further for longitudinal data.

The cell-mean based approach can be viewed as a screening tool or an exploratory idea about the interaction structure and longitudinal effects. In that sense, PIA for this problem provides an exploratory analysis of interaction structures. The idea of first fitting additive terms and then representing the residual matrix via a sparse decomposition appears to be a promising approach to study interaction. Further development of likelihood-based estimation approaches with asymptotic theory are warranted to properly account for repeated measures data.

Table 2.1: Descriptive characteristics of 662 study participants in the Normative Aging Study considered in our data analysis. Age, BMI, Health Status and Smoking variables are measured at baseline. PTA hearing threshold is averaged over all repeated measures.

Variable	Mean	SD
PTA hearing threshold (dB) (Y)	10.86	6.54
Age (years)	41.66	8.77
Body Mass Index (kg/m2)	25.71	2.76
	N	Percent
Race (white)	645	97.43
Education ($>$ 12 years)	381	57.55
Type-2 Diabetes	13	1.96
Hypertension	28	4.23
Pack-Years of Cigarettes		
0	205	30.97
<30	336	50.76
≥ 30	121	18.28
Genes (G)		
CAT(C/T) rs1001179		
CC	403	65.96
CT	179	29.3
TT	29	4.75
HMOX-1(T/A) rs2071746		
TT	171	27.67
TA	320	51.78
AA	127	20.55
Environment (E)		
Level of Noise Exposure		
1	120	18.13
2	95	14.35
3	182	27.49
4	153	23.11
5	112	16.92
Number of Repeated Measures on PTA Per Subject		
2	129	19.49
3	122	18.43
4	155	23.41
5	147	22.21
6	85	12.84
7	20	3.02
8	4	0.60

Table 2.2: Analysis results for gene-gene interaction and gene-environment interaction in the Normative Aging Study. Two SNPs, rs2071746 on *HMOX-1* and rs1001179 on *CAT* gene are considered for GGI analysis. The GEI analysis considers the interaction between the same SNP on *HMOX-1* and occupational noise exposure. Results from the four models Tukey's one df, Mandel's Row, Mandel's Column, and Principal interaction analysis via AMMI1 are presented.

Model	Hypothesis	Numerator df	F	p-value (cell mean)
Analysis results for *CAT*(C/T) × *HMOX-1*(T/A)				
Tukey's one df for Non-additivity	$H_0 : \theta$=0	1	0.87	0.20
Mandel's Row(*CAT*)-Regression	$H_0 : \eta_j = 0$	2	0.91	0.52
Mandel's Column(*HMOX-1*)-Regression	$H_0 : \lambda_i = 0$	2	14.84	0.06
AMMI	First PI	3.57	F* = 0.11	0.61
AMMI	First PI		LRT = 0.95	$0.1 < P < 0.2$
Mixed Model(random intercept, saturated)	*CAT*×*HMOX-1*	4	1.07	0.37
Analysis results for *HMOX-1*(T/A) × Noise Exposure				
Tukey's one df for Nonadditivity	$H_0 : \theta = 0$	1	1.06	0.34
Mandel's Row(*HMOX-1*)-Regression	$H_0 : \eta_j = 0$	4	0.19	0.93
Mandel's Column(Noise)-Regression	$H_0 : \lambda_i = 0$	2	0.97	0.43
AMMI	First PI	6.36	F* = 1.43	0.50
AMMI	First PI		LRT = 0.85	P >0.40
Mixed Model(random intercept, saturated)	*HMOX-1*×Noise	8	0.61	0.77

F* = Pseudo F Value with fractional DF (Mandel, 1971).
LRT is the likelihood ratio test statistic (Johnson and Graybill, 1972a).

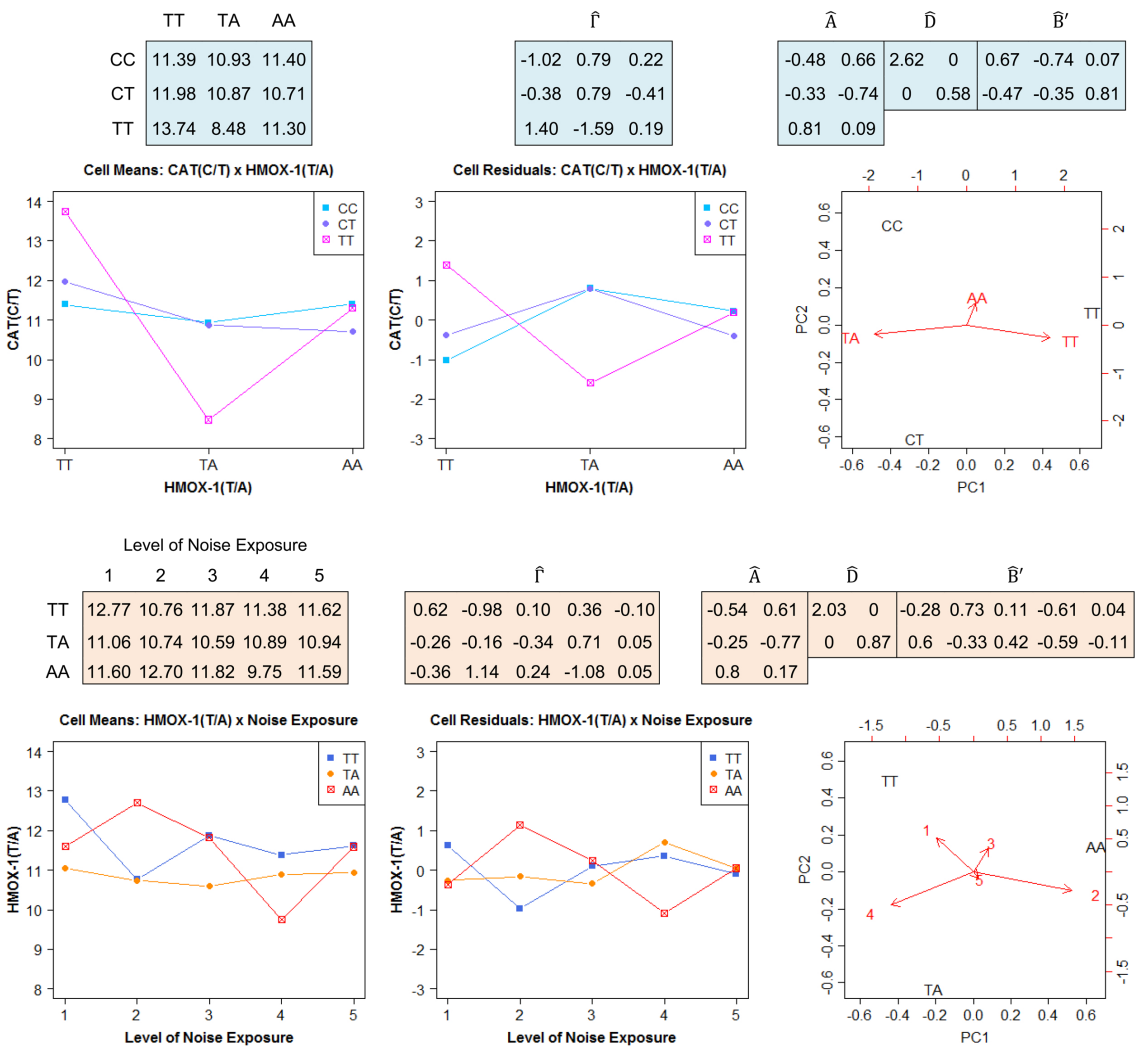


Figure 2.1: Cell means, residuals after eliminating additive row and column main effects and the SVD of the estimated $\hat{\mathbf{\Gamma}}$ matrix for GGI (top panel) and GEI (bottom panel) analyses. The numerical arrays are accompanied by graphical displays of the cell means, entries of $\hat{\mathbf{\Gamma}}$ and the biplot representation. Results are based on the Normative Aging Study data.

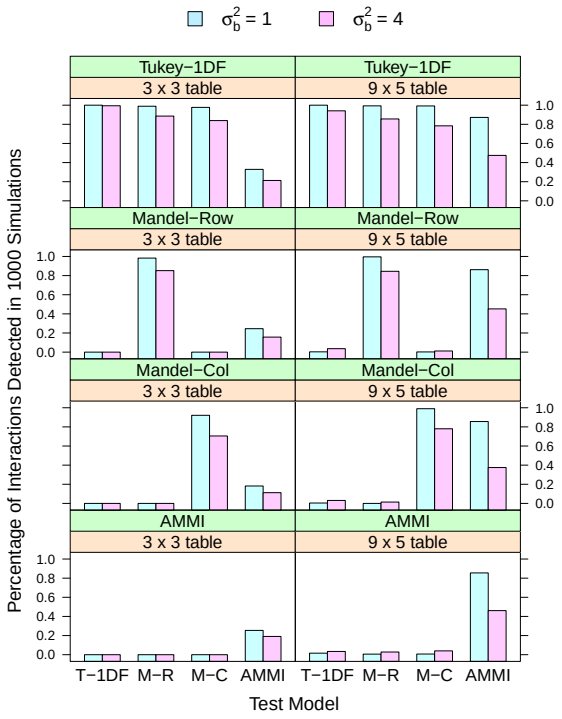


Figure 2.2: Percentage of interactions detected by each of the four tests in the simulation settings corresponding to 3×3 and 9×5 array from 1000 simulated datasets with $N = 3600$. The top label within each box represents the true simulation model whereas the horizontal axis labels indicate the tests. The error variance σ_e^2 is set at 4 in all cases.

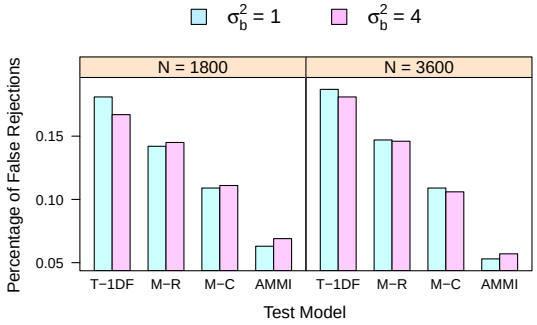


Figure 2.3: Empirical estimates of Type I error rates corresponding to the four interaction tests in a 3×3 array setting. Data are generated under additive model which has only main effects.

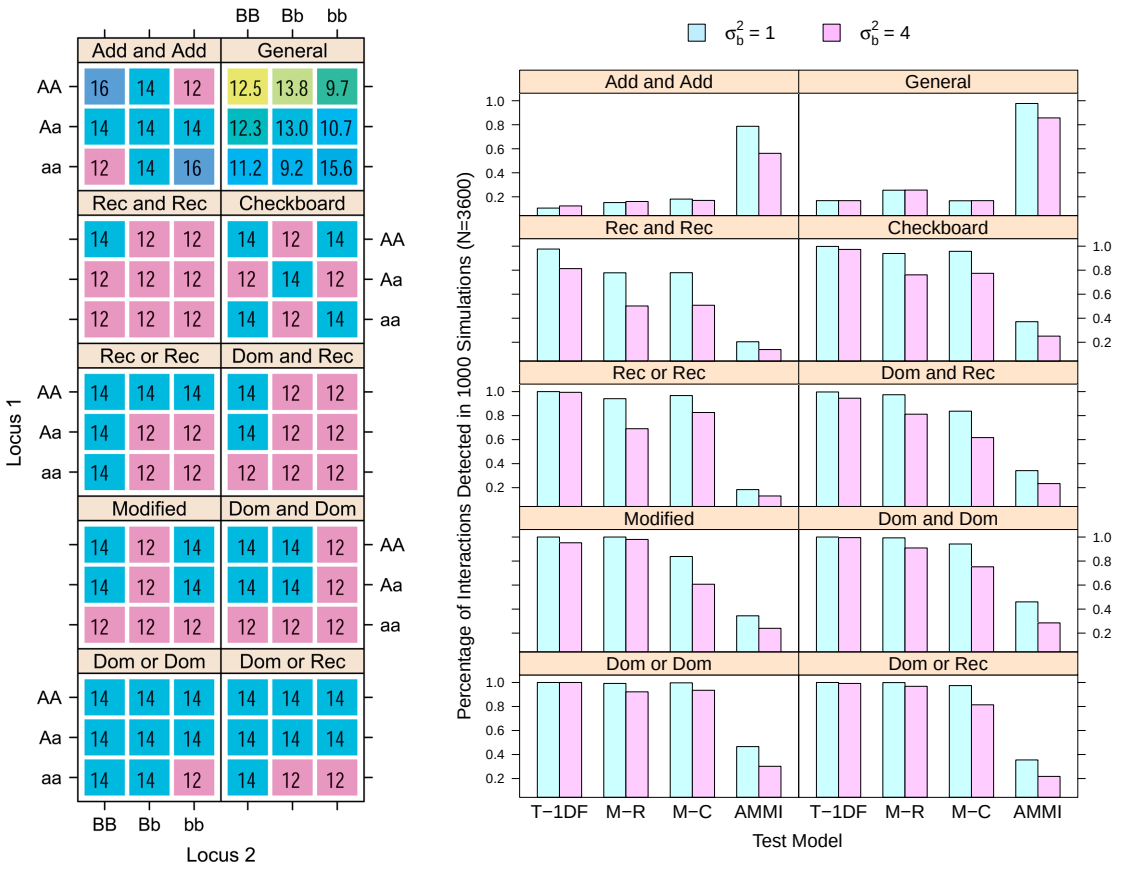


Figure 2.4: Percentage of interactions detected by each of the four tests in 1000 simulated datasets under 10 common epistasis models. The true models with cell means are displayed in different colors on the left hand panel. The top label within each box represents the true simulation model whereas the horizontal axis labels indicate the tests. The error variance σ_e^2 is set at 4 in all cases.

2.6 Appendix

2.6.1 Tables of Analysis of Variance

Table 2.3: ANOVA Table corresponding to Tukey's one df non-additivity test

Source	SS	df	F
Mean	$SSm = IJ\hat{\mu}^2$	1	
Row Main	$SS_R = J\sum_i \hat{R_i}^2$	$I-1$	MS_R/MSE
Column Main	$SS_C = I\sum_j \hat{C_j}^2$	$J-1$	MS_C/MSE
Interaction	$SS_{ll} = \hat{\theta}^2 \sum_i \sum_j \hat{R_i}^2 \hat{C_j}^2$	1	MS_{ll}/MSE
Error	$SSE = SST - SSm - SS_R - SS_C - SS_{ll}$	$(I-1)(J-1)-1$	
Total	$SST = \sum_i \sum_j y_{ij}^2$	IJ	

Table 2.4: ANOVA Table corresponding to Mandel's column-regression model

Source	SS	df	F
Mean	$SSm = IJ\hat{\mu}^2$	1	
Row Main	$SS_R = J\sum_i \hat{R_i}^2$	$I-1$	$MS_R/MSres$
Column Main	$SS_C = I\sum_j \hat{C_j}^2$	$J-1$	$MS_C/MSres$
Column Slopes	$SSs = \sum_i \sum_j \hat{\lambda_i}^2 \hat{C_j}^2$	$I-1$	$MSs/MSres$
Residuals	$SSres = SST - SSm - SS_R - SS_C - SSs$	$(I-1)(J-1) -(I-1)$	
Total	$SST = \sum_i \sum_j y_{ij}^2$	IJ	

Table 2.5: ANOVA Table corresponding to Tukey's row-column model

Source	SS	df	F
Mean	$SSm = IJ\hat{\mu}^2$	1	
Row Main	$SS_R = J\sum_i \hat{R_i}^2$	$I-1$	$MS_R/MSres$
Column Main	$SS_C = I\sum_j \hat{C_j}^2$	$J-1$	$MS_C/MSres$
Linear by Linear	$SS_{ll} = \hat{\theta}^2 \sum_i \sum_j \hat{R}_i^2 \hat{C}_j^2$	1	$MS_{ll}/MSres$
Row Slopes	$SS_{RS} = \sum_i \sum_j \hat{\lambda_i}^2 \hat{C_j}^2$	$I-2$	$MS_{RS}/MSres$
Column Slopes	$SS_{CS} = \sum_i \sum_j \hat{R_i}^2 \hat{\eta_j}^2$	$J-2$	$MS_{CS}/MSres$
Residuals	$SSres = SST - SSm - SS_R$ $-SS_C - SS_{ll} - SS_{RS} - SS_{CS}$	$(I-1)(J-1)-1$ $-(I-2)-(J-2)$	
Total	$SST = \sum_i \sum_j y_{ij}^2$	IJ	

Table 2.6: ANOVA Table corresponding to Gollob's F test for the AMMI model with $M < (I-1)$ components, $I < J$.

Source	SS	df	F
Mean	$SSm = IJ\hat{\mu}^2$	1	
Row Main	$SS_R = J\sum_i \hat{R_i}^2$	$I-1$	$MS_R/MSres$
Column Main	$SS_C = I\sum_j \hat{C_j}^2$	$J-1$	$MS_C/MSres$
Row by Column	$SS_{RC} = \sum_i \sum_j y_{ij}^2 - SSm$ $-SS_R - SS_C$	$(I-1)(J-1)$	
Fm	$SS_{Fm} = \hat{d}_m^2$	$I+J-1-2m$	$MS_{Fm}/MSres$
$Fres$	$SS_{Fres} = SS_{RC} - \sum_{m=1}^{M} SS_{Fm}$	$(I-1-M)\times$ $(J-1-M)$	
Total	$SST = \sum_i \sum_j y_{ij}^2$	IJ	

2.6.2 Maximum Likelihood Estimation

This section shows derivation of maximum likelihood estimators (MLEs) for parameters in the four models listed previously. Define the parameter vector $\nu^\top =$

$(\mu, R^\top, C^\top, \gamma^\top, \sigma^2)$, the likelihood can then be expressed as

$$
\begin{aligned}
L(\nu) &= (\frac{1}{2\pi\sigma^2})^{IJ/2} \exp\{-\frac{1}{2\sigma^2} \sum_i \sum_j (y_{ij} - \mu - R_i - C_j - \gamma_{ij})^2\} \\
&= (\frac{1}{2\pi\sigma^2})^{IJ/2} \exp\{-\frac{1}{2\sigma^2} \sum_i \sum_j [z_{ij} + (y_{..} - \mu) + (y_{i.} - y_{..} - R_i) + (y_{.j} - y_{..} - C_j) - \gamma_{ij}]^2\} \\
&\quad \text{where, } (z_{ij} = y_{ij} - y_{i.} - y_{.j} + y_{..}) \\
&= (\frac{1}{2\pi\sigma^2})^{IJ/2} \exp\{-\frac{1}{2\sigma^2} [IJ(y_{..} - \mu)^2 + J \sum_i (y_{i.} - y_{..} - R_i)^2 + I \sum_j (y_{.j} - y_{..} - C_j)^2 \\
&\quad + \sum_i \sum_j (z_{ij} - \gamma_{ij})^2]\} \\
&\leq (\frac{1}{2\pi\sigma^2})^{IJ/2} \exp\{-\frac{1}{2\sigma^2} [\sum_i \sum_j (z_{ij} - \gamma_{ij})^2]\}
\end{aligned}
$$

The maximum value of $L(\nu)$ is attained when

$$
\begin{aligned}
\hat{\mu} &= y_{..} \\
\hat{R}_i &= y_{i.} - y_{..}, i = 1, \ldots, I \\
\hat{C}_j &= y_{.j} - y_{..}, j = 1, \ldots, J
\end{aligned}
$$

With the above estimates $\hat{\mu}, \hat{R}, \hat{C}$ substituted in the likelihood, the log likelihood that needs to be maximized as a function of γ_{ij} is simply,

$$
l^*(\gamma) = -\sum_i \sum_j \gamma_{ij}^2 + 2(\sum_i \sum_j \gamma_{ij} z_{ij}). \tag{2.12}
$$

Special Cases: Tukey 1-df: In this case, $\gamma_{ij} = \theta R_i C_j$ (Tukey 1949),

$$
\begin{aligned}
l^*(\theta) &= -\theta^2 \sum_i \sum_j \hat{R}_i^2 \hat{C}_j^2 + 2\theta \sum_i \sum_j \hat{R}_i \hat{C}_j z_{ij} \\
\frac{\partial l^*(\theta)}{\partial \theta} &= -2\theta \sum_i \sum_j \hat{R}_i^2 \hat{C}_j^2 + 2 \sum_i \sum_j \hat{R}_i \hat{C}_j z_{ij} = 0 \\
\hat{\theta} &= \frac{\sum_i \sum_j \hat{R}_i \hat{C}_j z_{ij}}{\sum_i \sum_j \hat{R}_i^2 \hat{C}_j^2}
\end{aligned}
$$

Mandel's Column Model: In this case, $\gamma_{ij} = \lambda_i C_j$ (Mandel 1961),

$$\begin{aligned} l^*(\lambda) =& -\sum_i \sum_j \lambda_i^2 \hat{C}_j^2 + 2\sum_i \sum_j \lambda_i \hat{C}_j z_{ij} \\ \frac{\partial l^*(\lambda)}{\partial \lambda_i} =& -2\lambda_i \sum_j \hat{C}_j^2 + 2\sum_j \hat{C}_j z_{ij} = 0 \\ \hat{\lambda}_i =& \frac{\sum_j \hat{C}_j z_{ij}}{\sum_j \hat{C}_j^2} \end{aligned}$$

Tukey's Row-Column Model: If $\gamma_{ij} = \theta R_i C_j + \lambda_i C_j + R_i \eta_j$ (Tukey 1962),

$$\begin{aligned} l^*(\theta,\lambda,\eta) =& -\sum_i \sum_j (\theta \hat{R}_i \hat{C}_j + \lambda_i \hat{C}_j + \hat{R}_i \eta_j)^2 + 2\sum_i \sum_j (\theta \hat{R}_i \hat{C}_j + \lambda_i \hat{C}_j + \hat{R}_i \eta_j) z_{ij} \\ =& -\sum_i \sum_j [\theta^2 \hat{R}_i^2 \hat{C}_j^2 + 2\theta \hat{R}_i \hat{C}_j (\lambda_i \hat{C}_j + \hat{R}_i \eta_j) + (\lambda_i \hat{C}_j + \hat{R}_i \eta_j)^2] \\ & + 2\sum_i \sum_j (\theta \hat{R}_i \hat{C}_j + \lambda_i \hat{C}_j + \hat{R}_i \eta_j) z_{ij} \\ \frac{\partial l^*(\theta,\lambda,\eta)}{\partial \theta} =& -2\theta \sum_i \sum_j \hat{R}_i^2 \hat{C}_j^2 + 2\sum_i \sum_j \hat{R}_i \hat{C}_j z_{ij} = 0 \\ \hat{\theta} =& \frac{\sum_i \sum_j \hat{R}_i \hat{C}_j z_{ij}}{\sum_i \sum_j \hat{R}_i^2 \hat{C}_j^2} \\ \frac{\partial l^*(\theta,\lambda,\eta)}{\partial \lambda_i} =& -2\lambda_i \sum_j \hat{C}_j^2 + 2\hat{R}_i \sum_j \hat{C}_j \eta_j + 2\theta \hat{R}_i \sum_j \hat{C}_j^2 - 2\sum_j \hat{C}_j z_{ij} = 0 \\ \hat{\lambda}_i =& \frac{\sum_j \hat{C}_j z_{ij} - \hat{\theta} \hat{R}_i \hat{C}_j^2}{\sum_j \hat{C}_j^2} = \frac{\sum_j \hat{C}_j \hat{z}_{ij}}{\sum_j \hat{C}_j^2} - \hat{\theta} \hat{R}_i \\ \frac{\partial l^*(\theta,\lambda,\eta)}{\partial \eta_j} =& -2\eta_j \sum_i \hat{R}_i^2 + 2\hat{C}_j \sum_i \hat{R}_i \lambda_i + 2\theta \hat{C}_j \sum_i \hat{R}_i^2 - 2\sum_i \hat{R}_i z_{ij} = 0 \\ \hat{\eta}_j =& \frac{\sum_i \hat{R}_i z_{ij}}{\sum_i \hat{R}_i^2} - \hat{\theta} \hat{C}_j \end{aligned}$$

Principal Interactions or AMMI Model: In this case, $\gamma_{ij} = d_1 \alpha_i \beta_j$. Johnson and Graybill (1972) showed that the maximum likelihood for AMMI model with

$M = 1$ is attained when the MLE $\hat{d_1}^2$ is equal to the largest root of $Z^\top Z$ (l_1). We present the following simpler argument to derive MLEs for the interaction terms.

Note that in that case, from (2.12),

$$l^*(d_1, \alpha_i, \beta_j) = -d_1^2 \sum_i \alpha_i^2 \sum_j \beta_j^2 + 2d_1 \sum_i \sum_j \alpha_i \beta_j z_{ij}.$$

Under the constraints, $\sum_i \alpha_i^2 = \sum_j \beta_j^2 = 1$, this reduces to maximizing $(\sum_i \sum_j \alpha_i \beta_j z_{ij})^2$ subject to the normalization constraints. It is well-known from the eigen-theory of matrices that this quadratic form is maximized by the values of $\{\alpha_i, \beta_j\}$ that are give by the left and right normalized characteristic vectors corresponding to $Z'Z$ and ZZ' where Z has the (i, j)-th entry $z_{ij} = \hat{\gamma}_{ij} = y_{ij} - y_{i.} - y_{.j} + y_{..}$. The maximum value of $(\sum_i \sum_j \alpha_i \beta_j z_{ij})^2$ is l_1. With these estimates of α_i and β_j, it can be seen easily seen from the expression of $l^*(d_1, \alpha_i, \beta_j)$, that $\hat{d}_1^2 = l_1$. Hence the proof.

CHAPTER III

Novel Likelihood Ratio Tests for Screening Gene-Gene and Gene-Environment Interactions with Unbalanced Repeated-Measures Data

3.1 Introduction

Prospective cohort studies examining gene-gene or gene-environment interaction (GGI or GEI) effects on disease-related quantitative traits have received considerable attention in recent years (Bookman et al., 2011; Fan et al., 2012). The detection of GEI plays a critical role in identifying a sub-population of the genetically susceptible individuals that are strongly affected by an adverse exposure. A better understanding of GEI may lead to the development of more effective disease prevention and intervention strategies. Studies of GEI in relation to disease development are facilitated by life-time characterization of exposure data, which are often available in prospective cohort studies. Repeated measures design in a prospective cohort study may increase power to detect interaction effects (Wong et al., 2003) and provide better ways to handle exposure measurement error. In addition, repeated measures data provide valuable information for delineating potentially time-dependent form of GGI or GEI, thereby permitting a much more detailed assessment of the dynamic interplay between genes and environment.

Cohort studies for GGI or GEI are typically characterized by unequal sample size in each genotype-genotype or genotype-exposure configuration as a result of unbal-

anced allele frequencies and heterogeneous environmental exposure distributions in a population. A common analysis strategy for such unbalanced data involves modeling GGI or GEI by a product term in a regression setting, implying that the effect of two factors may not be purely additive in their contribution to the quantitative trait. Alternatively, one can try to model the interaction term in the generalized additive mixed model framework with nonlinear exposure and time effects (Lin and Zhang, 1999), but tests for such non-parametric, smoothed interaction terms may yield reduced power for moderate sample size. Therefore, flexible yet parsimonious modeling of GGI or GEI is of interest in the longitudinal setting. In this paper, we propose likelihood ratio tests (LRT) for GGI and GEI using a sparse representation of interaction borrowing ideas from the classical ANOVA literature.

Genetic factors (G) and environmental exposures (E) are frequently treated as binary or ordered categorical variables. Consequently, GGI and GEI are often analyzed in the form of a two-way table. Considering G as a row variable with I categories and E as a column variable with J categories, the mean structure of a general two-way classification model for analyzing row $\times$ column interactions is given by

$$\mu_{ij} = \mu + R_i + C_j + \gamma_{ij},\ i = 1, \ldots, I,\ j = 1, \ldots, J, \tag{3.1}$$

where μ_{ij} is the the expected (mean) value of a quantitative trait corresponding to the ith row and the jth column, μ is the grand mean, R_i is the additive main effect of the ith row, C_j is the additive main effect of the jth column, and γ_{ij} is the non-additive effect of the ith row and the jth column. The sum-to-zero conditions,

$$\sum_i R_i = \sum_j C_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0, \tag{3.2}$$

ensure identifiability of the parameters in (3.1), so the degrees of freedom (df) for testing γ_{ij} in a fully saturated model is $(I-1)(J-1)$. While a saturated model (3.1) is flexible for estimation of γ_{ij}, the df for interaction tests can increase considerably

for finely cross-classified tables, which is inefficient and may result in low power for detecting GGI or GEI.

To improve the power of the test for GGI and GEI in longitudinal cohort studies, we explore alternative parsimonious interaction structures that were proposed in the classical ANOVA literature for testing interaction with only one observation per cell. Several models are summarized in the following:

$$\text{Model (a): } \mu_{ij} = \mu + R_i + C_j + \theta R_i C_j \text{ (Tukey, 1949)}$$

$$\text{Model (b): } \mu_{ij} = \mu + R_i + C_j + \lambda_i C_j \text{ (Mandel, 1961)}$$

$$\text{Model (c): } \mu_{ij} = \mu + R_i + C_j + R_i \eta_j \text{ (Mandel, 1961)}$$

$$\text{Model (d): } \mu_{ij} = \mu + R_i + C_j + \theta R_i C_j + \lambda_i C_j + R_i \eta_j \text{ (Tukey, 1962)}$$

with constraints $\sum_{i=1}^{I} \lambda_i = \sum_{j=1}^{J} \eta_j = 0$ for models (b) and (c), respectively, and additional constraints $\sum_{i=1}^{I} \lambda_i R_i = \sum_{j=1}^{J} \eta_j C_j = 0$ for model (d). The multiplicative interaction term is proportional to the main effects of one or both factors. The null hypotheses of no interaction for models (a)–(d) are $\theta = 0, \lambda_i = 0$ $(i = 1, \ldots, I-1), \eta_j = 0$ $(j = 1, \ldots, J-1)$, and $\theta = \lambda_i = \eta_j = 0$ $(i = 1, \ldots, I-2; j = 1, \ldots, J-2)$, corresponding to $1, I-1$, $J-1$, and $I+J-3$ df for the tests of interaction effects, respectively. A more flexible potential alternative is the additive main effects and multiplicative interactions (AMMI) model (Gollob, 1968; Mandel, 1971)

$$\text{Model (e): } \mu_{ij} = \mu + R_i + C_j + \sum_{m=1}^{M} d_m \alpha_{im} \beta_{jm} + \gamma_{ij}^*,$$

where M represents the number of interaction factors being extracted, $M \leq \min(I-1, J-1)$, and a residual γ_{ij}^* remains if not all interaction factors are used. The terms $\{\alpha_{im}\beta_{jm}\}$ can the considered as the weights corresponding to a multiplicative contrast among $\{\gamma_{ij}\}$ with $\sum_i \alpha_{im} = \sum_j \beta_{jm} = 0$ and $\sum_i \alpha_{im}\alpha_{im'} = \sum_j \beta_{jm}\beta_{jm'} = 0$ for $m \neq m'$. For normalized contrasts ($\sum_i \alpha_{im}^2 = \sum_j \beta_{jm}^2 = 1$), $\{d_m, \alpha_{im}, \beta_{jm}\}$ can be obtained by applying singular value decomposition (SVD) to $\{\gamma_{ij}\}$. Since the

motivation for using an AMMI model is to extract a low rank approximation to the interaction matrix to save df and thus to enhance efficiency for the test, we focus on AMMI models with $M = 1$ (AMMI1). For all subsequent discussions, model (e) refers to AMMI1 model. The null hypothesis of no interaction for AMMI1 model is $H_0 : d_1 = 0$.

Models (a)–(e) were conceived from a statistical objective of reducing df and enhancing power of tests for interaction. They have been used in designed genotype-by-environment yield trials in agricultural studies (Freeman et al., 1973; Zobel et al., 1988; Crossa et al., 1990). These models were not conceived from a mechanistic or human biological perspective. Model (a) has recently been used to test for genetic effects in case-control studies (Chatterjee et al., 2006) and repeated measures data of complex traits (Maity et al., 2009). Models (a), (b), (c), and (e) have also been applied for GGI effects on quantitative traits in cross-sectional studies (Barhdadi and Dubé, 2010). In unbalanced designs, the sums of squares associated with the two factors and their interaction are not orthogonal to one another. Consequently, the difficulties that arise in applying these nonlinear interaction models to unbalanced data involve obtaining unbiased parameter estimates, partitioning the sums of squares, deriving the appropriate test statistics and their null distributions. Mukherjee et al. (2012) proposed a screening tool for GGI and GEI using cell means from an unbalanced repeated measures array. This approach is appealing due to a closed-form analytical expression of the test statistic. However, violations in the homoscedasticity assumption of cell mean error distributions result in inflated type I error. While their proposed resampling-based method recognizes unbalanced, repeated measures data structure, the test implemented for AMMI models lacks power because it was not based on a theoretically derived pivot but an ad hoc extension of the balanced, cross-sectional case.

To overcome some limitations of the previous methods, we propose alternate ap-

proaches to explore GGI and GEI using models (a)–(e). We first describe our improved cell-mean approach that properly handles unbalanced data. Specifically, we adapt and modify the test proposed by Boik (1989) under a reduced-rank model for application to GGI/GEI using AMMI models. Next, we extend models (a)–(e) to the repeated measures setting using a mixed-effects modeling framework. We then develop a parametric bootstrap resampling approach by replacing the ad hoc pivot in Mukherjee et al. (2012) with a LRT-based pivot derived from the maximum likelihood under a nonlinear mixed-effects model. The power and type I error of our proposed tests are examined through a series of simulation studies. Lastly, we apply the proposed methods to a GEI study concerning the modifying effects of polymorphisms in the hemochromatosis gene (*HFE*) on the association between cumulative lead exposure and pulse pressure (Zhang et al., 2010). Subject-specific and time-specific contributions to GEI are investigated using outputs from the AMMI model.

3.2 Methods

3.2.1 Likelihood Ratio Test based on Cell Means

Following the notations in (3.1), let y_{ijkh} be the hth measurement corresponding to the kth individual in the (i,j)th cell (or equivalently, row i and column j) in a longitudinal cohort study, $i = 1,\ldots,I, j = 1,\ldots,J, k = 1,\ldots,N_{ij}, h = 1,\ldots,n_{ijk}$. Let N denote the total number of individuals, $N = \sum_i \sum_j N_{ij}$. Let $\bar{\boldsymbol{Y}} = \{\bar{Y}_{ij}\}$ be the $I \times J$ matrix of sample means with $\bar{Y}_{ij} = \sum_{k=1}^{N_{ij}} \sum_{h=1}^{n_{ijk}} y_{ijkh} / \sum_{k=1}^{N_{ij}} n_{ijk}$. Let $\boldsymbol{L}$ be the matrix of main effects, parameterized as $\boldsymbol{L} = \mathbf{1}_I \mu \mathbf{1}_J' + \boldsymbol{R}\mathbf{1}_J' + \mathbf{1}_I \boldsymbol{C}'$, where $\mathbf{1}_\nu$ is a length-ν vector of ones, and $\boldsymbol{R} = (R_1,\ldots,R_I)'$, and $\boldsymbol{C} = (C_1,\ldots,C_J)'$ are the parameter vectors representing row and column effects, respectively. Let $\boldsymbol{\Gamma}$ be the $I \times J$ matrix of interaction effects, so the mean structures of models (a)–(e) can be expressed as $\boldsymbol{E}(\bar{\boldsymbol{Y}}) = \boldsymbol{L} + \boldsymbol{\Gamma}$. Throughout our treatment of the problem, we consider the drop-outs in longitudinal studies to be missing at random, leading to the

unbalanced data structure.

We propose to use an empirical variance estimate for the variance of $\bar{Y}_{ij}$ (denoted as δ_{ij}^2) that accounts for within-subject correlation. Let $\sigma^2 \boldsymbol{P}(\boldsymbol{\rho})$ be a symmetric $n_{ijk} \times n_{ijk}$ within-subject covariance matrix, where $\boldsymbol{\rho}$ is a $s \times 1$ parameter vector that fully characterizes the correlation matrix $\boldsymbol{P}(\boldsymbol{\rho})$, and σ^2 is a scale parameter. Both $\boldsymbol{\rho}$ and σ^2 can be estimated by Pearson residuals, namely, $\hat{r}_{ijkh} = y_{ijkh} - \bar{y}_{ijk}$ (Liang and Zeger, 1986). The pooled estimate of σ^2 is

$$\hat{\sigma}^2 = \sum_i \sum_j (N_{ij} - 1)\hat{\sigma}_{ij}^2 / (\sum_i \sum_j N_{ij} - IJ), \text{ where } \hat{\sigma}_{ij}^2 = \sum_{k=1}^{N_{ij}} \sum_{h=1}^{n_{ijk}} \hat{r}_{ijkh}^2 / (\sum_{k=1}^{N_{ij}} n_{ijk} - 1).$$

The estimation of $\boldsymbol{\rho}$ is conditional on the correlation structure. For a compound symmetric correlation structure, $s = 1$, $\text{corr}(y_{ijkh}, y_{ijkh'}) = \rho$ for $h \neq h'$. The pooled estimate for ρ is

$$\hat{\rho} = \sum_i \sum_j (N_{ij} - 1)\hat{\rho}_{ij} / (\sum_i \sum_j N_{ij} - IJ),$$
$$\text{where } \hat{\rho}_{ij} = \sum_{k=1}^{N_{ij}} \sum_{h>h'} \hat{r}_{ijkh}\hat{r}_{ijkh'} / \big\{\hat{\sigma}^2 \big[\sum_{k=1}^{N_{ij}} \frac{1}{2} n_{ijk}(n_{ijk} - 1) - 1 \big] \big\}.$$

Finally, the empirical variance estimate for $\bar{Y}_{ij}$ is given by

$$\hat{\delta}_{ij}^2 = \frac{\hat{\sigma}^2}{n_{ij}} + \frac{2\hat{\sigma}^2}{n_{ij}^2} \sum_{k=1}^{N_{ij}} \sum_{h>h'} \hat{\text{corr}}(y_{ijkh}, y_{ijkh'}), \quad \text{where } n_{ij} = \sum_{k=1}^{N_{ij}} n_{ijk}. \tag{3.3}$$

Given $\hat{\delta}_{ij}$, we maximize the likelihood of $\bar{\boldsymbol{Y}}$ under the normality assumption on the cell means, namely, $\text{Vec}(\bar{\boldsymbol{Y}}) \sim \mathcal{N}\big(\text{Vec}(\boldsymbol{L}) + \text{Vec}(\boldsymbol{\Gamma}), \text{Diag}(\hat{\delta}_{ij}^2)\big)$. Maximizing the log-likelihood is equivalent to least squares fitting of μ_{ij} subject to weights $1/\hat{\delta}_{ij}^2$. For classical interaction models (a)–(d) involving nonlinearity in the parameters, the maximum likelihood (ML) estimates for $\boldsymbol{L}$ and $\boldsymbol{\Gamma}$ are obtained using a quasi-Newton method in R (R Core Team, 2012) with function 'optim' and L-BFGS-B algorithm (Nocedal and Wright, 1999). Quasi-Newton methods are sequential line search algorithms, and generally require only the gradient of the objective to be computed at

each iterate. When convergence is reached, we calculate the log-likelihood under the null ($\hat{\ell}_0$) and under the alternative ($\hat{\ell}_1$) to construct the LRT statistic: $-2(\hat{\ell}_0 - \hat{\ell}_1)$. Under H_0, the LRT statistic approximately follows a central chi-square distribution with df $= 1, I-1, J-1$, and $I+J-3$ for models (a), (b), (c), and (d), respectively. The comparison of empirical quantiles of the LRT statistics with chi-square quantiles is presented in Appendix (Figure 3.7).

Boik (1989) proposed the likelihood ratio criterion to test the rank of $\boldsymbol{\Gamma}$ for unbalanced data without repeated measures. For AMMI1 models, the test for non-additivity is $H_0 : d_1 = 0$ vs. $H_a : d_1 = 1$, which is equivalent to H_0: rank$(\boldsymbol{\Gamma}) = 0$ vs. H_a: rank$(\boldsymbol{\Gamma}) = 1$. Let $\boldsymbol{H}_\nu$ be the row-space or column-space projection operator, $\boldsymbol{H}_\nu = \boldsymbol{I}_\nu - (1/\nu)\mathbf{1}_\nu\mathbf{1}_\nu'$ ($\nu = I, J$), and let $\boldsymbol{K}_\nu\boldsymbol{K}_\nu'$ be a full-rank factorization of $\boldsymbol{H}_\nu$ with dimension $\nu \times (\nu - 1)$ satisfying $\boldsymbol{K}_\nu'\boldsymbol{K}_\nu = \boldsymbol{I}$. We have

$$\boldsymbol{\Gamma} = \boldsymbol{H}_I\boldsymbol{\Gamma}\boldsymbol{H}_J = \boldsymbol{K}_I\boldsymbol{K}_I'\boldsymbol{\Gamma}\boldsymbol{K}_J\boldsymbol{K}_J' = \boldsymbol{K}_I\boldsymbol{\Phi}\boldsymbol{K}_J', \ \boldsymbol{\Phi} = \boldsymbol{K}_I'\boldsymbol{\Gamma}\boldsymbol{K}_J,$$

with rank$(\boldsymbol{\Gamma})$ = rank$(\boldsymbol{\Phi}) = r \le p = \min(I-1, J-1)$. The elements of $\boldsymbol{\Phi}$ form a basis for the set of interaction contrasts. Define $\boldsymbol{K} = (\boldsymbol{K}_J \otimes \boldsymbol{K}_I)$ so that $\boldsymbol{K}'\text{Vec}(\bar{\boldsymbol{Y}})$ is a linear function of Vec$(\bar{\boldsymbol{Y}})$ without containing the main effects (because $\boldsymbol{K}'\text{Vec}(\boldsymbol{L}) = 0$). Hence, $\boldsymbol{E}[\boldsymbol{K}'\text{Vec}(\bar{\boldsymbol{Y}})] = \boldsymbol{K}'\text{Vec}(\boldsymbol{\Gamma}) = \text{Vec}(\boldsymbol{\Phi})$ and var$(\boldsymbol{K}'\text{Vec}(\bar{\boldsymbol{Y}})) = \boldsymbol{K}'\text{Diag}(\delta_{ij}^2)\boldsymbol{K}$.

The goal is to maximize the likelihood function of $\bar{\boldsymbol{Y}}$ subject to the constraint rank$(\boldsymbol{\Gamma}) = r$, which is the same as computing

$$S(r) = \min_{\text{rank}(\boldsymbol{\Phi})=r} \big[\boldsymbol{K}'\text{Vec}(\bar{\boldsymbol{Y}}) - \text{Vec}(\boldsymbol{\Phi})\big]'\boldsymbol{W}^{-1}\big[\boldsymbol{K}'\text{Vec}(\bar{\boldsymbol{Y}}) - \text{Vec}(\boldsymbol{\Phi})\big], \tag{3.4}$$

where $\boldsymbol{W} = \boldsymbol{K}'\text{Diag}(\delta_{ij}^2)\boldsymbol{K}$, and δ_{ij}^2 is replaced by $\hat{\delta}_{ij}^2$ in (3.3). The constrained ML estimate $\hat{\boldsymbol{\Phi}}$ is the solution to $S(r)$. Due to the weight matrix $\boldsymbol{W}$, a direct SVD solution does not exist. Instead, $\hat{\boldsymbol{\Phi}}$ can be obtained by criss-cross regression (Gabriel and Zamir, 1979). Write $\boldsymbol{\Phi} = \boldsymbol{A}\boldsymbol{B}'$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are $(I-1) \times r$ and $(J-1) \times r$, respectively. Now (3.4) becomes a standard weighted least squares problem. Given

$\boldsymbol{A}^{(n)}$, $\boldsymbol{B}$ is updated as

$$\boldsymbol{B}^{(n+1)} = \left[(\boldsymbol{I}_{J-1} \otimes \boldsymbol{A}^{(n)})'\boldsymbol{W}^{-1}(\boldsymbol{I}_{J-1} \otimes \boldsymbol{A}^{(n)})\right]^{-1}(\boldsymbol{I}_{J-1} \otimes \boldsymbol{A}^{(n)})'\boldsymbol{W}^{-1}\boldsymbol{K}'\text{Vec}(\bar{\boldsymbol{Y}}) \quad (3.5)$$

In turn, given $\boldsymbol{B}^{(n+1)}$, $\boldsymbol{A}$ is updated as

$$\boldsymbol{A}^{(n+1)} = \left[(\boldsymbol{B}^{(n+1)} \otimes \boldsymbol{I}_{I-1})'\boldsymbol{W}^{-1}(\boldsymbol{B}^{(n+1)} \otimes \boldsymbol{I}_{I-1})\right]^{-1}(\boldsymbol{B}^{(n+1)} \otimes \boldsymbol{I}_{I-1})'\boldsymbol{W}^{-1}\boldsymbol{K}'\text{Vec}(\bar{\boldsymbol{Y}}) \quad (3.6)$$

We alternate (3.5) and (3.6) until convergence of (3.4) is reached, and $\hat{\boldsymbol{\Phi}} = \hat{\boldsymbol{A}}\hat{\boldsymbol{B}}'$. The LRT statistic is $S(0) - S(1)$, where $S(0) = \boldsymbol{K}'\text{Vec}(\bar{\boldsymbol{Y}})\boldsymbol{W}^{-1}\boldsymbol{K}'\text{Vec}(\bar{\boldsymbol{Y}})$. The asymptotic null distribution of this LRT statistic converges in distribution to the maximum root of a p–variate Wishart matrix with df $= \max(I-1, J-1)$ in balanced designs (Boik, 1989). The corresponding 95th and 99th percentiles of this distribution can be found in Hanumara and Thompson Jr (1968). With unbalanced data, under the assumption that $N_{ij} = \sum_j N_{ij} \sum_i N_{ij} / \sum_{ij} N_{ij}$, the null distribution of the LRT is known to be identical to that in balanced designs. Due to correlated nature of the outcome data, these approximations are not directly applicable to our context. However, our numerical work illustrates that using this reference distribution provides a conservative approximation to the test.

3.2.2 Parameter Estimation based on Individual Observations

The cell-means approach provides a quick way of summarizing interaction effects for repeated measures data. In presence of confounders and other covariates, a mixed-effects regression model uses all individual observations and provides a general framework for handling repeated measurements. Let $\boldsymbol{y}_{ijk}$ denote the length-n_{ijk} observation vector for subject (i, j, k),

$$\boldsymbol{y}_{ijk} = \mu_{ij}\mathbf{1}_{n_{ijk}} + \boldsymbol{Z}_{ijk}\boldsymbol{b}_{ijk} + \boldsymbol{e}_{ijk},\ i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, N_{ij}, \quad (3.7)$$

where μ_{ij} is the mean response value for the (i,j)th cell, $\boldsymbol{Z}_{ijk}$ is a $n_{ijk}\times q$ design matrix for the random effects, $\boldsymbol{e}_{ijk} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{ijk})$, not depending on i, j, or k except that its size is $n_{ijk} \times n_{ijk}$, and $\boldsymbol{b}_{ijk}$ is a length-q vector of subject-specific random effects, independent of $\boldsymbol{e}_{ijk}$. The random effects are distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is the $q\times q$ covariance matrix for the random effects. It follows that the variance-covariance matrix for $\boldsymbol{y}_{ijk}$ is $\boldsymbol{V}_{ijk} = \boldsymbol{Z}_{ijk}\boldsymbol{\Psi}\boldsymbol{Z}'_{ijk} + \boldsymbol{\Sigma}_{ijk}$.

Classical Interaction Models

To avoid computationally intensive iterations associated with ML estimation for models (a)–(d), we propose a two-step regression procedure to approximate the interaction parameters. The idea is similar to Milliken and Johnson (1989), who applied a two-step regression procedure in two-way tables to estimate nonlinear interaction effects. Let $\boldsymbol{X}_{ijk}$ be the design matrix with dimension $n_{ijk}\times IJ$ that allows estimation of all plausible effects from the row and column factors. In the first step, we fit a saturated interaction model to the data using a linear mixed-effects model:

$$\boldsymbol{y}_{ijk} = \boldsymbol{X}_{ijk}\boldsymbol{\xi} + \boldsymbol{Z}_{ijk}\boldsymbol{b}_{ijk} + \boldsymbol{e}_{ijk}, \tag{3.8}$$

where $\boldsymbol{\xi} = (\mu, R_1, \ldots, R_{I-1}, C_1, \ldots, C_{J-1}, \gamma_{11}, \ldots, \gamma_{(I-1)(J-1)})'$. The log-likelihood function is

$$\ell(\boldsymbol{\xi}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}) = -\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}}\left[n_{ijk}\log(2\pi)+\log(|\boldsymbol{V}_{ijk}|)+(\boldsymbol{y}_{ijk}-\boldsymbol{X}_{ijk}\boldsymbol{\xi})'\boldsymbol{V}_{ijk}^{-1}(\boldsymbol{y}_{ijk}-\boldsymbol{X}_{ijk}\boldsymbol{\xi})\right]. \tag{3.9}$$

The variance components are estimated by restricted maximum likelihood (REML) (Patterson and Thompson, 1971), and the $I \times J$ fixed effect estimates are

$$\hat{\boldsymbol{\xi}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{N_{ij}}(\boldsymbol{X}'_{ijk}\hat{\boldsymbol{V}}_{ijk}^{-1}\boldsymbol{X}_{ijk})^{-1}\boldsymbol{X}'_{ijk}\hat{\boldsymbol{V}}_{ijk}^{-1}\boldsymbol{y}_{ijk}. \tag{3.10}$$

In the second step, we extract the main effect estimates from $\hat{\boldsymbol{\xi}}$ and compute the residuals

$$\boldsymbol{r}_{ijk} = \boldsymbol{y}_{ijk} - \hat{\mu}\boldsymbol{1}_{n_{ijk}} - \hat{R}_i\boldsymbol{1}_{n_{ijk}} - \hat{C}_j\boldsymbol{1}_{n_{ijk}}. \tag{3.11}$$

Since the interaction term of Tukey's and Mandel's models involves main effects, we perform a second regression (without intercept) where the residuals $\boldsymbol{r}_{ijk}$ are treated as the response variable and the respective specific forms of main effect estimates are treated as the regressors to obtain the corresponding slope estimates. The second-step regression equations for models (a)–(c) are:

$$\boldsymbol{r}_{ijk} = \theta \hat{R}_i \hat{C}_j \mathbf{1}_{n_{ijk}} + \boldsymbol{\epsilon}_{ijk} \tag{3.12}$$

$$\boldsymbol{r}_{ijk} = \lambda_i \hat{C}_j \mathbf{1}_{n_{ijk}} + \boldsymbol{\epsilon}_{ijk} \tag{3.13}$$

$$\boldsymbol{r}_{ijk} = \hat{R}_i \eta_j \mathbf{1}_{n_{ijk}} + \boldsymbol{\epsilon}_{ijk} \tag{3.14}$$

with $\boldsymbol{\epsilon}_{ijk} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\epsilon}(n_{ijk} \times n_{ijk})})$. One can select a covariance structure $\boldsymbol{\Omega}_{\boldsymbol{\epsilon}}$ depending on the criterion of model fitting. Note that parameter constraints in (3.2) are handled in the regressors, and there are $I-1$ and $J-1$ regression equations in (3.13) and (3.14), respectively. For model (d), we first obtain the estimates of λ_i and η_j then compute the second-step residuals using $\{\hat{R}_i, \hat{C}_j, \hat{\lambda}_i, \hat{\eta}_j\}$, and finally estimate θ (see Appendix in Section 3.4 for details).

AMMI1 Model

Given that the interaction structure of model (e) is derived from a SVD of the matrix of residuals after removing additive effects, we propose to perform SVD to the saturated $\hat{\boldsymbol{\Gamma}}$ matrix as obtained from (3.10). The resulting largest singular value of $\hat{\boldsymbol{\Gamma}}$ is an approximation of $\hat{d}_1$. The corresponding left and right singular vectors are approximations of $\hat{\alpha}_i$ and $\hat{\beta}_j$, for $i = 1, \ldots, I, j = 1, \ldots, J$.

Remark: We evaluated the bias and mean squared error (MSE) properties of the two-step regression estimators through simulation. The empirical results indicate that the two-step regression estimators appear to be unbiased, even under misspecified correlation structures (Table 3.4 in Appendix). The estimator of d_1 for AMMI1 models (obtained by SVD of the estimated saturated interaction matrix), however, slightly over-estimates d_1.

3.2.3 Parametric Bootstrap using a LRT Pivot

We construct a LRT statistic based on the non-iterative two-step regression estimates. First, the log-likelihood under the null hypothesis is obtained by fitting an additive mixed model (denoted as $\hat{\ell}_0$), and the log-likelihood under the alternative hypothesis is obtained by the previous two-step regression procedure (denoted as $\hat{\ell}_1$). Specifically for calculating $\hat{\ell}_1$, we extract $\hat{\mu}, \hat{R}_i$, and $\hat{C}_j$ from (3.10) and obtain the interaction effect estimates from the second-step regression of residuals on a pre-specified structure of main effect estimates for models (a)–(d). Subsequently, an approximate LRT pivot is created, $\tilde{\Lambda} = -2(\hat{\ell}_0 - \hat{\ell}_1)$. Because the parameter estimates used in $\tilde{\Lambda}$ are not proper ML estimates, the resulting test statistic does not have a standard asymptotic distribution. We use parametric bootstrap to elicit the null distribution of this LRT-based pivot. Since permuting Y or subjects across the configurations of G and E factors can remove both interaction and main effects, we generate pseudo data $\boldsymbol{y}^*_{ijk}$ under the null hypothesis of no interaction while preserving the main effects using the model: $\boldsymbol{y}^*_{ijk} = \hat{\mu}\mathbf{1}_{n_{ijk}} + \hat{R}_i\mathbf{1}_{n_{ijk}} + \hat{C}_j\mathbf{1}_{n_{ijk}} + \boldsymbol{Z}_{ijk}\boldsymbol{b}^*_{ijk} + \boldsymbol{e}^*_{ijk}$, where $\boldsymbol{b}^*_{ijk} \sim N(0, \hat{\boldsymbol{\Psi}})$, $\boldsymbol{e}^*_{ijk} \sim N(0, \hat{\boldsymbol{\Sigma}})$. $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Sigma}}$ are REML estimates from the saturated mixed-effects model in (3.8). For each simulated null sample, a $\tilde{\Lambda}$ is computed. Repeating the procedure for a large number of times (e.g., we use 1000) provides an approximate distribution of $\tilde{\Lambda}$ under H_0. Finally, an empirical p-value is obtained by calculating the proportion of all $\tilde{\Lambda}$ that exceeds the observed $\tilde{\Lambda}$.

3.2.4 Simulation Settings

We carried out a series of simulation studies to examine the following properties of the proposed tests: [1] type I error for the LRT using cell means (LRT-CM) and for the parametric bootstrap approach with the LRT-based statistic $\tilde{\Lambda}$ (LRT-PB); [2] power comparison of AMMI1 to saturated interaction model if AMMI1 model holds; [3] power comparison of LRT-PB to LRT-CM; and [4] performance of models (a)–(e)

across classical interaction structures. In addition, we compared relative performances of our proposed tests to existing strategies for testing GGI or GEI, including the standard saturated interaction model and the naive cell mean approach of Barhdadi and Dubé (2010) (not accounting for correlated data). Furthermore, we evaluated the performance of each model in detecting GGI with repeated measures on a quantitative trait under 12 epistasis patterns.

Individual-level outcome Y with n_{ijk} repeated measures on subject (i, j, k) were generated for a total of N subjects. The general description of the model is given by

$$\boldsymbol{y}_{ijk} = \mu_{ij}\mathbf{1}_{n_{ijk}} + b_{ijk}\mathbf{1}_{n_{ijk}} + \boldsymbol{e}_{ijk}, \quad k = 1, \ldots, N_{ij}, \tag{3.15}$$

where $\boldsymbol{e}_{ijk} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I})$, $b_{ijk} \sim \mathcal{N}(0, \sigma_b^2)$, and $\{\boldsymbol{e}, \boldsymbol{b}\}$ are mutually independent. Cell means were first generated according to models (a)–(e), and the data vector for each individual with a given mean and a covariance structure was generated from a multivariate normal distribution. The interaction terms in all models were scaled in such a way that they contributed to 15% of the total variation explained by the model, and the remainder was attributed to row and column main effects. While simulating data under an AMMI1 model, we assigned the entire contribution due to interaction effect to the first interaction factor.

To evaluate model performance in terms of detecting common patterns of GGI, data were simulated according to 12 epistasis models (Barhdadi and Dubé, 2010): (1) dominant or dominant; (2) dominant or recessive; (3) modified model; (4) dominant and dominant; (5) recessive or recessive; (6) dominant and recessive; (7) recessive and recessive [(1)–(7) from (Jung et al., 2009)]; (8) checkerboard; (9) diagonal [(8) and (9) from (Culverhouse et al., 2004)]; (10) threshold; (11) additive and additive; and (12) a general model. The additive and additive model and the general model are purely epistatic models, that is, the quantitative trait depends on genotype from two loci in the absence of any marginal effects. Figure 3.1 gives a visual representation of

the true cell means for the 12 epistasis patterns in our simulation.

We considered 3×3 table settings for all simulations to mimic studies of GGI. In addition, we evaluated the power of AMMI1 models in 9×5 table settings as described in [2]. For 3×3 tables, minor allele frequencies for the two loci were set at 0.3 and 0.4, respectively. For 9×5 tables, combinations of the two loci with allele frequencies 0.3 and 0.4 (resulting in nine categories) along with an environmental exposure with five levels (each with probability 0.2) were considered. Hardy-Weinberg equilibrium was assumed to hold for both loci. We set (i) $\sigma^2 = 8, \rho = 0.5$ (or $\sigma_b^2 = \sigma_e^2 = 4$) and (ii) $\sigma^2 = 16, \rho = 0.5$ (or $\sigma_b^2 = \sigma_e^2 = 8$). We also considered $\rho = \{0.2, 0.5, 0.8\}$ for the power evaluation of AMMI1 models. Under each simulation setting, 1000 datasets were generated with 1800 and 3600 subjects for 3×3 and 9×5 tables, respectively. The number of repeated measurements per subject was generated from a multinomial distribution similar to the analysis dataset: $n_{ijk} \in \{2, 3, 4, 5, 6\}$, $\boldsymbol{n} = \{n_{ijk} : 1 \leq k \leq N_{ij}, 1 \leq i \leq I, 1 \leq j \leq J\} \sim \text{mult}(N, \boldsymbol{p})$, $\boldsymbol{p} = (0.15, 0.2, 0.3, 0.2, 0.15)$. This is equivalent to generating outcome data missing completely at random.

3.3 Results

3.3.1 Simulation Findings

Type I Error

We generated data under an additive model, $H_0 : \gamma_{ij} = 0$ (while $R_i, C_j \neq 0$) as well as under a completely null model, $H_0 : \gamma_{ij} = R_i = C_j = 0$ for all i, j. Figure 3.2 shows the percentage of false rejections for the five interaction models from 1000 simulations at 5% significance level. Under the additive model, the type I error rates for all models using LRT-CM and LRT-PB are maintained at the nominal 5%. Under the null model, type I error rates for models (a)–(e) using LRT-PB as well as for model (e) using LRT-CM are still maintained at 5%. LRT-CM for classical models (a)–(d), however, are either too liberal or too conservative (>12% for model (a), >8% for

models (b) and (c), and <3% for model (d)).

Power

The gain in power using an AMMI1 model compared to a saturated interaction model increases as the table dimension increases. In the 3×3 array setting, saturated models (4 df for the interaction effects) appear to have similar power to AMMI1 models (data not shown). In the 9×5 array setting, AMMI1 models using LRT-CM and LRT-PB clearly have greater power than saturated models when the true interaction only has one interaction factor (Figure 3.3). The highest observed gain in power for AMMI1 using LRT-PB compared to the saturated model (32 df for the interaction effects) is 11% under three correlation settings. As ρ increases from 0.2 to 0.8, AMMI1 begins to show power gain across a wider range of d_1.

Figure 3.4 shows the percentage of interactions detected by each test across a set of true simulation models. Overall, the power of LRT-PB is increased by 2–5% compared to LRT-CM. When Tukey's model (a) is the true model, all other models are able to capture some interactions (70–82% when $\sigma_b^2 = \sigma_e^2 = 4$, 38–53% when $\sigma_b^2 = \sigma_e^2 = 8$). Under Mandel's column model (b), Tukey's row-column (d) and AMMI1 (e) are able to detect the interaction (both power >99% when $\sigma_b^2 = \sigma_e^2 = 4$ and >90% when $\sigma_b^2 = \sigma_e^2 = 8$); whereas Tukey's 1-df model (a) and Mandel's row model (c) have very low power (both <50% with LRT-CM and <6% with LRT-PB). Similar properties are observed for simulations under Mandel's row model (c). With Tukey's row-column model (d) being the simulation model, all alternatives, except model (a), are able to detect the interaction with power greater than 60%. When the true model is an AMMI1 model (e), models (a)–(c) have relatively low power to detect interaction (<50% when $\sigma_b^2 = \sigma_e^2 = 4$ and <32% when $\sigma_b^2 = \sigma_e^2 = 8$). Saturated model has lower power than AMMI1 in most cases.

Figure 3.5 shows the percentages of interaction detected by six interaction models using LRT-PB under 12 common epistasis models. Given the robust performance of

model (e), AMMI1 model appears to be a desirable approach for evaluating common epistasis structures, especially when main effects do not exist (e.g., epistasis models (10) and (12)).

We also compared our proposed methods with those in Barhdadi and Dubé (2010) in terms of type I error and power. As expected, the tests in Barhdadi and Dubé (2010) assuming balanced data structure and not accounting for within-subject correlation yield inflated type I error (especially for Tukey's and Mandel's models) and lower power (see Figure 3.8 in the Appendix).

3.3.2 Application to the Normative Aging Study (NAS)

The NAS is a multidisciplinary longitudinal study initiated by the U.S. Veterans Administration in 1963 (Bell et al., 1966). We analyzed 671 participants from a subset of the NAS data who were successfully genotyped for the *HFE* gene and had baseline measurements of tibia bone lead (a measure of cumulative lead exposure) (Zhang et al., 2010). The analysis goal was to investigate effect modification by the different *HFE* alleles on the association between lead exposure and pulse pressure (PP), which is a strong predictor of heart problems for older adults. Since 1991, data had been collected every 3–5 years until 2011 with a median follow-up time of 12 years, including physical examination, blood pressure and laboratory measurements, and questionnaire data. The majority (97%) of the participants were Caucasian. The average age was 66.29 $\pm$ 7.14 (range 48–93) at the time of tibia bone lead measurement. More than 96% of subjects had repeated measurements on blood pressure, and over 65% of them had at least four measurements during the study period contributing to a total of 2914 observations.

Two major mutations in the *HFE* gene (*C282Y* and *H63D* mutations) were considered for analysis following Zhang et al. (2010). Let (*AA, Aa, aa*) and (*BB, Bb, bb*) denote wild type, having one variant allele, and having two variant alleles for *C282Y*

and *H63D*, respectively. As a result of small sample sizes in certain homozygote genotypes (N=5 for *aaBB*, N=17 for *AAbb*) and compound heterozygotes (N=14 for *AaBb* and N=0 for *Aabb, aaBb, aabb*), we were unable to test for interaction between the two loci. Since the research interest was to compare three mutually exclusive groups (wild type, *H63D, C282Y*), 14 subjects with compound heterozygotes (*AaBb*) were excluded from analysis. Consequently, the *HFE* genotypes were classified into three categories for analysis: *AABB, AaBB* or *aaBB*, and *AABb* or *AAbb*. The environmental exposure (cumulative lead) was a continuous variable, but to illustrate the proposed methods, we categorized bone lead levels into three groups (Low: $\leq$15, Medium: (15, 25], and High: >25 $\mu g/g$). Table 3.1 lists the observed cell means of PP and the number of participants for each G $\times$ E configuration.

We applied LRT-CM and LRT-PB to test this GEI effect. According to the Akaike information criterion (AIC) for model fit, we chose a random-intercept mixed-effects model for analysis:

$$\boldsymbol{y}_{ijk} = \mu_{ij}\mathbf{1}_{n_{ijk}} + b_{ijk}\mathbf{1}_{n_{ijk}} + \boldsymbol{e}_{ijk}, \quad \text{with} \quad \mu_{ij} = \mu + R_i + C_j + \gamma_{ij}, \tag{3.16}$$

where $b_{ijk} \sim \mathcal{N}(0, \sigma_b^2)$ is the random-effect coefficient for subject (i, j, k), $\boldsymbol{e}_{ijk} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I})$ is the random error term, and $\{\sigma_b^2, \sigma_e^2\}$ are assumed to be constant across individuals. We also considered the model adjusting for baseline age, time since baseline in years, and squared time.

$$\boldsymbol{y}_{ijk} = \mu_{ij}\mathbf{1}_{n_{ijk}} + \beta_1 \text{Age}_{ijk} + \beta_2 \mathbf{Time}_{ijk} + \beta_3 \mathbf{Time}^2_{ijk} + b_{ijk}\mathbf{1}_{n_{ijk}} + \boldsymbol{e}_{ijk}. \tag{3.17}$$

For LRT-CM adjusting for covariates, cell means were formed by the residuals from a regression of the outcome on covariates other than G and E. This is an ad hoc approach for covariate adjustment since correlations of covariates with G and E are ignored. In general, LRT-PB based on a full regression model with *G, E*, and covariates will yield more power.

Results for GEI

Table 3.2 shows the p-values for testing *HFE* $\times$ Lead Exposure interaction using LRT-CM and LRT-PB and the saturated interaction model. Using LRT-CM without adjusting for any covariate, the interaction was significant in all four classical models ($p < 0.05$), whereas (e) gave a p-value between 0.05 and 0.10. After adjusting for the covariates, model (e) detected the interaction using LRT-CM ($p < 0.01$). Regardless of covariate adjustment, the interaction was significant for models (a)–(e) using LRT-PB ($p < 0.01$), and also for the saturated interaction model ($p < 0.02$). P-values for the GEI effect decreased further for all tests with adjustment for baseline age, time since baseline, and squared time. Given the significant GEI on all models, this interaction may be real and not model dependent.

According to the SVD of $\hat{\mathbf{\Gamma}}_{G\times E}$ under a saturated interaction model with covariate adjustment, the first and second characteristic roots of $\hat{\mathbf{\Gamma}}_{G\times E}$ were $\hat{d}_1 = 5.65$ and $\hat{d}_2 = 1.24$, respectively (Table 3.5). The first interaction factor contributed to over 80% of the total contribution to the interaction term. The association between PP and bone lead levels was strongest among *H63D* variant ($AABb$ or $AAbb$) carriers, compared to *C282Y* variant ($AaBB$ or $aaBB$) carriers and wild-type ($AABB$) participants. Based on the saturated interaction model estimates, the estimated difference in mean PP for *H63D* variant carriers with High versus Low lead levels was 9.6 mmHg [95% confidence interval (CI), 0.43–14.83 mmHg]. The same estimated differences were 3.52 mmHg [95% CI, -3.39–10.43 mmHg] and -0.33 mmHg [95% CI, -2.95–2.29 mmHg] for *C282* variant carriers and wild-type participants, respectively.

Subject-specific and Age-specific Contributions to GEI

Using the estimates of singular vectors ($\hat{\alpha}_{im}, \hat{\beta}_{jm}$), we investigated subject-specific and age-specific contributions to GEI in the first and the second interaction factors ($m = 1, 2$) via the sum of squared deviations (Mukherjee et al., 2012). Briefly, the variation due to subject (i, j, k) can be calculated by $\hat{d}_{ijkm} = \hat{\alpha}_{im}\hat{\beta}_{jm}\hat{r}_{ijk.}$, where $\hat{r}_{ijk.}$

is the mean of n_{ijk} subject-level residuals from (3.11) after removing main effects. The variation in the contribution of subject (i, j, k) is then $(\hat{d}_{ijkm} - \hat{d}_{.m})^2$, where $\hat{d}_{.m} = \sum_i \sum_j \sum_k \hat{d}_{ijkm}/N$. For the age-specific contribution, we constructed eight three-year age intervals. The first and last intervals contained observations from those who were younger than 65 and who were 83 or older, respectively. The cell means of PP and numbers of participants for genotypes and lead levels based on different age categories are presented in Figure 3.9 in Appendix. The contribution due to the t-th age interval is calculated as $\hat{d}_{tm} = \sum_i \sum_j \hat{\alpha}_{im}\hat{\beta}_{jm}\hat{r}_{tij.}$, where $\hat{r}_{tij.}$ is the average of residuals (3.11) in the t-th age interval among individuals in the (i, j)th cell, $t = 1, \ldots, 8$. The variation in the contribution of the t-th interval is $(\hat{d}_{tm} - \hat{d}_{.m})^2$, where $\hat{d}_{.m} = \sum_{t=1}^{8} \hat{d}_{tm}/8$.

Figure 3.6 displays subject-specific contributions from the 671 individuals (left panel) and contributions of eight age intervals to the first interaction factor of GEI (right panel). The plot indicates that the modifying effect of the *HFE* gene on the effect of cumulative lead exposure on PP spiked around age 75. This was due to the fact that the mean difference in PP between the Low and the High bone lead groups became largest in that age interval with *H63D* ($AABb$ or $AAbb$), whereas the difference in PP among those with wild-type of *HFE* ($AABB$) was the smallest. A stratified analysis by baseline age also indicates time-dependent evidence of interaction effects (see Appendix). Figure 3.10 shows patterns corresponding to the second interaction factor with substantially less subject-specific and age-specific variability for the GEI. These graphical diagnostics can provide important insight into longitudinal features of the interaction factors.

3.4 Discussion

We have proposed new likelihood ratio tests for GGI and GEI effects in longitudinal cohort studies using a sparse representation of interaction structure via Tukey's

and Mandel's models as well as AMMI1 models. AMMI1 appears to be a robust and flexible model in detecting interaction effects across a spectrum of interaction structures. Moreover, it is relatively powerful in detecting certain epistasis structures with no appreciable main effects but potential interaction. In contrast, Tukey's and Mandel's models fail to detect interactions if the interaction structure is misspecified.

Both of our approaches require prior assumptions of the mean structure under the null hypothesis of no interaction and an underlying correlation structure for within-subject measurements. When either part of the model is misspecified, the power and the false rejection rate might be affected. Although this is a generic limitation of parametric modeling, we performed additional simulations to evaluate the influence of misspecification of covariance structure on the proposed tests. We generated data under several common correlation structures (e.g., compound symmetry, autoregressive, unstructured) and analyzed interactions assuming a different correlation structure. The results show that under a misspecification of covariance structure, type I error rates are maintained for LRT-PB but can be slightly inflated for LRT-CM.

In our simulation studies, we did not see a vast difference in the power between LRT-CM and LRT-PB. The correlation across repeated measurements in the LRT-CM approach is accounted for by the weight matrix $\boldsymbol{W}$. Therefore, the test is not based on naive subject-level averages as in Mukherjee et al. (2012). The main advantage of LRT-PB is the flexible regression structure that allows all readily available mixed model estimation tools to be used.

We have focused on developing valid tests for the five interaction models, yet there are some limitations. First, a caveat of the ML estimation for classical interaction models (a)–(d) based on cell means is that when the underlying main effects are relatively small, the estimation for interaction parameters would become numerically unstable. Depending on initial values, the final converged estimates might be local ML estimates instead of global ML estimates. Second, SVD of the estimated saturated

interaction matrix yields approximate estimates rather than proper ML estimates for AMMI model parameters. Our simulation results indicate that this estimator leads to slight over-estimation of d1. Nevertheless, LRT-PB for AMMI1 still maintains the nominal type I error rate and in general possesses greater power than a saturated interaction model. In addition, how to connect the parameters of the AMMI model to directly interpretable quantities for biological interactions are not clear. At this stage, AMMI model remains a screening strategy for testing non-additivity. Third, covariate adjustment was not considered in our simulation studies. In practice, one can incorporate time effect and other (time-varying) covariates with LRT-CM and LRT-PB, as we did in our data example. Lastly, to our knowledge, no replication study has examined the interaction effect between the *HFE* gene and cumulative lead exposure on pulse pressure. We randomly split the data in half and analyzed the two halves for GEI as an assessment of internal consistency, and the results were consistent with our findings. As discussed in Zhang et al. (2010), the conclusion from the NAS data analysis may not be generalizable to other populations given that the study population was exclusively white men. Besides, unmeasured confounding factors and interactions with other genetic polymorphisms or environmental factors were not considered.

The proposed analysis strategies are useful for detecting GGI and GEI effects in longitudinal data. A full likelihood-based approach using a general nonlinear mixed-effects model set-up would be more appealing if the appropriate test statistics and their closed form null distributions can be obtained. Further work is required to investigate specialized nonlinear optimization algorithms in the ML framework to replace the two-step estimation and to construct a valid and more efficient test. It is also important to develop a formal test for individual- and time-specific contributions to interactions, which will ultimately lead to better understanding of GGI and GEI.

Table 3.1: Cell means corresponding to pulse pressure and number of participants (in parentheses) for each configuration of the *HFE* genotypes and bone lead levels in the Normative Aging Study

	Tibia lead levels (μg/g)		
HFE gene	Low: ≤ 15	Medium: > 15 and ≤ 25	High: >25
Wild-type *(AABB)*	52.94 (161)	56.16 (149)	56.61 (131)
C282Y (AaBB or *aaBB)*	51.89 (23)	56.65 (39)	59.10 (23)
H63D (AABb or *AAbb)*	52.58 (54)	57.72 (53)	64.49 (38)

Table 3.2: P-values for testing GEI between *HFE* genotypes and tibia lead levels in the Normative Aging Study using the proposed likelihood ratio test with cell means (LRT-CM) and the parametric bootstrap (LRT-PB) approach (1000 replicates simulated under the null hypothesis)

Model	Hypothesis	LRT-CM[a]	LRT-CM[b]	LRT-PB[a]	LRT-PB[b]
Model (a)	$H_0 : \theta = 0$	0.008	0.002	0.002	0.003
Model (b)	$H_0 : \lambda_i = 0$ (Lead)	0.029	0.007	0.009	0.008
Model (c)	$H_0 : \eta_j = 0$ (*HFE*)	0.015	0.002	0.002	0.001
Model (d)	$H_0 : \theta = \lambda_i = \eta_j = 0$	0.035	0.005	0.007	0.002
Model (e)	$H_0 : d_1 = 0$	<0.10	<0.01	0.009	0.002
Saturated	*HFE* $\times$ Lead			0.015	0.006

[a] No covariate adjustment.

[b] Adjusting for baseline age, time since baseline, and squared time. For LRT-CM, residuals from a regression of pulse pressure on all other covariates except lead levels and genotype were used to form the cell means.

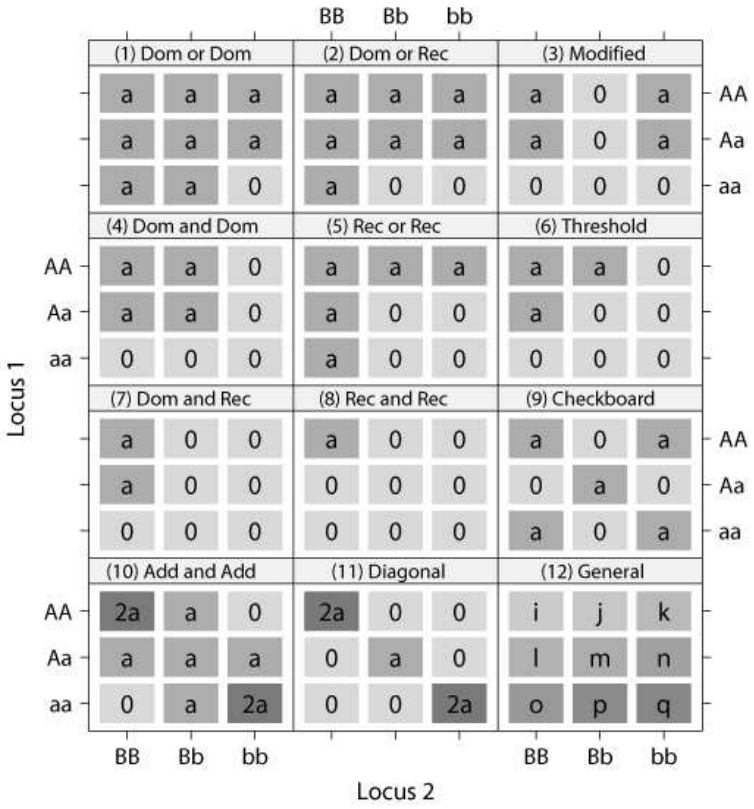


Figure 3.1: Cell means ($a = 0.5$) for 12 common epistasis models

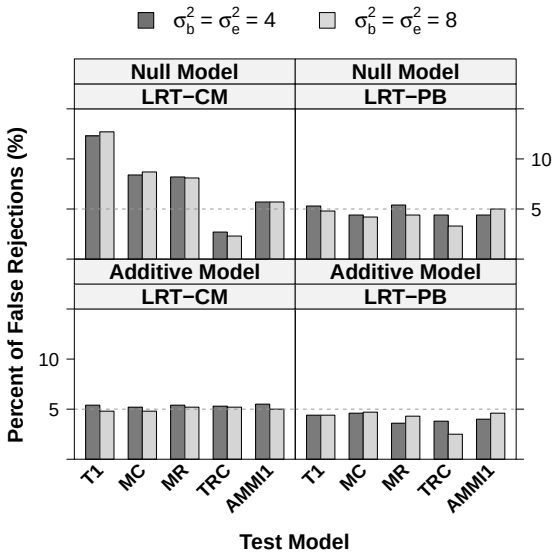


Figure 3.2: Type I error for the five interaction tests in a 3×3 array setting using the likelihood ratio test with the cell-mean approach (LRT-CM) and the parametric bootstrap test (LRT-PB). 1000 simulation datasets are generated under an additive model (only main effects) and under a completely null model (no main or interaction effects). T1 = Tukey's one degree-of-freedom non-additivity test (a), MC = Mandel's column model (b), MR = Mandel's row model (c), TRC = Tukey's row-column model (d), AMMI1 = model (e).

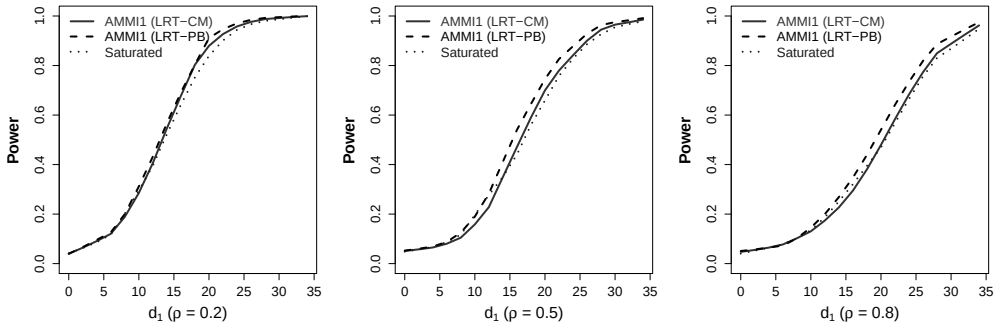
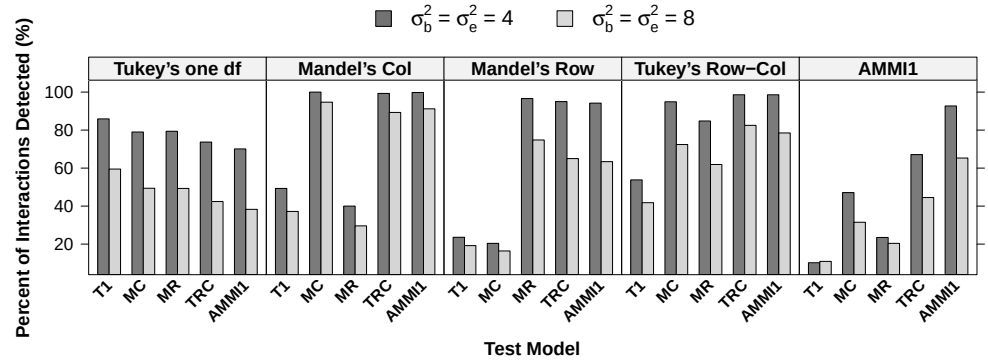
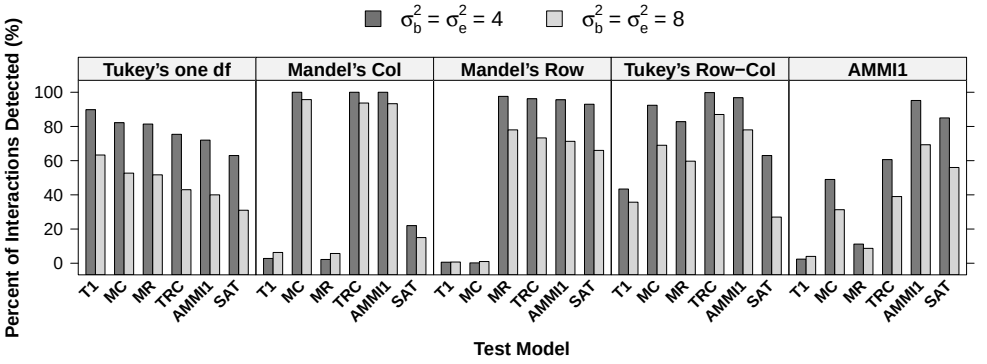


Figure 3.3: Empirical power (or true positive rate) of AMMI1 model (at α=0.05) using the likelihood ratio test with the cell-mean approach (LRT-CM) and the parametric bootstrap test (LRT-PB), and a saturated interaction model in a 9×5 array setting with $\sigma^2 = 8$ and $\rho = 0.2, 0.5$, and 0.8.



(a)



(b)

Figure 3.4: Percentage of interactions detected by different interaction models in the simulation settings corresponding to a 3×3 array. Results are based on (a) the likelihood ratio test with the cell-mean approach (LRT-CM) and (b) the parametric bootstrap test (LRT-PB) with test results of using a saturated model for interaction as a comparison. The top label within each box represents the true simulation model. The horizontal-axis labels indicate the models used for testing interaction. T1 = Tukey's one degree-of-freedom non-additivity test (a), MC = Mandel's column model (b), MR = Mandel's row model (c), TRC = Tukey's row-column model (d), AMMI1 = model (e), SAT = saturated interaction model.

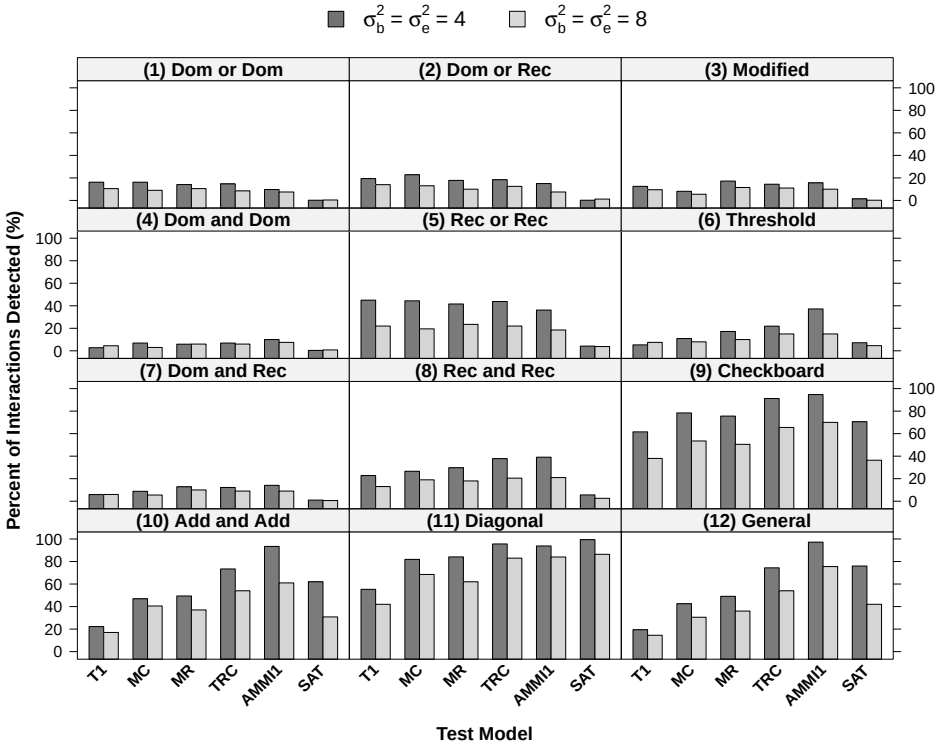


Figure 3.5: Percentage of interactions detected (or null hypotheses of no interaction rejected) by each of the interaction models using parametric bootstrap test (LRT-PB) and a saturated model for interaction under 12 common epistasis models. T1 = Tukey's one degree-of-freedom non-additivity test (a), MC = Mandel's column model (b), MR = Mandel's row model (c), TRC = Tukey's row-column model (d), AMMI1 = model (e), SAT = saturated interaction model.

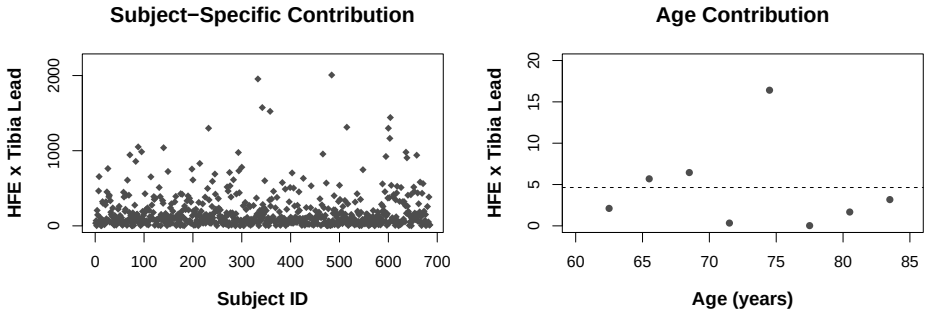


Figure 3.6: Subject-specific contributions (left) and age-specific contributions (right) to the *first* interaction factor in the *HFE* × Lead interaction based on the Normative Aging Study data.

3.5 Appendix

3.5.1 Sensitivity of Using the Empirical Variance Estimate for LRT-CM

We studied the sensitivity of using the empirical variance estimates by comparing the power and type I error under true and misspecified correlation structures across a variety of commonly used correlation structures (e.g., compound symmetric, autoregressive, unstructured). Table 3.3 shows the simulation results of the LRT-CM method with and without a misspecified correlation structure under compound symmetric (CS) and autoregressive-1 (AR-1) correlation structures.

Overall, the power of the tests decreases when the assumed correlation structure is more complicated (i.e., more parameters need to be estimated) than the true underlying correlation structure. In our simulation setting, the power of LRT-CM decreases by 16% (at most) when assuming an AR-1 correlation while the true correlation structure is a CS. In contrast, the power of the LRT-CM is not affected as much (less than 6%) when the assumed correlation structure requires fewer parameters to be estimated than the true underlying correlation structure. Concerning the type I error under misspecified correlation structure, we again investigated the type I error under two null hypotheses: (1) no interaction with the presence of main effects and (2) no interaction without the presence of main effects. Under the null hypothesis of (2), the estimates can become quite unstable for models (a)–(d). The type I error can be inflated or deflated under misspecified correlation structure but always remain less than 10% in our simulations.

3.5.2 Estimation for Tukey's Row-Column Model in Two-Step Regression

We can express the interaction term $\theta R_i C_j + \lambda_i C_j + R_i \eta_j$ in the model as $\gamma_{ij} = (\theta R_i + \lambda_i) C_j + R_i \eta_j$ or $(\theta C_j + \eta_j) R_i + \lambda_i C_j$. If we regress the residuals $\boldsymbol{r}_{ijk}$ after removing the additive main effects (from a saturated model fit) on $\hat{C}_j$ and $\hat{R}_i$ (again

without intercept) separately:

$$\boldsymbol{r}_{ijk} = u_i \hat{C}_j \mathbf{1}_{n_{ijk}} + \boldsymbol{\epsilon}_{ijk}, \tag{3.18}$$

$$\boldsymbol{r}_{ijk} = v_j \hat{R}_i \mathbf{1}_{n_{ijk}} + \boldsymbol{\epsilon}_{ijk}, \tag{3.19}$$

we have $\hat{u}_i = \tilde{\theta}\hat{R}_i + \tilde{\lambda}_i$, and $\hat{v}_j = \tilde{\theta}\hat{C}_j + \tilde{\eta}_j$. Model (d) has a total of $I + J + 1$ interaction parameters. Together with four sum-to-zero identifiability constraints, $I+J-3$ parameters (i.e., $\lambda_1 \ldots, \lambda_{I-2}, \eta_1, \ldots, \eta_{J-2}$) are left to be estimated. By (3.18) and (3.19), we have $(I-1)+(J-1)$ equations, which are sufficient for estimating the $I + J - 3$ parameters. After obtaining $\hat{u}_i$ and $\hat{v}_j$ from (3.18) and (3.19), each $\hat{\lambda}_i$ and $\hat{\eta}_j$ can be calculated using the constraints. Finally, we estimate θ by regressing the residuals from the second step, $\boldsymbol{s}_{ijk} = \boldsymbol{r}_{ijk} - \hat{R}_i\hat{\eta}_j\mathbf{1}_{n_{ijk}} - \hat{\lambda}_i\hat{C}_j\mathbf{1}_{n_{ijk}}$, on $\hat{R}_i\hat{C}_j$,

$$\boldsymbol{s}_{ijk} = \theta \hat{R}_i \hat{C}_j \mathbf{1}_{n_{ijk}} + \boldsymbol{\varepsilon}_{ijk},$$

where $\boldsymbol{\varepsilon}_{ijk} \sim N(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\varepsilon}})$. Again, $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon}}$ can be a user-defined covariance structure based on model fitting criterion.

3.5.3 Comparison with Other Exisiting GGI/GEI Methods

The existing GGI or GEI methods for handling (longitudinal) continuous traits are very limited. Barhdadi and Dubé (2010) have applied Tukey's and Mandel's models as well as AMMI models to testing GGI effects on quantitative traits for unbalanced data. They reduced data to cell means and applied F tests that assume equal variance of all cell means as described in the original papers of Tukey (1949) and Mandel (1961). The likelihood ratio test proposed by Johnson and Graybill (1972a) was used for GGI tests with AMMI models, which is also based on single observation per cell. Despite these complex classical models, a saturated model for interaction is commonly used for testing GGI and GEI in practice for its computational simplicity and flexibility.

We generated interaction data in the same simulation setting as described in the main text (unbalanced correlated data in 3×3 table settings) and applied the GGI tests summarized in Barhdadi and Dubé (2010) for Tukey's, Mandel's, and AMMI models (any within-subject correlation is ignored). Figure 3.8 shows type I error (left panel) and power (right panel) for each of the five multiplicative models using tests in Barhdadi and Dubé and our proposed tests (LRT-CM and LRT-PB) under the same simulation settings as described in the section of Simulation Settings in the main text. As expected, the tests in Barhdadi and Dubé (2010) assuming balanced data structure and not accounting for within-subject correlations yield inflated type I error (especially for Tukey's and Mandel's models) and low power, compared to our proposed methods. For example, when the simulation model is AMMI1 with $\sigma_b^2 = \sigma_e^2 = 8$, AMMI1 has 65% and 69% power for detecting interactions using our proposed LRT-CM and LRT-PB, respectively; whereas AMMI1 using the test by Barhdadi and Dubé only has 8% power (far right column).

3.5.4 Stratified Analysis of GEI in the NAS Data

To further investigate the potential three-way interaction (age contributions to *HFE* $\times$ Lead interaction), we performed stratified analysis for by baseline age: one for those who started the study at age < 66 years old (N=316) and the other one with those who started at age ≥ 66 years old (N=355). We then analyzed the two subsets separately. The p-values for GEI using models (a)–(e) are shown in Table 3.6. The results indicate that *HFE* $\times$ Lead interaction was found for the older group of participants but not for the younger group. The stratified analysis results may indicate some evidence of three-way (age-dependent) interaction.

Table 3.3: Power and type I error of the LRT-CM method with and without a misspecified correlation structure. Two covariance structures were compound symmetric and autoregressive-1 correlation ($\sigma^2 = 16, \rho = 0.5$).

	Corr. Structure		Model				
	True	Assumed	(a) T1	(b) MC	(c) MR	(d) TRC	(e) AMMI1
Power	CS	CS	0.592	0.947	0.747	0.825	0.651
	CS	AR-1	0.528	0.924	0.660	0.750	0.544
	AR-1	AR-1	0.693	0.973	0.858	0.921	0.779
	AR-1	CS	0.722	0.977	0.888	0.940	0.830
Type I Error (Additive)	CS	CS	0.054	0.052	0.054	0.053	0.055
	CS	AR-1	0.044	0.034	0.035	0.024	0.027
	AR-1	AR-1	0.055	0.057	0.053	0.058	0.056
	AR-1	CS	0.074	0.079	0.082	0.087	0.088
Type I Error (Null)	CS	CS	0.127	0.087	0.081	0.023	0.057
	CS	AR-1	0.091	0.057	0.050	0.012	0.027
	AR-1	AR-1	0.125	0.083	0.080	0.023	0.056
	AR-1	CS	0.147	0.116	0.098	0.037	0.088

CS = compound symmetric; AR-1 = autoregressive-1

Table 3.4: Percent bias and mean squared error (MSE) corresponding to the interaction parameter estimates from Tukey's 1-df model (θ) and AMMI1 model (d_1) using a two-step regression procedure under compound symmetric and autoregressive-1 correlation structures (both with $\rho = 0.5$)

			Assumed Correlation Structure for Analysis				
Parameter	True	σ^2	CS	AR-1	ARH	UN	IND
Percent Bias (%)							
		1	-0.2	-0.2	-0.2	-0.2	-0.2
θ	CS	4	0.2	0.3	0.3	0.2	0.3
		8	0.8	0.8	0.8	0.8	0.7
		1	0.1	0.1	0.1	0.1	0.1
	AR-1	4	0.7	0.7	0.7	0.7	0.7
		8	-0.4	-0.4	-0.4	-0.4	-0.4
		1	1.2	1.2	1.2	1.2	1.2
d_1	CS	4	3.0	3.0	3.0	3.0	3.1
		8	6.8	6.8	6.9	6.8	7.0
		1	1.0	1.0	1.0	1.0	1.0
	AR-1	4	1.9	1.9	1.9	1.9	1.8
		8	3.6	3.3	3.3	3.3	3.7
MSE							
		1	0.090	0.090	0.090	0.090	0.090
θ	CS	4	0.182	0.185	0.185	0.183	0.185
		8	0.263	0.268	0.268	0.263	0.270
		1	0.078	0.077	0.077	0.077	0.078
	AR-1	4	0.163	0.162	0.162	0.162	0.163
		8	0.239	0.239	0.239	0.239	0.239
	CS	1	0.101	0.102	0.103	0.101	0.104
d_1		4	0.199	0.204	0.204	0.200	0.206
		8	0.278	0.282	0.282	0.278	0.284
	AR-1	1	0.096	0.093	0.093	0.093	0.095
		4	0.178	0.175	0.176	0.176	0.179
		8	0.252	0.250	0.250	0.251	0.253

CS = compound symmetric; AR-1 = autoregressive-1; ARH = autoregressive heterogeneous; UN = unstructured; IND = independence
True $\theta = d_1 = 1$

Table 3.5: Estimated interaction matrices from fitting a saturated model (adjusted for baseline age, time, and squared time) and the corresponding singular value decompositions: $\hat{\mathbf{\Gamma}}_{G\times E}$ for gene-environment ($HFE \times$ Lead) interaction analysis based on the Normative Aging Study data

$\hat{\mathbf{\Gamma}}_{G\times E}$			$\hat{\boldsymbol{A}}_{\text{HFE}}$		$\hat{\boldsymbol{D}}$		$\hat{\boldsymbol{B}}'_{\text{Lead}}$		
1.92	0.77	-2.68	-0.59	0.57	5.65	0	-0.43	-0.38	0.82
-0.19	1.14	-0.94	-0.20	-0.79	0	1.24	0.69	-0.72	0.03
-1.73	-1.90	3.63	0.79	0.22					

Table 3.6: P-values corresponding to different tests for GEI between *HFE* genotypes and tibia lead levels in the Normative Aging Study stratified by baseline age at the time of recruitment are reported. LRT-CM and LRT-PB stand for the two likelihood ratio tests based on cell means and al mixed-effects regression model, respectively. The model adjusts for baseline age (years), time since baseline, and squared time. For LRT-CM, the residuals from the adjusted model were used to form cell means corresponding to G×E cross-tables.

		Baseline Age <66		Baseline Age ≥ 66	
Model	Hypothesis	LRT-CM	LRT-PB	LRT-CM	LRT-PB
Model (a)	$H_0 : \theta = 0$	0.054	0.208	0.001	0.001
Model (b)	$H_0 : \lambda_i = 0$ (Lead)	0.143	0.080	0.003	0.001
Model (c)	$H_0 : \eta_j = 0$ (*HFE*)	0.133	0.142	0.001	0.002
Model (d)	$H_0 : \theta = \lambda_i = \eta_j = 0$	0.234	0.184	0.002	$<$.0001
Model (e)	$H_0 : d_1 = 0$	$<$0.10	0.250	$<$0.005	0.001
Saturated	$HFE \times$ Lead	NA	0.284	NA	0.002

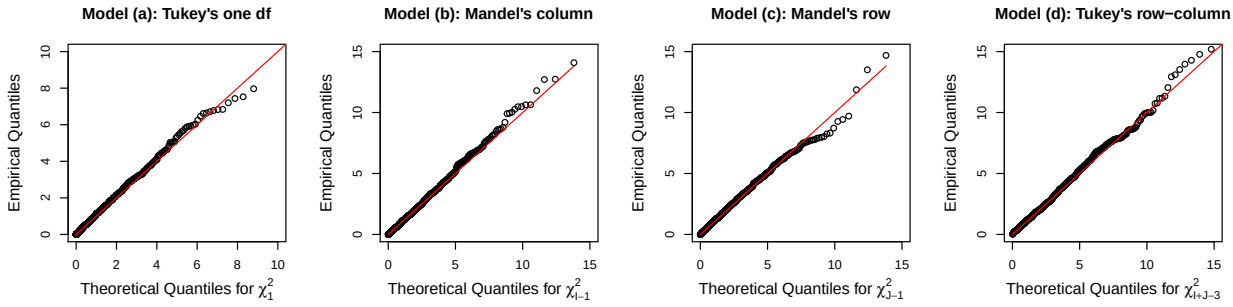
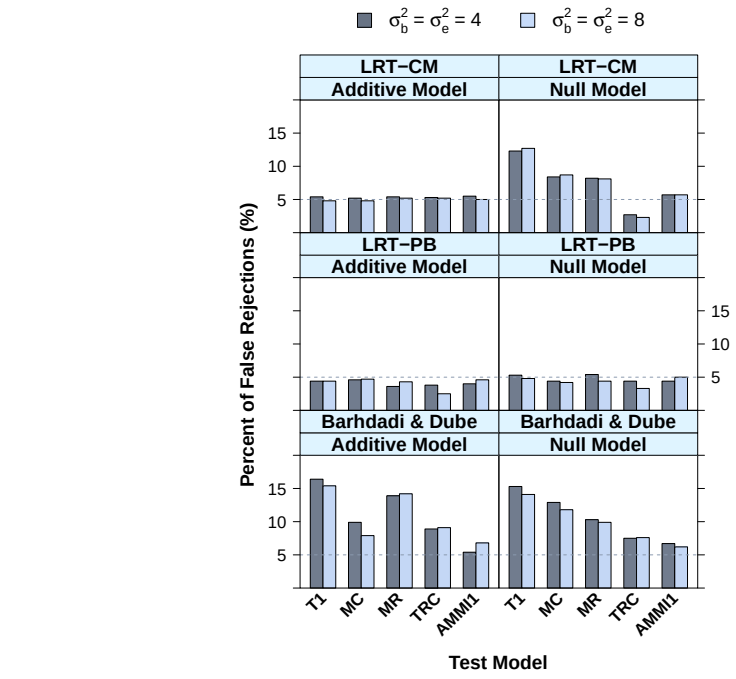
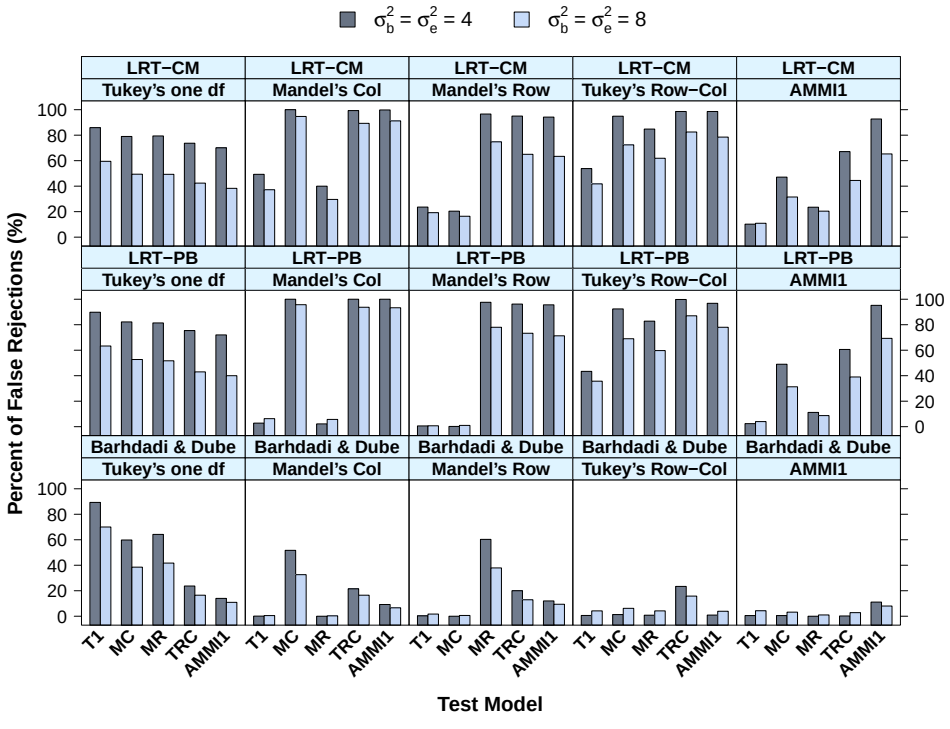


Figure 3.7: Comparison of empirical quantiles of the likelihood ratio test (LRT) statistics to the corresponding theoretical quantiles of chi-squares under the null hypothesis based on $I \times J$ cell means. The LRT statistic follows a chi-square distribution with df $= 1, I-1, J-1$, and $I+J-3$ for models (a), (b), (c), and (d), respectively ($I = J = 3$).



(a)



(b)

Figure 3.8: (a) Type I error and (b) percentage of interactions detected by each of the five multiplicative models using tests in Barhdadi and Dubé (2010) and the proposed methods in the same simulation settings as described in the section of Simulation Settings. The top label within each box represents the true simulation model. The horizontal-axis labels indicate the tests carried out.

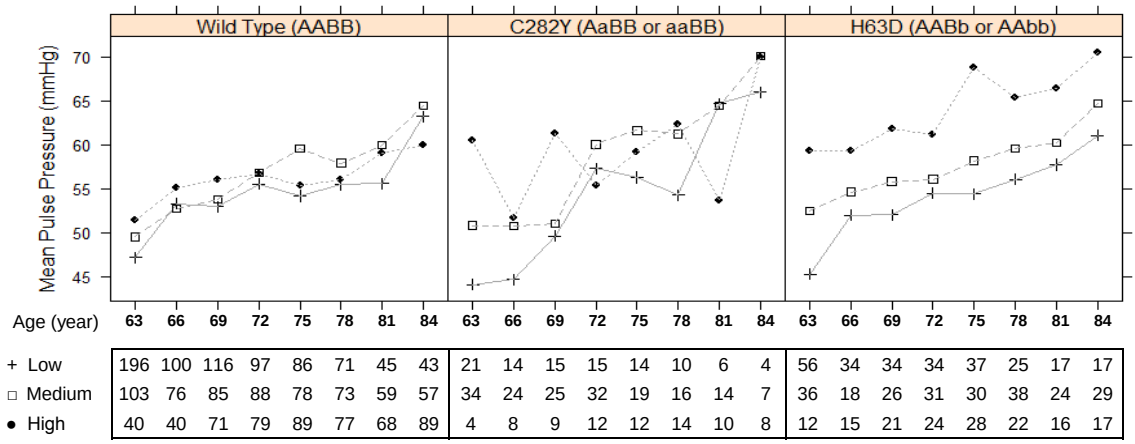


Figure 3.9: Cell means of pulse pressure and numbers of observations (shown in table below the graph) for three genotypes of the *HFE* gene and lead exposure levels (Low, Medium, High) across eight age intervals in the Normative Aging Study

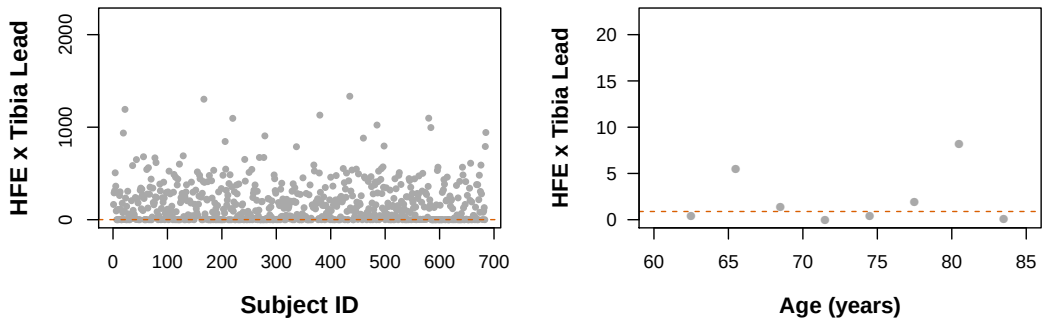


Figure 3.10: Subject-specific contributions and age-specific contributions to the *second* interaction factor in the *HFE* × Lead interaction based on the Normative Aging Study data

CHAPTER IV

Testing Departure from Additivity in Tukey's Model using Shrinkage: Application to a Longitudinal Setting

4.1 Introduction

The presence of gene-environment interactions (GEI) implies that the effect of an environmental exposure (E) is enhanced or reduced for a sub-group with a certain genotype or vice versa. Investigation of GEI is essential to better understand the etiology and development of common, complex diseases. Many longitudinal environmental epidemiology studies have been collecting genetic data with the goal of identifying GEI. In these cohort studies, GEI is often investigated by focusing on an established association between an exposure biomarker (e.g., lead levels in blood or bone) and a quantitative trait (e.g., pulse pressure), and how this association is modified by a selected set of genetic markers. The set of genes (candidate genes) to be studied is often determined by the metabolic pathway related to the exposure instead of an agnostic search across the genome.

While there has been extensive literature on GEI regarding ways to enhance the efficiency of interaction test in case-control studies (Kraft et al., 2007; Mukherjee and Chatterjee, 2008; Mukherjee et al., 2012), statistical methods for GEI in longitudinal settings remain limited. Methods to study disease-gene association in longitudinal settings, however, have started to receive attention. For instance, Wang et al. (2012) proposed to estimate and test for time-varying genetic effects using semiparametric

models with penalized splines. Fan et al. (2012) also used penalized spline models to estimate the mean function and genetic regression coefficients with extensions to linkage disequilibrium (LD) mapping. Nevertheless, limited number of studies have focused on testing of gene-gene interactions (GGI) or GEI for complex traits in longitudinal settings. The multivariate adaptive splines presented by Zhang (1997, 2004) have been applied to analyze GEI in longitudinal cohort studies (e.g., Zhu et al., 2009). Xu (2007) developed an empirical Bayes method to estimate GGI effects under the mixed model framework and compared it with several variable selection procedures. Malzahn et al. (2010) developed a nonparametric test for investigation of GGI in repeated measures data using a rank procedure. Mukherjee et al. (2012) proposed to explore the GEI structure with various parsimonious classical ANOVA models for non-additivity by taking the average of repeated measurements and forming cell means of a two-way GEI table. Along the same lines, Ko et al. (2013) extended the classical ANOVA models under a mixed model framework and developed a resampling-based test for GEI that accounts for correlation within repeated measures.

Typically, an interaction model including cross-product terms of gene and environment under the mixed model framework is used for testing GGI and GEI in longitudinal studies (Moreno-Macias et al., 2010). In considering the estimation of GEI for longitudinal data where both the genetic factor (G) and E are categorical variables, this conventional modeling approach involves distinct parameter estimation for each configuration of GEI (i.e., a saturated interaction form) with sum-to-zero type constraints to ensure identifiability. Estimation bias is minimized since the model does not impose any structural assumptions on the interaction term. However, the number of parameters and hence the corresponding degrees of freedom (df) for the interaction test can become substantially large as the number of categories of G and/or E increases. In addition, under a saturated interaction model only observations in a cell can contribute to the parameter estimation for that cell. This may result in

reduced efficiency and loss of power for detecting interactions because of small cell sample size in human studies involving a gene with a modest minor allele frequency.

Tukey's one df model for non-additivity (Tukey, 1949), originally proposed for data with no replication per cell, has been applied to the modeling of GGI in cohort studies (Maity et al., 2009). The interaction term in Tukey's model is treated as a scaled product of main effects, implying that the existence of interaction is conditional on the presence of main effects. When a GEI study is based on a two-stage strategy, namely, the candidate genes are selected based on marginal genetic associations (Kooperberg and LeBlanc, 2008; Murcray et al., 2009), it may be reasonable to adopt Tukey's interaction form for GEI. Chatterjee et al. (2006) proposed that Tukey's model is also consistent with the notion that individual markers within a gene are associated with disease through a common biological mechanism. However, when candidate genes are chosen in relation to an exposure pathway, genes may not necessarily have main effects. Also, when the assumption of Tukey's interaction structure is violated (e.g., absence of genetic main effects), the estimate for the interaction effect using Tukey's model will be biased and the corresponding one-df test can result in extremely low power (Mukherjee et al., 2012; Barhdadi and Dubé, 2010).

When searching for GEI across multiple genetic markers, it is possible that GEIs exhibit distinct interaction patterns, departing from Tukey's model. Conducting multiple tests under a fixed interaction structure (e.g., Tukey) may not capture interactions of alternative forms. At the same time, it would be advantageous to leverage the power of Tukey's test if it is indeed a plausible model. Given as such, we propose to model GEI using a shrinkage estimator that combines estimates from Tukey's model and from the saturated interaction model. An adaptive framework is utilized similar to Mukherjee and Chatterjee (2008). This estimator will shrink the maximum likelihood estimates (MLEs) under a flexible interaction structure toward Tukey's model estimates. The amount of shrinkage is data adaptive, so that in large samples, such

estimator is unbiased even if Tukey's assumption is violated. More importantly, when compared to a saturated model, the shrinkage estimator has reduced mean squared error (MSE) for small samples (Chen et al., 2009). Although Tukey's model has been used to model GEI or GGI under a generalized linear model setting (Maity et al., 2009; Chatterjee et al., 2006; Barhdadi and Dubé, 2010), no prior work has been carried out to data-adaptively combine Tukey's model and saturated interaction model to take advantage of both models for testing GEI. Thus, the shrinkage approach is not only novel for longitudinal data but also a new approach for cross-sectional data.

In Section 4.2, we introduce notations for GEI models using a mixed-effects model framework. The parameter estimation for Tukey's model with repeated measures data is described in Section 4.3. In Section 4.4, we propose a shrinkage estimator and derive its approximate variance estimate. In Section 4.5, we summarize the test for interaction corresponding to each method. In Section 4.6, we evaluate the performance of our proposed methods via simulation studies. In particular, we compare the average performance by generating GEIs with different interaction structures to mimic a hypothetical GEI search study involving multiple genetic markers. In Section 4.7, we apply the proposed methods to search GEI between 105 single-nucleotide polymorphisms (SNPs) within 22 genes in the iron metabolism pathway and cumulative lead exposure on pulse pressure using the Normative Aging Study (NAS) data. We also test GEI between 27 SNPs and energy intake and intentional exercise on body mass index (BMI) using data from the Multi-Ethnic Study of Atherosclerosis (MESA). These 27 SNPs have been shown to be significantly associated with BMI in previous genome wide association studies (GWAS). In NAS, genes are chosen in relation to the exposure pathway. In MESA, the question is whether the loci identified by GWAS (with marginal effects) modify the effect of certain exposures. Another distinction between the two data examples is that one of the exposure variables considered in MESA, intentional exercise, is a time-varying variable, while the other two, energy

intake in MESA and cumulative lead exposure in NAS, are time-invariant (i.e., both are baseline measurements).

4.2 Model

Let y_{kt} be the value of the t-th repeated measure on a phenotypic response Y corresponding to the k-th individual ($t = 1, \ldots, n_k$, $k = 1, \ldots, N$). Define a mixed-effects model for the $n_k \times 1$ response vector $\boldsymbol{y}_k = (y_{k1}, y_{k2}, \ldots, y_{kn_k})^\top$ such that it is related to an $n_k \times \nu$ matrix of explanatory variables $\boldsymbol{X}_k = (\boldsymbol{x}_{k1}, \boldsymbol{x}_{k2}, \ldots, \boldsymbol{x}_{kn_k})^\top$, with each $\boldsymbol{x}_{kt}$ a $\nu \times 1$ vector associated with y_{kt}, through some nonlinear function f. Namely,

$$\boldsymbol{y}_k = \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k) + \boldsymbol{Z}_k \boldsymbol{b}_k + \boldsymbol{e}_k, \tag{4.1}$$

where $\boldsymbol{\eta}$ is the p-dimensional vector of fixed effects, $\boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)$ is the $n_k \times 1$ mean vector, $\boldsymbol{b}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$ is the q-dimensional vector of random effects, $\boldsymbol{Z}_k$ is the design matrix of size $n_k \times q$ for the random effects satisfying $\text{rank}(\boldsymbol{Z}_k) = q \leq n_k$ for all k, and $\boldsymbol{e}_k = (e_{k1}, \ldots, e_{kn_k})^\top \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_k)$ is the n_k-dimensional vector of random errors. The random effects $\boldsymbol{b}_k$ are assumed to be independent of $\boldsymbol{e}_k$. Let $\boldsymbol{V}_k(\boldsymbol{\omega})$ be the variance matrix of $\boldsymbol{y}_k$, $\boldsymbol{V}_k(\boldsymbol{\omega}) = \boldsymbol{Z}_k \boldsymbol{\Psi} \boldsymbol{Z}_k^\top + \boldsymbol{\Sigma}_k$. Here $\boldsymbol{\omega}$ consists of parameters in $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}_k$.

We use (4.1) to model the association between the phenotypic response of interest and genetic and environmental exposure factors. Let G_k be the genotype and E_{kt} be the exposure level for the k-th subject at the t-th measurement, $G_k = i, i = 1, 2, \ldots, I$, $E_{kt} = j, j = 1, 2, \ldots, J$. Both G_k and E_{kt} are assumed to be categorical variables. Without considering any covariates, the mean structure for y_{kt} under Tukey's model (Tukey, 1949) has the following form

$$\begin{aligned} f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = f(\boldsymbol{\beta}, \theta, \boldsymbol{x}_{kt}) = \beta_0 + \sum_{i=1}^{I} \beta_i^G I(G_k = i) + \sum_{j=1}^{J} \beta_j^E I(E_{kt} = j) + \\ \theta \sum_{i=1}^{I} \sum_{j=1}^{J} \beta_i^G \beta_j^E I(G_k = i, E_{kt} = j). \end{aligned} \tag{4.2}$$

Here $\boldsymbol{\eta}$ has two components, $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \theta)^\top$. $\boldsymbol{\beta}$ consists of the intercept β_0, the parameters for genetic main effects, $\boldsymbol{\beta}^G = (\beta_1^G, \ldots, \beta_I^G)^\top$, and exposure main effects, $\boldsymbol{\beta}^E = (\beta_1^E, \ldots, \beta_J^E)^\top$. θ is a scale parameter representing the interaction effect. A saturated interaction model, on the other hand, allows for separate interaction parameters for each GEI configuration:

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = f(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{x}_{kt}) = \beta_0 + \sum_{i=1}^{I} \beta_i^G I(G_k = i) + \sum_{j=1}^{J} \beta_j^E I(E_{kt} = j) + \sum_{i=1}^{I}\sum_{j=1}^{J} \tau_{ij} I(G_k = i, E_{kt} = j), \tag{4.3}$$

where $\boldsymbol{\tau} = (\tau_{11}, \ldots, \tau_{IJ})^\top$ is the interaction parameter vector with length IJ. Due to the constraints for parameter identifiability, $\sum_i \beta_i^G = \sum_j \beta_j^E = 0$, $\boldsymbol{\beta}^G$ and $\boldsymbol{\beta}^E$ are left with $(I-1)$ and $(J-1)$ independent parameters to be estimated, respectively. Similarly, $\sum_i \tau_{ij} = \sum_j \tau_{ij} = 0$, so $(I-1)(J-1)$ parameters in $\boldsymbol{\tau}$ are left to be estimated.

4.3 Parameter Estimation for Tukey's Model with Repeated Measures Data

We describe the estimation strategy for the parameters in Tukey's model. The log-likelihood for the data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ is

$$\ell(\boldsymbol{\eta}, \boldsymbol{\omega}, \sigma^2) \propto \sum_{k=1}^{N} \log|\boldsymbol{V}_k(\boldsymbol{\omega})| - \sum_{k=1}^{N} \Big\{ \big[\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)\big]^\top \boldsymbol{V}_k(\boldsymbol{\omega})^{-1} \big[\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)\big] \Big\}. \tag{4.4}$$

Given $\boldsymbol{V}_k(\boldsymbol{\omega})$, maximizing the likelihood is equivalent to minimizing the objective function

$$\boldsymbol{Q}(\boldsymbol{\eta}|\boldsymbol{\omega}) = \sum_{k=1}^{N} \big[\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)\big]^\top \boldsymbol{V}_k(\boldsymbol{\omega})^{-1} \big[\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)\big] \tag{4.5}$$

with respect to $\boldsymbol{\eta}$. The solution for $\boldsymbol{\eta}$ is the generalized least squares (GLS) estimator. Since the estimation for fixed effects in Tukey's model does not have a closed-form

solution, the iterative linearization method is considered.

The linearization method uses a first-order Taylor series expansion to approximate solutions of a general function by a linear function (Bates and Watts, 1988), which has been applied to nonlinear mixed-effects models (Lindstrom and Bates, 1990; Vonesh and Carter, 1992; Crainiceanu and Ruppert, 2004). Let $\boldsymbol{\eta}^* = \hat{\boldsymbol{\eta}}^{(0)} = (\hat{\boldsymbol{\beta}}^{(0)\top}, \hat{\theta}^{(0)})^\top$ denote the initial estimate of $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \theta)^\top$. The first-order Taylor series expansion of $\boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)$ about $\boldsymbol{\eta} = \boldsymbol{\eta}^*$ is

$$\boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k) \approx \boldsymbol{f}(\boldsymbol{\eta}^*, \boldsymbol{X}_k) + \boldsymbol{D}_k^*(\boldsymbol{\eta} - \boldsymbol{\eta}^*), \tag{4.6}$$

where $\boldsymbol{D}_k^*$ is an $n_k \times p$ matrix $\boldsymbol{D}_k^{*\top} = \boldsymbol{D}_k^\top(\boldsymbol{\eta}^*) = \left\{\partial \boldsymbol{f}(\boldsymbol{\eta})/\partial\eta_1, \ldots, \partial \boldsymbol{f}(\boldsymbol{\eta})/\partial\eta_p\right\}\Big|_{\boldsymbol{\eta}^*}$. Initial values of $\boldsymbol{\eta}^*$ can be obtained by fitting a saturated interaction model (via standard linear mixed effects model) and using the main effect estimates as $\boldsymbol{\beta}^*$. After removing main effects, the residuals can then be regressed on the product term $\beta_i^{G*}\beta_j^{E*}$ (without intercept) to obtain θ^*. The mean function of Tukey's model for the k-th subject at the t-th measurement is

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) \approx f(\boldsymbol{\eta}^*, \boldsymbol{x}_{kt}) + (\beta_0 - \beta_0^*) + \sum_i \sum_j \big[(1 + \theta^*\beta_j^{E*})(\beta_i^G - \beta_i^{G*}) + (1 + \theta^*\beta_i^{G*})(\beta_j^E - \beta_j^{E*}) + \beta_i^{G*}\beta_j^{E*}(\theta - \theta^*)\big] I(G_k = i, E_{kt} = j),$$

where $f(\boldsymbol{\eta}^*, \boldsymbol{x}_{kt}) = \beta_0^* + \sum_i \beta_i^{G*} I(G_k = i) + \sum_j \beta_j^{E*} I(E_{kt} = j) + \theta^* \sum_i \sum_j \beta_i^{G*}\beta_j^{E*} I(G_k = i, E_{kt} = j)$. Following (4.1), the expansion in (4.6) yields the approximation

$$\boldsymbol{y}_k = \boldsymbol{f}(\boldsymbol{\eta}^*, \boldsymbol{X}_k) + \boldsymbol{D}_k^*(\boldsymbol{\eta} - \boldsymbol{\eta}^*) + \boldsymbol{Z}_k\boldsymbol{b}_k + \boldsymbol{e}_k,$$

which can be expressed as a linear model

$$\boldsymbol{y}_k^* = \boldsymbol{D}_k^*\boldsymbol{\eta} + \boldsymbol{Z}_k\boldsymbol{b}_k + \boldsymbol{e}_k, \tag{4.7}$$

where $\boldsymbol{y}_k^* = \boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}^*, \boldsymbol{X}_k) + \boldsymbol{D}_k^*\boldsymbol{\eta}^*$. Then the GLS estimator for $\boldsymbol{\eta}$ is given by

$$\hat{\boldsymbol{\eta}}_{GLS} = \big(\sum_{k=1}^{N} \boldsymbol{D}_k^{*\top} \hat{\boldsymbol{V}}_k^{*-1} \boldsymbol{D}_k^*\big)^{-1} \sum_{k=1}^{N} \boldsymbol{D}_k^{*\top} \hat{\boldsymbol{V}}_k^{*-1} \boldsymbol{y}_k^*, \tag{4.8}$$

where $\hat{\boldsymbol{V}}_k^*$ is the assumed covariance matrix of $\boldsymbol{y}_k^*$ evaluated at $\boldsymbol{\omega} = \boldsymbol{\omega}^*$. When $\boldsymbol{\eta}$ and $\boldsymbol{\omega}$ are unknown, a common strategy is to replace $\boldsymbol{V}(\boldsymbol{\omega})$ with a consistent estimate and minimize the corresponding weighted sum of squares to yield an initial estimate of $\boldsymbol{\eta}$. The MLE of $\boldsymbol{\omega}$ is obtained by maximizing (4.4) with respect to $\boldsymbol{\omega}$, after $\boldsymbol{\eta}$ is replaced by the estimate in (4.8).

This iteratively reweighted generalized least-squares (IRGLS) algorithm involves iterations between [a] Taylor series linearization – given the w-th iterates $\hat{\boldsymbol{\eta}}^{(w)}$ and $\hat{\boldsymbol{\omega}}^{(w)}$, construct $\boldsymbol{D}_k^{(w)} = \boldsymbol{D}(\hat{\boldsymbol{\eta}}^{(w)})$ and $\hat{\boldsymbol{r}}_k^{(w)} = \boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}^{(w)}, \boldsymbol{X}_k) + \boldsymbol{D}_k^{(w)}\hat{\boldsymbol{\eta}}^{(w)}$ to yield a pseudo model that is of the form of (4.7) – and [b] updating estimates $\hat{\boldsymbol{\eta}}^{(w+1)}$ in (4.8) and $\hat{\boldsymbol{\omega}}^{(w+1)}$. Steps [a] and [b] are repeated until a convergence criterion is achieved.

The linearization method provides an easy calculation for nonlinear models by translating the nonlinear estimation problem into a linear model. Only the first-order derivatives are required. Though the assumption of normality is not required for estimates from this IRGLS procedure, minimizing the objective function (4.5) is equivalent to maximizing the joint log-likelihood function of $\boldsymbol{y}_k$ in (4.4). Hence, this procedure yields MLEs (Gallant, 2009). Vonesh et al. (2001) argued that the IRGLS estimator is consistent and asymptotically normal even when the variance-covariance structure is misspecified if the mean function $\boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)$ is correctly specified. Our experience is that the proposed estimation algorithm for Tukey's model converges relatively fast and the final estimates are insensitive to initial values. Nevertheless, seriously slow convergence or possibly non-convergence could occur when one or both of the main effects are truly absent, a situation where θ is not identifiable.

4.4 Shrinkage Estimator

We now construct a shrinkage estimator for interaction that is a weighted average of the estimators from Tukey's model and a saturated interaction model. Denote the interaction parameters to be estimated for an $I \times J$ GEI table by $\boldsymbol{\tau} = (\tau_{11}, \tau_{21}, \ldots, \tau_{(I-1)1}, \tau_{12}, \ldots, \tau_{(I-1)(J-1)})^\top$. Let $\boldsymbol{\tau}_{tuk}$ and $\boldsymbol{\tau}_{sat}$ be the asymptotic limits of the estimator of $\boldsymbol{\tau}$ from Tukey's model and saturated interaction model, respectively, each being a length-$(I-1)(J-1)$ vector. When the true model is a Tukey's one-df model, we have $\boldsymbol{\tau}_{tuk} - \boldsymbol{\tau}_{sat} = \boldsymbol{\delta}(\text{say}) = \mathbf{0}$. To relax the model assumption, let $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta})$. A conservative estimate of $\boldsymbol{\Theta}$ is given by $\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^\top$, where $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\tau}}_{tuk} - \hat{\boldsymbol{\tau}}_{sat}$ and $\hat{\boldsymbol{\tau}}_{tuk} = \hat{\theta}(\hat{\beta}_1^G\hat{\beta}_1^E, \hat{\beta}_2^G\hat{\beta}_1^E, \ldots, \hat{\beta}_{I-1}^G\hat{\beta}_{J-1}^E)^\top$. We define $\boldsymbol{B} = \hat{\boldsymbol{V}}_\tau\big(\hat{\boldsymbol{V}}_\tau + \hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^\top\big)^{-1}$, where $\hat{\boldsymbol{V}}_\tau$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\tau}}_{sat}$. Then the proposed shrinkage estimator for $\boldsymbol{\tau}$ is given by

$$\hat{\boldsymbol{\tau}}_{shk} = \hat{\boldsymbol{\tau}}_{sat} + \boldsymbol{B}(\hat{\boldsymbol{\tau}}_{tuk} - \hat{\boldsymbol{\tau}}_{sat}), \tag{4.9}$$

where $\hat{\boldsymbol{\tau}}_{tuk}$ and $\hat{\boldsymbol{\tau}}_{sat}$ are MLEs from (4.2) and (4.3), respectively.

The shrinkage factor $\boldsymbol{B}$ in (4.9) determines the amount of shrinkage of $\hat{\boldsymbol{\tau}}_{sat}$ toward $\hat{\boldsymbol{\tau}}_{tuk}$. As $\hat{\boldsymbol{\delta}} \to \mathbf{0}$ and $\boldsymbol{B} \to \boldsymbol{I}$, $\hat{\boldsymbol{\tau}}_{shk} \to \hat{\boldsymbol{\tau}}_{tuk}$ (data are indicative of a Tukey's interaction structure). On the other hand, as the bias of Tukey's model estimator $\hat{\boldsymbol{\delta}}$ increases, the largest eigenvalue of $\boldsymbol{B}$ goes to 0 and $\hat{\boldsymbol{\tau}}_{shk} \to \hat{\boldsymbol{\tau}}_{sat}$ (data are not in favor of Tukey's form of interaction). Now express the shrinkage estimator in (4.9) as

$$\hat{\boldsymbol{\tau}}_{shk} = \hat{\boldsymbol{\tau}}_{sat} + \hat{\boldsymbol{V}}_\tau\left(\hat{\boldsymbol{V}}_\tau^{-1} - \frac{\hat{\boldsymbol{V}}_\tau^{-1}\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^\top\hat{\boldsymbol{V}}_\tau^{-1}}{1 + \hat{\boldsymbol{\delta}}^\top\hat{\boldsymbol{V}}_\tau^{-1}\hat{\boldsymbol{\delta}}}\right)\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\tau}}_{sat} + \hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}\left(\frac{\hat{\boldsymbol{\delta}}^\top\hat{\boldsymbol{V}}_\tau^{-1}\hat{\boldsymbol{\delta}}}{1 + \hat{\boldsymbol{\delta}}^\top\hat{\boldsymbol{V}}_\tau^{-1}\hat{\boldsymbol{\delta}}}\right).$$

When data are under Tukey's model, $\hat{\boldsymbol{\delta}} \to \mathbf{0}$ as $N \to \infty$. When data are not under Tukey's model, the largest eigenvalue of $\hat{\boldsymbol{V}}_\tau$ goes to 0 and $\hat{\boldsymbol{\delta}}^\top\hat{\boldsymbol{V}}_\tau^{-1}\hat{\boldsymbol{\delta}} \to \infty$ as $N \to \infty$. So, the term $(\hat{\boldsymbol{\delta}}^\top\hat{\boldsymbol{V}}_\tau^{-1}\hat{\boldsymbol{\delta}})/(1 + \hat{\boldsymbol{\delta}}^\top\hat{\boldsymbol{V}}_\tau^{-1}\hat{\boldsymbol{\delta}})$ converges to 1. This indicates that $\hat{\boldsymbol{\tau}}_{shk}$ is asymptotically equivalent to $\hat{\boldsymbol{\tau}}_{sat}$, which is an unbiased estimator of $\boldsymbol{\tau}$. But with

moderate sample size, $\hat{\boldsymbol{\delta}}$ creates a small bias in $\hat{\boldsymbol{\tau}}_{shk}$ that can be traded for a larger decrease in variance, leading to an improvement in finite sample MSE (Mukherjee and Chatterjee, 2008). In addition, when main effects are not present, the shrinkage estimator will guard against the instability of parameter estimates under Tukey's model by shrinking $\hat{\boldsymbol{\tau}}_{shk}$ toward $\hat{\boldsymbol{\tau}}_{sat}$.

4.4.1 Variance Estimation for the Shrinkage Estimator

We proceed to estimate the covariance matrix for $\hat{\boldsymbol{\tau}}_{shk}$. As a result of asymptotic equivalence of $\hat{\boldsymbol{\tau}}_{shk}$ and $\hat{\boldsymbol{\tau}}_{sat}$, the covariance matrix for $\hat{\boldsymbol{\tau}}_{sat}$ can be used as an estimator for the covariance matrix of $\hat{\boldsymbol{\tau}}_{shk}$ in large samples. Since this estimator is often too conservative in finite samples, we develop an approximate covariance matrix estimator for $\hat{\boldsymbol{\tau}}_{shk}$ using the delta method.

Define $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\tau}}_{sat}^\top, \hat{\boldsymbol{\eta}}_{tuk}^\top)^\top$ as the MLEs under a saturated form of interaction and Tukey's model with $\hat{\boldsymbol{\eta}}_{tuk} = (\hat{\beta}_1^G, \ldots, \hat{\beta}_{I-1}^G, \hat{\beta}_1^E, \ldots, \hat{\beta}_{J-1}^E, \hat{\theta})^\top$. Further define $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\tau}}_{sat}, \hat{\boldsymbol{\tau}}_{tuk})^\top = \boldsymbol{h}(\hat{\boldsymbol{\phi}})$ such that $\hat{\boldsymbol{\tau}}_{shk} = \boldsymbol{g}(\hat{\boldsymbol{\xi}}) = \boldsymbol{g}(\boldsymbol{h}(\hat{\boldsymbol{\phi}}))$, where $\hat{\boldsymbol{\xi}}$ and $\boldsymbol{g}(\hat{\boldsymbol{\xi}})$ have $2(I-1)(J-1)$ and $(I-1)(J-1)$ elements, respectively. We first derive the joint distribution of the components in $\hat{\boldsymbol{\phi}}$. Let $\boldsymbol{\mathcal{I}}$ be the information matrix with dimension $(I-1)(J-1) \times (I-1)(J-1)$ and ℓ be the log-likelihood corresponding to a saturated interaction model (4.3). Let $\boldsymbol{\mathcal{I}}_0$ be the information matrix with dimension $(I+J-1) \times (I+J-1)$ and ℓ_0 be the log-likelihood for Tukey's model (4.2). By the consistency of $\hat{\boldsymbol{\phi}}$, the MLE $\hat{\boldsymbol{\tau}}_{sat}$ has an asymptotic linear representation

$$\sqrt{N}(\hat{\boldsymbol{\tau}}_{sat} - \boldsymbol{\tau}) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \boldsymbol{\mathcal{I}}^{-1} \dot{\ell}_k + o_p(1) \text{ as } N \to \infty, \text{ where } \dot{\ell}_k = \partial \ell(\boldsymbol{X}_k)/\partial \boldsymbol{\tau}.$$

Similarly,

$$\sqrt{N}(\hat{\boldsymbol{\eta}}_{tuk} - \boldsymbol{\eta}) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \boldsymbol{\mathcal{I}}_0^{-1} \dot{\ell}_{0k} + o_p(1) \text{ as } N \to \infty, \text{ where } \dot{\ell}_{0k} = \partial \ell_0(\boldsymbol{X}_k)/\partial \boldsymbol{\eta}_{tuk}.$$

Denote the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\phi}}$ by $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\phi}}}$. Then by multi-

variate Taylor series expansion, the variance-covariance matrix of $\hat{\boldsymbol{\xi}} = \boldsymbol{h}(\hat{\boldsymbol{\phi}})$ is approximated by

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\xi}}} \approx \{\bigtriangledown \boldsymbol{h}(\hat{\boldsymbol{\phi}})\}^{\top} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\phi}}} \bigtriangledown \boldsymbol{h}(\hat{\boldsymbol{\phi}}),$$

where $\bigtriangledown \boldsymbol{h} = \partial \boldsymbol{h}/\partial \boldsymbol{\phi}$ is the gradient matrix of $\boldsymbol{h}$ evaluated at $\hat{\boldsymbol{\phi}}$. Finally, the variance-covariance matrix of $\hat{\boldsymbol{\tau}}_{shk}$ is approximated by applying the delta method:

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\tau}}_{shk}} = \hat{\text{cov}}(\hat{\boldsymbol{\tau}}_{shk}) = \hat{\text{cov}}\big(\boldsymbol{g}(\hat{\boldsymbol{\xi}})\big) \approx \{\bigtriangledown \boldsymbol{g}(\hat{\boldsymbol{\xi}})\}^{\top} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\xi}}} \bigtriangledown \boldsymbol{g}(\hat{\boldsymbol{\xi}}), \tag{4.10}$$

where $\bigtriangledown \boldsymbol{g} = \partial \boldsymbol{g}/\partial \boldsymbol{\xi}$ evaluated at $\hat{\boldsymbol{\xi}}$ (refer to Appendix for $\bigtriangledown \boldsymbol{h}(\hat{\boldsymbol{\phi}})$ and $\bigtriangledown \boldsymbol{g}(\hat{\boldsymbol{\xi}})$.) Comparing $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\tau}}_{shk}}$ to the empirical estimate of variance-covariance matrix through simulations, we found that variance components can be estimated very well by $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\tau}}_{shk}}$ but not necessarily the covariance. Either a small variance for the random measurement errors or a large sample size is needed to obtain accurate estimates of covariance terms (see Table 4.5 in Appendix). Since the magnitudes of covariance estimates are smaller compared to the variance estimates, the influence of covariance estimates on the Wald test statistic is expected to be small. Thus, the proposed shrinkage test (see below) is still an approximately valid test with conservative Type 1 error rates.

4.5 Tests for Interaction Effects

We are interested in testing the null hypothesis of no interaction effects $H_0 : \boldsymbol{\tau} = \mathbf{0}$ versus $H_1 : \boldsymbol{\tau} \neq \mathbf{0}$. For Tukey's model, it is equivalent to $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. A likelihood ratio test (LRT) statistic is given by $T_L = -2(l_0 - l_1)$, where l_0 and l_1 are the maximized log-likelihoods obtained under H_0 and H_1, respectively. Under regularity conditions, $T_L \sim \chi^2_1$ for Tukey's model and $T_L \sim \chi^2_{(I-1)(J-1)}$ for saturated model under H_0 for large samples. Based on (4.9) and Chen et al. (2009), the limiting distribution of the shrinkage estimator is technically not normal. The simulation results, however, reveal that this estimator is well approximated by a normal density and the amount of departure from normality is small (see Figure 4.1 in Appendix).

Hence, the Wald test is used as an approximate test for interaction. The test statistic for $H_0 : \boldsymbol{\tau} = \mathbf{0}$ is given by $\tilde{T}_W = \hat{\boldsymbol{\tau}}_{shk}^{\top} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\tau}}_{shk}}^{-1} \hat{\boldsymbol{\tau}}_{shk}$, where $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\tau}}_{shk}}$ can be found in (4.10).

4.6 Simulation Study

4.6.1 Evaluation of Test Properties for a Single GEI Test

We investigated the Type I error and power properties of the following three test procedures for interaction: the LRT under Tukey's model of interaction, the Wald test using the proposed adaptive shrinkage estimator, and the LRT using a saturated interaction model. Two null hypotheses of no interactions were considered: (i) the genetic main effects were present (additive) and (ii) the genetic main effect were absent (null). The main effects of the exposure were always present in our simulations to represent a study looking for genetic modification effects on an established phenotype-exposure association. For these comparisons, we used 3×3 table settings for GEI with N=1200. The number of repeated measurements per subject was generated from a multinomial distribution similar to the example data: $n_{ijk} \in \{2, 3, 4, 5, 6\}$, $\boldsymbol{n} = \{n_{ijk} : 1 \leq k \leq N_{ij}, 1 \leq i \leq I, 1 \leq j \leq J\} \sim \text{Mult}(N, \boldsymbol{p})$, $\boldsymbol{p} = (0.15, 0.2, 0.3, 0.2, 0.15)$, which implies that dropouts are missing completely at random. Data were were simulated under a first-order autoregressive (AR-1) correlation structure for $\boldsymbol{\Sigma}_k$ ($\sigma^2 = 4, 8$ and $\rho = 0.7$). Additionally, the test properties were evaluated under misspecification of correlation structure. Again, data were still generated under the AR-1 correlation structure but were analyzed using a compound symmetric covariance structure. A total of 1000 datasets were generated for each setting. Type I error and power were estimated by the sample proportions of null hypothesis being rejected under various simulation settings.

In the 3×3 GEI table settings, three genotype categories were considered for G with minor allele frequency 0.4 and following the Hardy-Weinberg equilibrium. An environmental exposure with three categories (with probabilities 0.25, 0.25, and 0.50)

was considered. Cell means for all GEI configurations were first generated under a pre-specified interaction model. Given a mean and covariance structure, the vector of observations per individual were generated from a multivariate normal distribution. In addition to Tukey's and saturated models, we considered simulations under additive main effects and multiplicative interaction models (Gollob, 1968; Mandel, 1971). AMMI models are a class of interaction models that have a flexible structure, which essentially entails a singular value decomposition (SVD) of the cell residual matrix after removing the additive main effects. Following the notations in (4.2), the mean structure for y_{kt} under an AMMI model is given by

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = f(\boldsymbol{\beta}, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}_{kt}) = \beta_0 + \sum_{i=1}^{I} \beta_i^G I(G_k = i) + \sum_{j=1}^{J} \beta_j^E I(E_{kt} = j) + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{m=1}^{M} d_m \alpha_{im} \gamma_{jm} I(G_k = i, E_{kt} = j).$$

The m-th interaction factor is subject to the constraints $\sum_{i=1}^{I} \alpha_{im}^2 = \sum_{j=1}^{J} \gamma_{jm}^2 = 1$ and $\sum_{i=1}^{I} \alpha_{im} = \sum_{j=1}^{J} \gamma_{jm} = 0$, as well as the $2(M-1)$ orthogonality restrictions $\sum_i \alpha_{im}\alpha_{im'} = \sum_i \gamma_{jm}\gamma_{jm'} = 0$ for $m \neq m'$. Specifically, AMMI models with $M = 1$ (AMMI1) were considered in the simulation as an intermediate model between Tukey and saturated model. AMMI2 would be equivalent to a saturated interaction model in the 3×3 table settings. We compared test performance under AMMI1 models because Tukey's test may not be capable of capturing interaction of AMMI1 form. Though AMMI1 is nested within the saturated interaction model, the test based on a saturated interaction model may not have as much power to detect the interaction.

4.6.2 Assessment of Average Performance for Multiple GEI Tests

When GEI tests are conducted across a moderately large number of SNPs within several gene regions, the average performance of each method over many GEI tests is of particular interest rather than a single specific GEI test. As such, we assessed

the Type I error and power of the tests for interaction using Tukey's model, saturated interaction model, and the proposed shrinkage estimator, averaged over a set of genetic markers. We based our simulation studies on the setting of the NAS data example where the candidate genes were chosen based on some pathway analysis. For each dataset, one exposure factor and 100 independent SNPs (without LD) were generated, with the minor allele frequencies ranging from 0.3 to 0.5. The exposure had five categories, each with probability 0.2. Thus, a 3×5 table was constructed for each GEI test.

We considered two simulation schemes for multiple GEI tests: (i) 100 marginal models, $Y_i|G_i, E, i = 1, \ldots, 100$, were generated with a common E for each subject, and (ii) a joint multivariate model, $Y|G_1, G_2, \ldots, G_{100}, E$, was generated. In both (i) and (ii), 15 out of 100 SNPs were assigned to have GEI effects on Y. Another five SNPs were generated to have only additive main effects on Y. The rest 80 SNPs were not associated with Y. The simulation design represents a study where GEI over multiple SNPs are being tested, the majority of SNPs do not have GEI effects and only a relatively small number of SNPs exhibit GEI effects.

To assess the sensitivity of tests in response to the underlying composition of different interaction models, we created three scenarios by assigning each of the 15 GEI to have either a Tukey's or a saturated form of interaction: Scenario (A): all 15 were of Tukey's form of interaction; Scenario (B): 10 were of Tukey's form, and 5 had saturated interaction structures; Scenario (C): 10 had saturated interaction structures, and 5 were of Tukey's form. For example, the mean function of the simulation model for subject k under scenario (B) in simulation scheme (ii), following the notations in (4.3), is given by

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = \beta_0 + \sum_{j=1}^{J} \beta_j^E I(E_{kt} = j) + \sum_{i=1}^{I} \sum_{s=1}^{5} \beta_i^{Gs} I(G_{sk} = i)$$
$$+ \sum_{i=1}^{I} \sum_{j=1}^{J} \Bigg\{ \sum_{s=6}^{15} \left[\beta_i^{Gs} I(G_{sk} = i) + \theta^s \beta_i^{Gs} \beta_j^E I(G_{sk} = i, E_{kt} = j)\right]$$
$$+ \sum_{s=16}^{20} \left[\beta_i^{Gs} I(G_{sk} = i) + \tau_{ij}^s (G_{sk} = i, E_{kt} = j)\right] \Bigg\},$$

where β_i^{Gs} represents the genetic main effect of the i-th genotype from the s-th SNP, G_{sk} is the genotype of the s-th SNP for the k-th subject, and θ^s and τ_{ij}^s are the interaction parameter corresponding to the s-th SNP. An individual-level outcome Y with repeated measures were generated for 1000 subjects in each simulation using (4.1) with $\boldsymbol{e}_k \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I}_{n_k}), \boldsymbol{b}_k = b_k \mathbf{1}_{n_k}, b_k \sim \mathcal{N}(0, \sigma_b^2)$. We set $\sigma_b^2 = 2.8$, and $\sigma_e^2 = 1.2$. The number of repeated measurements per subject was generated using the same multinomial distribution described previously.

The average performance for each test procedure was quantified by true positive rate (TPR) and false positive rate (FPR). The TPR is defined as the proportion of interactions detected in the 15 simulated SNPs with GEI associations. The FPR is the proportion of interactions detected among the 85 simulated SNPs without GEI effects. The TPR and FPR were then averaged over 10,000 simulation datasets. To control the family-wise error rate (FWER), the significance level was adjusted according to the total number of SNPs (i.e., number of GEI tests) using Bonferroni correction, $\alpha^* = 0.05/100 = 5 \times 10^{-4}$.

4.6.3 Power and Type I Error

The upper panel of Table 4.1 shows the power and Type I error of tests using Tukey's, the saturated model, and the shrinkage estimator for GEI. In general, the saturated interaction model has less power to detect interactions when the true interaction has a Tukey's form. For example, the LRT for Tukey's form of interaction

has power 0.76 for $\sigma^2 = 4$, while the saturated model has a power of 0.54. On the other hand, when the true interaction has a saturated form, Tukey's model can hardly detect the interaction effects. The saturated model has a power of 0.81 for $\sigma^2 = 4$, but Tukey's model using the LRT only has power 0.09. Under both situations, the interaction test using the shrinkage estimator has power 0.69. When the true interaction has an AMMI1 form, the saturated interaction and the shrinkage estimator can detect 82% and 72% of interactions, respectively, but Tukey's model can only detect 30% of interactions. The Type I error rates are maintained at the nominal level for all testing procedures under additive models except the Wald test using the shrinkage estimator being a slightly conservative test. However, both Tukey's test and the shrinkage estimator have inflated Type I error under the completely null model when one of the main effects is not present.

When the within-subject correlation structure is misspecified (lower panel of Table 4.1), the patterns of power comparison are similar to the upper panel. Under the null hypothesis of an additive model where both main effects are present, the Type I error rates for the two LRTs are still maintained at the 0.05 level when $\sigma^2 = 4$ but are inflated when $\sigma^2 = 8$. Only the proposed Wald test using the shrinkage estimator maintains the nominal level of Type I erorr. Under the null that genetic main effects are absent, the Type I error is no longer maintained at 0.05 for all of the tests.

4.6.4 Average Performance for Multiple GEI Tests

The upper panel of Table 4.2 shows the average performance of the three GEI tests for marginal models under three scenarios. Under scenario (A) where all 15 simulated GEI are of Tukey's form, the LRT using Tukey's model has a TPR of 0.72, whereas the saturated model has a TPR of 0.43. Under scenario (B) where 2/3 of the simulated GEI are of Tukey's form, the LRT using Tukey's model and the saturated interaction test have comparable performance. Under scenario (C) where 2/3 of

the interactions are of saturated forms, the Wald test using the shrinkage estimator and the saturated interaction tests have comparable performance, but the TPR for the LRT using Tukey's model is substantially lower. The FPRs are maintained at the nominal level for the tests using a saturated model and slightly inflated for the shrinkage estimator. However, the LRT for Tukey's model has the highest FPR.

The lower panel of Table 4.2 shows the results of a multivariate model (single outcome) from 100 simulated GEI. The LRT using a saturated interaction form yields relatively low TPRs. The test based on the shrinkage estimator still maintains at the same level of TPR across scenarios. In summary, the GEI test using the shrinkage estimator has the most robust average performance with respect to various GEI structures compared to the tests using Tukey's and saturated interaction models.

4.7 Application

4.7.1 Normative Aging Study (NAS)

The Normative Aging Study (NAS) is a multidisciplinary longitudinal study initiated by the U.S. Veterans Administration in 1963 to investigate the effects of aging on various health outcomes (Bell et al., 1966). We focus on pulse pressure (PP), which is an important risk factor for heart disease (Franklin et al., 1999). Several studies have indicated a relationship between iron deficiency and increased lead absorption (Kwong et al., 2004; Bradman et al., 2001), and increased cumulative lead exposure has been shown to be associated with elevated PP (Perlstein et al., 2007). Thus, it may be reasonable to hypothesize that genes responsible for iron metabolism could potentially alter lead absorption and modify the effect of lead exposure on PP. The objective of this pathway-driven GEI study was to test the GEI between cumulative lead exposure and the iron metabolic genes on PP.

Zhang et al. (2010) observed a significant interaction between polymorphisms in the hemochromatosis (*HFE*) gene (rs1799945) and cumulative lead exposure on PP.

We revisited the study to include 105 SNPs in 22 genes with minor allele frequency>0.1 in the iron metabolic pathway to test for GEI using the proposed shrinkage estimation framework. Candidate genes were chosen based on *a priori* knowledge of iron metabolism and previous studies on iron-related genes (Knutson and Wessling-Resnick, 2003; Chung and Wessling-Resnick, 2003). We analyzed 729 participants from a subset of the NAS data who were successfully genotyped for the iron metabolism genes and had baseline measurements of cumulative lead concentrations (measured at the tibia bone and patella bone). The majority (97%) of the participants were Caucasian. The average age was 66.37±7.12 (range 48–93) at the time of bone lead measurement. Since 1991, blood pressure had been measured every 3–5 years until 2011 with a median follow-up time of 12 years. More than 94% of subjects had repeated measurements of blood pressure, and over 48% of them had at least four measurements during the study period contributing to a total of 3013 observations (see Table 4.8 in Appendix).

Each of the 105 SNPs had three possible genotypes (homozygous wild-type, heterozygous, and homozygous mutant). For illustration purposes, we categorized bone lead concentrations into three groups – Low: $\leq$15, Medium: (15, 25], and High: >25 μg/g for the tibia bone lead and Low: $\leq$22, Medium: (22, 30], and High: >30 μg/g for the patella bone lead. We used Tukey's model, saturated interaction model, and the shrinkage approach to model the GEI structures for each *SNP*×*Lead* interaction. Covariates in the model included baseline age, time since baseline, and squared time. According to the Akaike information criterion (AIC) for model fit, we chose a random-intercept mixed-effects model for analysis given by $\boldsymbol{y}_k = f(\boldsymbol{\eta}, \boldsymbol{X}_k) + b_k \mathbf{1}_{n_k} + \boldsymbol{e}_k$, where $b_k \sim \mathcal{N}(0, \sigma_b^2), \boldsymbol{e}_k \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I}_{nk})$.

Given that these SNPs are located in a small number of genomic regions, they are in close proximity to each other and thus may exhibit LD. To control for the FWER while accounting for the potentially correlated SNPs in the multiple testing

procedure, we adjusted the significance level according to the effective number of independent tests (denoted by M_{eff}) using the simple$M$ method (Gao et al., 2008). This method involves first estimating the correlation matrix among the 105 SNPs by the composite LD, calculating the corresponding eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{105}$, and then finding M_{eff} through principal component analysis: $\sum_{s=1}^{M_{\text{eff}}} \lambda_s / \sum_{s=1}^{105} \lambda_s > C$. We chose $M_{\text{eff}} = 89$ so that the corresponding eigenvalues explained at least $C = 99.5\%$ of the variation for the SNP data. Thus, the adjusted significance level was $\alpha^* = 0.05/M_{\text{eff}} = 0.05/89 = 5.6 \times 10^{-4}$.

Table 4.3 lists the smallest p-values of GEI tests for the three top-ranked SNPs by using Tukey's model, the proposed shrinkage estimator, and saturated interaction model within iron gene regions in the NAS data. The Wald test via the shrinkage estimator yielded the smallest p-values across all top ranked SNPs listed in the table (and three of which reached statistical significance), compared to Tukey's and saturated interaction models. For tibia bone lead, we found a significant modifying effect of SNP rs1799945 in the *HFE* gene using the shrinkage estimator ($p = 1 \times 10^{-4}$). For the wild-type participants, mean PP remained nearly unchanged between the High and the Low tibia lead groups. In contrast, mean PP was estimated to be 20.35 mmHg (95% CI = [14.53, 26.17]) higher for the High tibia lead group than the Low tibia lead group among the homozygous mutant carriers. The results replicate the findings in Zhang et al. (2010) that the positive association between PP and lead exposure was strongest among *HFE* homozygous mutant carriers. For patella bone lead, significant modifying effects of SNP rs17484524 in the IREB2 (iron-responsive element binding protein 2) gene ($p = 3 \times 10^{-4}$) and SNP rs7165535 in the B2M (beta-2-microglobulin) gene ($p = 4 \times 10^{-4}$) were detected using the Wald test based on the shrinkage estimator (but were not captured by the LRTs using Tukey's or saturated interaction model). For the wild-type and the heterozygous mutant participants, higher lead levels corresponded to higher mean PP (the estimated difference in mean

PP between High and Low patella lead groups ranged from 3.12 to 4.32 mmHg at both SNPs). However, mean PP was estimated to be 3.90 (95% CI = [1.45, 6.35]) and 7.73 (95% CI = [1.88, 13.58]) mmHg *lower* for the High lead group than the Low lead group among the homozygous mutant carriers at SNP rs17484524 in the IREB2 gene and SNP rs7165535 in the B2M gene, respectively. As such, the two homozygous mutant genotypes may indicate protective effects (i.e., preventing PP from elevating with increased lead exposure).

4.7.2 Multi-Ethnic Study of Atherosclerosis (MESA)

The Multi-Ethnic Study of Atherosclerosis (MESA) is a longitudinal study to investigate characteristics related to progression of subclinical to clinical cardiovascular disease (Bild et al., 2002). More than 6,800 men and women aged 45-84 years were recruited from six U.S. communities. Participants had a baseline examination (exam 1) in 2000-2002 and three additional follow-up examinations 18-24 months apart (exams 2-4). We aimed to explore GEI effects on BMI in the four race groups: Caucasians (N=2526), Chinese (N=775), African Americans (N=1611), and Hispanics (N=1449). Most (84%) of the participants had four BMI measurements, and over 92% had at least two measurements during the study period from 2000 to 2007 (see Table 4.9 in Appendix). A total of 27 SNPs that have demonstrated significant and replicated evidence of marginal association with BMI were selected as the candidate SNPs (Speliotes et al., 2010). The environmental exposures of interest were energy intake, measured at exam 1, and total intentional exercise, measured at exams 1–3. Both exposure variables were categorized into five groups: 0, (0, 7], (7, 14], (14, 28], >28 (hr/week) for total intentional exercise and <1000, (1000, 1300], (1300, 1600], (1600, 2000], >2000 (kcal/day) for energy intake.

We applied Tukey's model, saturated interaction model, and the shrinkage test to examine the GEI structure for each SNP×Energy Intake and SNP×Exercise interac-

tion. Covariates in the model included age at the time of data collection (centered), squared age, gender, having a college degree, household income, and the exposure variable (either intentional exercise or energy intake). We also accounted for population stratification by including the first two principal components. Except age, BMI, and intentional exercise that changed with time, all other variables were time-invariant. We chose an unstructured covariance matrix for this analysis based on AIC. A random gender effect was added to allow men and women to have different variances in BMI. Let $\mathbf{F} = \mathbf{1}_{n_k}$ for women and $\mathbf{F} = \mathbf{0}_{n_k}$ for men. The analysis model is given by $\boldsymbol{y}_k = f(\boldsymbol{\eta}, \boldsymbol{X}_k) + \mathbf{F}_k b_k + \boldsymbol{e}_k$, where $b_k \sim (0, \sigma_b^2), \boldsymbol{e}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$. We first analyzed data by race group (see Table 4.10 in Appendix) and then applied Fisher's method (Fisher, 1925) to combine four race groups into a single meta-analysis p-value for each SNP. Not every race group allowed for GEI tests across all 27 SNPs because of small sample size in certain GEI configurations. The df for deriving the combined p-values was based on the number of available race groups. The adjusted p-value to control for the FWER was set at $0.05/27 = 0.0019$.

Table 4.10 lists the combined p-values for significant SNPs using the three interaction tests. For the association of energy intake with BMI, significant modifying effect of SNP rs543874 on the *SEC16B* gene was observed using all three tests. SNP rs1558902 within the *FTO* gene was detected by Tukey's model ($p = 4.8 \times 10^{-5}$) and the shrinkage test ($p = 7.4 \times 10^{-4}$). SNP rs10767664 (on the *BDNF* gene) was also detected by Tukey's model ($p = 1.2 \times 10^{-3}$). For the association between intentional exercise and BMI, we found significant modifying effect of SNP rs206936 within the *NUDT3* and *HMGA1* genes using Tukey's model ($p = 1.4 \times 10^{-4}$). Overall, only one interaction was detected by a standard saturated interaction model used in the current practice. Both the examples illustrate the utility of enhancing power of a test for interaction by leveraging Tukey's model. The shrinkage estimator also offers protection against false positive.The findings require further replication studies.

4.8 Discussion

We proposed a novel adaptive shrinkage estimator that combines estimates from Tukey's one-df model and a saturated interaction model for GEI effects. The shrinkage estimator shrinks the MLEs under a general, saturated interaction structure toward Tukey's one-df model estimator that allows for data-adaptive relaxation of the structural assumption in Tukey's product form.

The unique simulation setting of multiple GEI tests represents the search for GEI over many candidate SNPs with different interaction patterns. The results indicate that the test based on the shrinkage estimator can be considered as a robust and unified approach for interaction detection. More importantly, the shrinkage method not only can be applied to the context of GEI or GGI detection but also can be extended to any two-way table.

We evaluated MSE and bias of these estimators of interaction effects through simulations (Table 4.6 in Appendix). The performance of the shrinkage estimator was compared with the MLE under a general saturated interaction model using the ratio of MSE, $\hat{E}\Big\{ \sum_i \sum_j (\hat{\tau}_{shk_{ij}} - \tau_{ij})^2 \Big\} / \hat{E}\Big\{ \sum_i \sum_j (\hat{\tau}_{sat_{ij}} - \tau_{ij})^2 \Big\}$. Based on simulation results, the ratio is uniformly less than 1, suggesting an efficiency advantage for the shrinkage estimator via bias-variance trade-off. In our simulation studies, we noted that the Wald test using the shrinkage estimator is slightly conservative, so the small bias of the shrinkage estimator in finite samples does not lead to inflated Type I error. In addition, we compared the shrinkage estimates of interaction parameters using only the diagonal elements of $\boldsymbol{B}$ (i.e., scalar shrinkage) versus using the whole $\boldsymbol{B}$ matrix (i.e., multivariate shrinkage). We found that multivariate shrinkage is required under certain situations (see Table 4.7 in Appendix). Chen et al. (2009) proposed both multivariate and scalar shrinkage estimators in case-control studies, and they also found that the scalar shrinkage estimator can lead to appreciable bias.

Although the methods we discussed have been developed for a two-dimensional

interaction structure (i.e., the genetic and interaction effects are assumed to be invariant with time), they can be easily modified to allow for time-dependent effects. To allow for temporal changes in the main effects and interaction effects, one may use spline functions. For example, the mean function for Tukey's model at time (or age) of measurement t can be expressed as

$$f(\boldsymbol{\eta}(t), \boldsymbol{x}_{kt}) = f(\boldsymbol{\beta}(t), \theta(t), \boldsymbol{x}_{kt}) = \beta_0(t) + \beta^G(t)g_k + \beta^E(t)e_{kt} + \theta(t)\beta^G(t)\beta^E(t)g_k e_{kt},$$

where the genotype g_k and the exposure variable e_{kt} for subject k at time t can be treated as continuous, $\beta_0(t)$ is the baseline function, $\beta^G(t)$ and $\beta^E(t)$ are the time-varying genetic and exposure function, and $\theta(t)$ is the time-varying interaction function. These functions can be approximated by a linear combination of basis functions (Hoover et al., 1998). We plan to address the issues of estimation and testing for the temporal dynamic changes in interaction effects using alternative models in future studies.

We have proposed a new approach in the area of longitudinal GEI cohort studies. The Tukey's one-df test for non-additivity can be very powerful in terms of detecting GEI for studies where the search for GEI is based on the presence of genetic main effects (e.g., MESA), but the test can suffer from misspecification of interaction structure. The proposed shrinkage estimation procedure, on the other hand, is useful for pathway-driven GEI studies (e.g., NAS) where there is no prior knowledge of the existence of genetic main effects. It also performs well across many scenarios. Despite the advantage of efficiency, the adaptive shrinkage estimation approach still uses the same df for interaction parameters as a saturated model. As such, the increase in power by shrinking parameter estimates toward Tukey's model estimates may be limited. However, the robust performance across multiple loci with different interaction structures remain an appealing feature of such adaptive screening tests.

Table 4.1: Power for detecting GEI and Type I error rates using Tukey's model, the proposed adaptive shrinkage estimator, and saturated interaction models under different interaction structures in 3×3 table settings (N=1200). Data were simulated under an autoregressive-1 (AR-1) correlation structure while analysis was performed under correctly specified and misspecified correlation structures (see Section 4.6.1 for simulation details).

	$\sigma^2 = 4$			$\sigma^2 = 8$		
Test Model / True Model	Tukey LRT	Shrinkage Wald	Saturated LRT	Tukey LRT	Shrinkage Wald	Saturated LRT
Correctly Specified Correlation Structure (AR-1)						
Tukey's one-df	0.758	0.686	0.540	0.479	0.409	0.273
AMMI1	0.303	0.720	0.817	0.211	0.398	0.514
Saturated	0.094	0.690	0.806	0.086	0.325	0.459
$H_0 : \theta = 0$ (Additive)	0.047	0.042	0.053	0.053	0.043	0.051
$H_0 : \theta = 0$ (Null)	0.104	0.081	0.049	0.107	0.087	0.052
Misspecified Correlation Structure (Compound Symmetric)						
Tukey's one-df	0.730	0.640	0.496	0.435	0.376	0.244
AMMI1	0.287	0.708	0.792	0.185	0.372	0.489
Saturated	0.060	0.646	0.775	0.065	0.308	0.432
$H_0 : \theta = 0$ (Additive)	0.048	0.046	0.053	0.070	0.048	0.060
$H_0 : \theta = 0$ (Null)	0.118	0.064	0.053	0.143	0.065	0.061

Table 4.2: Average performance of tests using Tukey's model, saturated interaction model, and the adaptive shrinkage estimator for detecting GEI across 100 simulated SNPs under scenarios (A): all simulated GEI are of Tukey's form, (B): 2/3 of simulated GEI are of Tukey's form and 1/3 are of saturated form, and (C): 2/3 of simulated GEI are of saturated form and 1/3 are of Tukey's form

Measure	Scenario	Tukey LRT	Shrinkage Wald	Saturated LRT
Marginal Models				
True Positive Rate	(A)	0.7221	0.5766	0.4302
	(B)	0.5611	0.6317	0.5769
	(C)	0.4923	0.6699	0.7357
False Positive Rate		0.0024	0.0007	0.0006
Multivariate Models				
True Positive Rate	(A)	0.3264	0.2810	0.0706
	(B)	0.2882	0.2602	0.2247
	(C)	0.2073	0.2507	0.2911
False Positive Rate		0.0045	0.0027	0.0006

Table 4.3: The p-values of GEI tests for the top three (ranks in parentheses) single-nucleotide polymorphisms (SNPs) by using Tukey's model, the proposed shrinkage estimator, and saturated interaction model within iron gene regions in the NAS data (adjusted $\alpha = 5.6 \times 10^{-4}$).

Bone Lead	SNP ID	Gene	Tukey LRT	Shrinkage Wald	Saturated LRT
Tibia	rs1799945	HFE	0.003 (1)	1×10^{-4} (1)	0.006 (1)
	rs2285228	DMT1	0.005 (2)	0.001 (2)	0.017 (2)
	rs3821716	MFI2	0.014 (3)	0.012	0.120
	rs422982	DMT1	0.016	0.003 (3)	0.072 (3)
Patella	rs7165535	B2M	0.001 (1)	4×10^{-4} (2)	0.014 (1)
	rs17484524	IREB2	0.002 (2)	3×10^{-4} (1)	0.021 (2)
	rs7866419	ACO1	0.009 (3)	0.005	0.054
	rs1358024	TF	0.016	0.004 (3)	0.038
	rs2304704	SLC40A1	0.044	0.030	0.034 (3)

Table 4.4: Findings of GEI with significant meta-analysis p-values for the single-nucleotide polymorphisms (SNPs) that have demonstrated significant and replicated evidence of marginal association with BMI in the MESA data (adjusted $\alpha = 1.9 \times 10^{-3}$).

Exposure	SNP ID	Gene	Tukey LRT	Shrinkage Wald	Saturated LRT
Energy Intake	rs543874	SEC16B	$<1.0\times10^{-8}$	$<1.0\times10^{-8}$	1.8×10^{-4}
	rs1558902	FTO	4.8×10^{-5}	7.4×10^{-4}	0.130
	rs10767664	BDNF	1.2×10^{-3}	0.103	0.124
Exercise	rs206936	NUDT3,HMGA1	1.4×10^{-4}	0.006	0.005

4.9 Appendix

4.9.1 Variance Estimation for the Shrinkage Estimator

The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\phi}}$, denoted by $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\phi}}}$, is given by

$$\begin{bmatrix} \boldsymbol{\mathcal{I}}^{-1}\text{var}\big(\sum_k \dot{\ell}_k\big)(\boldsymbol{\mathcal{I}}^{-1})^\top & \boldsymbol{\mathcal{I}}^{-1}\text{cov}\big(\sum_k \dot{\ell}_k, \sum_k \dot{\ell}_{0k}\big)(\boldsymbol{\mathcal{I}}_0^{-1})^\top \\ \boldsymbol{\mathcal{I}}_0^{-1}\text{cov}\big(\sum_k \dot{\ell}_{0k}, \sum_k \dot{\ell}_k\big)(\boldsymbol{\mathcal{I}}^{-1})^\top & \boldsymbol{\mathcal{I}}_0^{-1}\text{var}\big(\sum_k \dot{\ell}_{0k}\big)(\boldsymbol{\mathcal{I}}_0^{-1})^\top \end{bmatrix},$$

where $\dot{\ell}_k$ and $\dot{\ell}_{0k}$ are the individual score functions for the saturated interaction model and Tukey's model, respectively. When $N \to \infty$, $\text{var}(\sum_k \dot{\ell}_k) \to \boldsymbol{\mathcal{I}}$ and $\text{var}(\sum_k \dot{\ell}_{0k}) \to \boldsymbol{\mathcal{I}}_0$. To estimate $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\phi}}}$, we replace $\boldsymbol{\mathcal{I}}$ and $\boldsymbol{\mathcal{I}}_0$ by observed information matrices evaluated at $\hat{\boldsymbol{\phi}}$. The covariance of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\phi}}}$ can be estimated as

$$\hat{\text{cov}}\big(\sum_k \dot{\ell}_k, \sum_k \dot{\ell}_{0k}\big) = \frac{1}{N^2}\Big\{\sum_k \dot{\ell}_k(\hat{\boldsymbol{\eta}}_{sat})[\dot{\ell}_{0k}(\hat{\boldsymbol{\eta}}_{tuk})]^\top - \frac{1}{N}\sum_k \dot{\ell}_k(\hat{\boldsymbol{\eta}}_{sat})\sum_k \dot{\ell}_{0k}(\hat{\boldsymbol{\eta}}_{tuk})\Big\}$$

and its transpose. Recall that

$$\boldsymbol{h}(\hat{\boldsymbol{\phi}}) = \hat{\boldsymbol{\xi}} = \begin{bmatrix} \hat{\boldsymbol{\tau}}_{sat} \\ \hat{\boldsymbol{\tau}}_{tuk} \end{bmatrix}.$$

The gradient matrix $\bigtriangledown \boldsymbol{h}(\hat{\boldsymbol{\phi}})$ is given by

$$\bigtriangledown \boldsymbol{h}(\hat{\boldsymbol{\phi}}) = \begin{bmatrix} \boldsymbol{I}_{(I-1)(J-1)} & & & \mathbf{0} & & & \\ & \hat{\lambda}\hat{\beta}_1^E & 0 & \dots & \hat{\lambda}\hat{\beta}_2^E & 0 & \dots \\ & 0 & \hat{\lambda}\hat{\beta}_1^E & \dots & \hat{\lambda}\hat{\beta}_2^E & 0 & \dots \\ \mathbf{0} & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & \hat{\lambda}\hat{\beta}_1^G & \hat{\lambda}\hat{\beta}_2^G & \dots & 0 & 0 & \dots \\ & 0 & 0 & \dots & \hat{\lambda}\hat{\beta}_1^G & \hat{\lambda}\hat{\beta}_2^G & \dots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & \hat{\beta}_1^G\hat{\beta}_1^E & \hat{\beta}_2^G\hat{\beta}_1^E & \dots & \dots & \dots & \dots \end{bmatrix}.$$

Next, we want to derive $\bigtriangledown \boldsymbol{g}(\hat{\boldsymbol{\xi}})$. The shrinkage estimator can be expressed as

$$\begin{aligned}
\hat{\boldsymbol{\tau}}_{shk} = g(\hat{\boldsymbol{\xi}}) &= g(\hat{\boldsymbol{\tau}}_{tuk}, \hat{\boldsymbol{\tau}}_{sat}) = \hat{\boldsymbol{\tau}}_{sat} + \boldsymbol{B}(\hat{\boldsymbol{\tau}}_{tuk} - \hat{\boldsymbol{\tau}}_{sat}) \\
&= \hat{\boldsymbol{\tau}}_{sat} + \hat{\boldsymbol{V}}_{\tau}(\hat{\boldsymbol{V}}_{\tau} + \hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\top})^{-1}\hat{\boldsymbol{\delta}} \\
&= \hat{\boldsymbol{\tau}}_{sat} + \hat{\boldsymbol{V}}_{\tau}\left(\hat{\boldsymbol{V}}_{\tau}^{-1} - \frac{\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}}{1 + \hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}}\right)\hat{\boldsymbol{\delta}} \\
&= \hat{\boldsymbol{\tau}}_{sat} + \hat{\boldsymbol{\delta}} - \frac{\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}}{1 + \hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}} \\
&= \hat{\boldsymbol{\tau}}_{tuk} - \frac{\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}}{1 + \hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}}, \quad \hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\tau}}_{tuk} - \hat{\boldsymbol{\tau}}_{sat}.
\end{aligned}$$

Then the $(I-1)(J-1) \times 2(I-1)(J-1)$ matrix $\bigtriangledown \boldsymbol{g}(\hat{\boldsymbol{\xi}}) = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\xi}}\Big|_{\hat{\boldsymbol{\xi}}}$ is given by

$$\bigtriangledown \boldsymbol{g}(\hat{\boldsymbol{\xi}}) = \left[\frac{-2\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}}{1 + \hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}} + \frac{1}{1 + \hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}}\boldsymbol{I}_{(I-1)(J-1)} \quad \frac{\hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}}{1 + \hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}}\boldsymbol{I}_{(I-1)(J-1)} + \frac{2\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}}{1 + \hat{\boldsymbol{\delta}}^{\top}\hat{\boldsymbol{V}}_{\tau}^{-1}\hat{\boldsymbol{\delta}}}\right].$$

4.9.2 Estimates of Variance and Covariance Components for the Shrinkage Estimator

We compared model-based covariance estimates with empirical covariance estimates corresponding to the shrinkage estimator in a simulation study. The estimates for the off-diagonal entries in the dispersion matrix do not work uniformly well across simulation scenarios as the variance estimates of the diagonal entries of the same matrix. We noted that a much larger sample is required to obtain unbiased estimates of the covariance terms. Table 4.5 shows the simulation results of comparisons between empirical estimates and model-based estimates of the variances and covariances for the vector of shrinkage estimator (i.e., $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\tau}}_{shk}}$) under the same simulation settings as described in Section 4.6.1.

4.9.3 Empirical Distribution of the Shrinkage Estimator and the Approximate Wald Test Statistic

Though the limiting distribution of the shrinkage estimator is technically not normal, the simulation results reveal that this shrinkage estimator is approximately normal and the amount of departure from normality is small. Figure 4.1 shows the quantile-quantile plots of comparing the distribution of shrinkage estimator with the normal distribution (refer to Section 4.6.1 for simulation settings).

Figure 4.2 shows the quantile-quantile plots of comparing the distribution of $\tilde{T}_W$ with a χ^2 distribution, indicating that $\tilde{T}_W$ approximately follows a χ^2 with $df = (I-1)(J-1)$ under H_0. In fact, using the χ^2 null distribution would result in a slightly conservative test.

4.9.4 Efficiency and Bias

Table 4.6 shows the bias and MSE for the interaction estimators $\hat{\boldsymbol{\tau}}$ from three models. We report only the results with $\rho = 0.5$ to save space. The results indicate that the linearization and IRGLS gives numerically consistent and unbiased parameter estimates for Tukey's one-df model. However, when the underlying interaction model is not a Tukey's model (e.g., AMMI1), Tukey's model yields severely biased estimates. In contrast, the saturated model has the least biased estimates. The performance of the proposed shrinkage estimator always lies between Tukey's and the saturated model.

4.9.5 Multivariate Shrinkage versus Scalar Shrinkage

We compared the shrinkage estimates of interaction parameters using only the diagonal elements of $\boldsymbol{B}$ versus using the whole $\boldsymbol{B}$ matrix (Table 4.7). Using scalar shrinkage (or so-called "component-wise shrinkage", the table shows the mean estimated weights $\hat{W}_{tuk}$ corresponding to $\hat{\boldsymbol{\tau}}_{tuk}$, $\hat{W}_{sat}$ corresponding to $\hat{\boldsymbol{\tau}}_{sat}$, and the

resulting mean shrinkage estimator $\hat{\boldsymbol{\tau}}^*_{shk}$. We found that using the scalar version of $\boldsymbol{B}$, $\hat{\boldsymbol{\tau}}^*_{shk}$ is very close to $\hat{\boldsymbol{\tau}}_{shk}$ under Tukey's model since more weights are assigned to $\hat{\boldsymbol{\tau}}_{tuk}$ compared to $\hat{\boldsymbol{\tau}}_{sat}$. In fact, $\hat{\boldsymbol{\tau}}_{tuk} \approx \hat{\boldsymbol{\tau}}_{sat} \approx \hat{\boldsymbol{\tau}}_{shk} \approx \hat{\boldsymbol{\tau}}^*_{shk}$ under Tukey's model. However, this is not the case under AMMI1 or saturated interaction structures. In these cases, $\hat{W}_{tuk}$ dominates over $\hat{W}_{sat}$. As such, $\hat{\boldsymbol{\tau}}^*_{shk}$ is a biased estimate. The results indicate important contributions of the off-diagonal elements (covariances) of $\boldsymbol{B}$ and that multivariate shrinkage is required under certain situations.

Table 4.5: Comparison between empirical and model-based estimates of variance and covariance components for the shrinkage estimator under Tukey's and saturated interaction structures in 3×3 table settings (N=1200 with repeated measures). Data were simulated under an autoregressive-1 correlation structure.

Simulation Model: Tukey's single degree of freedom

σ^2	Estimation	Variance				Covariance					
1	Empirical	0.0017	0.0013	0.0016	0.0014	-0.0006	-0.0001	0.0001	0.0001	-0.0001	-0.0005
1	Model-Based	0.0019	0.0014	0.0017	0.0012	-0.0007	-0.0003	0.0001	0.0001	-0.0003	-0.0006
4	Empirical	0.0070	0.0057	0.0069	0.0061	-0.0024	-0.0007	0.0008	0.0006	-0.0004	-0.0017
4	Model-Based	0.0078	0.0056	0.0069	0.0049	-0.0029	-0.0014	0.0005	0.0005	-0.0011	-0.0026
8	Empirical	0.0140	0.0113	0.0136	0.0111	-0.0044	-0.0017	0.0008	0.0016	-0.0009	-0.0038
8	Model-Based	0.0155	0.0112	0.0137	0.0097	-0.0057	-0.0027	0.0009	0.0009	-0.0021	-0.0051

Simulation Model: Saturated interaction

σ^2	Estimate	Variance				Covariance					
1	Empirical	0.0027	0.0021	0.0026	0.0019	-0.0010	-0.0005	0.0002	0.0002	-0.0004	-0.0010
1	Model-Based	0.0028	0.0021	0.0026	0.0019	-0.0010	-0.0005	0.0002	0.0002	-0.0004	-0.0010
4	Empirical	0.0104	0.0079	0.0087	0.0080	-0.0037	-0.0016	0.0009	0.0016	-0.0016	-0.0038
4	Model-Based	0.0101	0.0080	0.0097	0.0073	-0.0037	-0.0018	0.0006	0.0014	-0.0021	-0.0041
8	Empirical	0.0193	0.0162	0.0193	0.0158	-0.0079	-0.0037	0.0016	0.0034	-0.0035	-0.0073
8	Model-Based	0.0187	0.0151	0.0183	0.0136	-0.0068	-0.0034	0.0012	0.0029	-0.0041	-0.0078

Table 4.6: Mean squared error (MSE) and bias of interaction estimators from Tukey's one-df model, saturated interaction model, and the shrinkage method under different simulation models in 3×3 table settings (N=1200).

Model	True Parm.	$\sigma^2 = 4$						$\sigma^2 = 8$					
		Tukey		Shrinkage		Saturated		Tukey		Shrinkage		Saturated	
		MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
Tukey	0.04	0.0004	-0.0045	0.0113	-0.0047	0.0070	-0.0047	0.0009	-0.0047	0.0226	-0.0103	0.0141	-0.0093
	0.07	0.0012	0.0015	0.0086	0.0010	0.0057	0.0012	0.0023	-0.0002	0.0171	0.0036	0.0113	0.0019
	0.07	0.0013	0.0012	0.0105	-0.0003	0.0069	0.0000	0.0026	0.0016	0.0205	0.0006	0.0136	0.0013
	0.14	0.0033	0.0032	0.0079	0.0042	0.0062	0.0039	0.0062	0.0024	0.0144	0.0040	0.0111	0.0040
AMMI1	0.14	0.0233	-0.1523	0.0113	0.0074	0.0101	-0.0044	0.0226	-0.1493	0.0231	0.0051	0.0195	-0.0130
	0.14	0.0271	-0.1634	0.0088	0.0073	0.0082	-0.0052	0.0261	-0.1579	0.0165	0.0066	0.0145	-0.0130
	-0.14	0.0105	0.0952	0.0098	-0.0018	0.0088	0.0052	0.0120	0.0929	0.0190	-0.0096	0.0156	0.0027
	-0.14	0.0075	0.0549	0.0069	-0.0071	0.0064	-0.0026	0.0125	0.0499	0.0160	-0.0074	0.0143	-0.0010
Saturated	0.00	0.0001	0.0030	0.0120	0.0027	0.0104	0.0026	0.0004	0.0056	0.0242	-0.0011	0.0193	-0.0006
	0.20	0.0390	-0.1958	0.0083	0.0002	0.0081	-0.0139	0.0382	-0.1916	0.0179	0.0048	0.0165	-0.0176
	0.20	0.0394	-0.1967	0.0091	0.0030	0.0088	-0.0115	0.0391	-0.1935	0.0218	0.0079	0.0195	-0.0145
	-0.20	0.0442	0.1983	0.0077	-0.0043	0.0081	0.0103	0.0484	0.1947	0.0157	-0.0066	0.0161	0.0160

Table 4.7: Parameter estimates using Tukey's model, the proposed adaptive shrinkage estimator, and saturated interaction models under Tukey's and saturated interaction structures in 3×3 table settings (N=1200 with repeated measures). Data were simulated under an autoregressive-1 (AR-1) correlation structure (see Section 4.6.1 for simulation details).

Model	True Parm.	$\hat{E}[\hat{\boldsymbol{\tau}}_{tuk}]$	$\hat{E}[\hat{\boldsymbol{\tau}}_{sat}]$	Multivariate Shrinkage					Scalar Shrinkage		
				$\hat{E}[\boldsymbol{B}]$				$\hat{E}[\hat{\boldsymbol{\tau}}_{shk}]$	$\hat{W}_{tuk}$	$\hat{W}_{sat}$	$\hat{E}[\hat{\boldsymbol{\tau}}^*_{shk}]$
Tukey	0.04	0.036	0.035	0.796	0.025	0.025	0.024	0.035	0.796	0.204	0.036
	0.07	0.072	0.071	0.030	0.825	0.033	0.062	0.071	0.825	0.175	0.071
	0.07	0.071	0.070	0.028	0.031	0.822	0.064	0.070	0.822	0.178	0.070
	0.14	0.143	0.144	0.036	0.070	0.079	0.912	0.144	0.912	0.088	0.143
AMMI1	0.14	-0.012	0.147	0.684	-0.314	0.107	0.105	0.136	0.684	0.316	0.059
	0.14	-0.023	0.147	-0.254	0.591	0.126	0.118	0.135	0.591	0.409	0.064
	-0.14	-0.045	-0.142	0.153	0.206	0.881	-0.041	-0.135	0.881	0.119	-0.066
	-0.14	-0.085	-0.147	0.105	0.141	-0.019	0.931	-0.143	0.931	0.069	-0.094
Saturated	0.00	0.003	0.003	0.928	0.005	0.004	0.006	0.003	0.928	0.072	0.012
	0.20	0.004	0.200	-0.150	0.624	-0.168	0.215	0.186	0.624	0.376	0.092
	0.20	0.003	0.203	-0.156	-0.317	0.766	0.219	0.188	0.766	0.234	0.062
	-0.200	-0.002	-0.204	0.174	0.347	0.202	0.767	-0.190	0.767	0.233	-0.053

Table 4.8: Baseline characteristics of 729 study participants in the Normative Aging Study (NAS)

Variable	Mean ± SD, N (percent)
Pulse Pressure (mmHg)	54.26 ± 14.62
Age (years)	66.37 ± 7.12
Body Mass Index (kg/m^2)	27.97 ± 3.77
Race (white)	705 (97%)
Type-2 Diabetes	95 (13%)
Hypertension	397 (54%)
Pack-Years of Cigarette Smoking	
0	226 (31%)
< 30	283 (40%)
≥ 30	206 (29%)
Cumulative Lead Exposure (μg/g): Tibia Bone	
$\leq$15	259 (36%)
(15,25]	263 (36%)
>25	207 (28%)
Cumulative Lead Exposure (μg/g): Patella Bone	
$\leq$20	244 (33%)
(20,32]	223 (31%)
>32	262 (36%)
Number of Repeated Measures on Pulse Pressure Per Subject	
1–2	147 (20%)
3–4	236 (32%)
5–6	286 (39%)
7–8	60 (9%)

Table 4.9: Baseline characteristics of 6361 participants in the Multi-Ethnic Study of Atherosclerosis (MESA)

	Caucasian (N=2526)	Chinese (N=775)	African American (N=1611)	Hispanic (N=1449)
Age (years)	62.66 ± 10.24	62.38 ± 10.38	62.29 ± 10.06	61.38 ± 10.30
Body Mass Index (kg/m^2)	27.74 ± 5.06	23.99 ±3.29	30.15 ± 5.89	29.45±5.14
Gender (female)	1320 (52%)	394 (51%)	868 (54%)	910 (62%)
Education (college and above)	1522 (60%)	392 (51%)	772 (48%)	312 (22%)
Gross Family Income ($\geq$ 50,000)	1397 (55%)	218 (28%)	554 (37%)	104 (7%)
Energy Intake (kcal/day)				
$\leq$ 1000	511 (22%)	360 (47%)	429 (17%)	315 (24%)
(1000, 1300]	486 (20%)	174 (23%)	354 (14%)	250 (19%)
(1300, 1600]	437 (18%)	91 (12%)	383 (15%)	210 (16%)
(1600, 2000]	435 (18%)	76 (10%)	502 (20%)	209 (16%)
>2000	512 (22%)	61 (8%)	851 (34%)	339 (26%)
Total Intentional Exercise (min/wk)				
$\leq$ 0	429 (17%)	197 (25%)	364 (27%)	448 (31%)
(0, 420]	354 (14%)	123 (16%)	252 (18%)	190 (13%)
(420, 840]	383 (15%)	122 (16%)	220 (16%)	230 (16%)
(840, 1680]	502 (20%)	166 (21%)	207 (15%)	225 (16%)
>1680	851 (34%)	331 (24%)	331 (24%)	355 (25%)
Number of Repeated Measures on BMI				
1	104 (4%)	69 (9%)	128 (8%)	141 (10%)
2	107 (4%)	32 (4%)	103 (6%)	90 (6%)
3	154 (6%)	38 (5%)	125 (8%)	35 (2%)
4	2161 (86%)	636 (82%)	1255 (88%)	1183 (82%)

Table 4.10: BMI-associated single-nucleotide polymorphisms (SNPs) with significant meta-analysis p-values for GEI tests in the MESA data (adjusted $\alpha = 0.0019$) for the four race groups (*** denotes $p < 1 \times 10^{-8}$).

		Caucasian			Chinese			African American			Hispanic		
		Tuk	Shk	Sat	Tuk	Shk	Sat	Tuk	Shk	Sat	Tuk	Shk	Sat
Exposure	SNP ID	LRT	Wald	LRT	LRT	Wald	LRT	LRT	Wald	LRT	LRT	Wald	LRT
Energy	rs2815752	0.607	0.993	0.949	0.009	0.032	0.319	0.198	0.500	0.590	0.665	0.997	0.967
Intake	rs543874	0.115	0.441	0.309	NA	NA	NA	NA	NA	0.554	***	***	***
	rs2867125	0.239	0.462	0.281	0.529	0.963	0.841	0.350	0.717	0.495	NA	NA	0.314
	rs987237	0.286	0.370	0.214	0.090	0.399	0.436	0.206	0.455	0.644	0.853	1.000	0.989
	rs3817334	0.108	0.189	0.093	0.028	0.121	0.418	0.017	0.093	0.423	0.629	0.992	0.941
	rs3810291	0.562	0.617	0.360	NA	NA	0.509	NA	NA	0.021	0.997	0.930	0.721
Intentional	rs543874	NA	NA	0.117	0.002	0.014	0.057	NA	NA	0.259	0.005	0.039	0.121
Exercise	rs1514175	NA	NA	0.793	0.238	0.711	0.829	NA	NA	0.800	0.055	0.147	0.104
	r10938397	0.133	0.357	0.255	0.031	0.293	0.585	0.022	0.217	0.185	NA	NA	0.170
	rs7359397	0.747	0.104	0.037	0.017	0.203	0.240	NA	NA	0.536	0.715	0.698	0.424

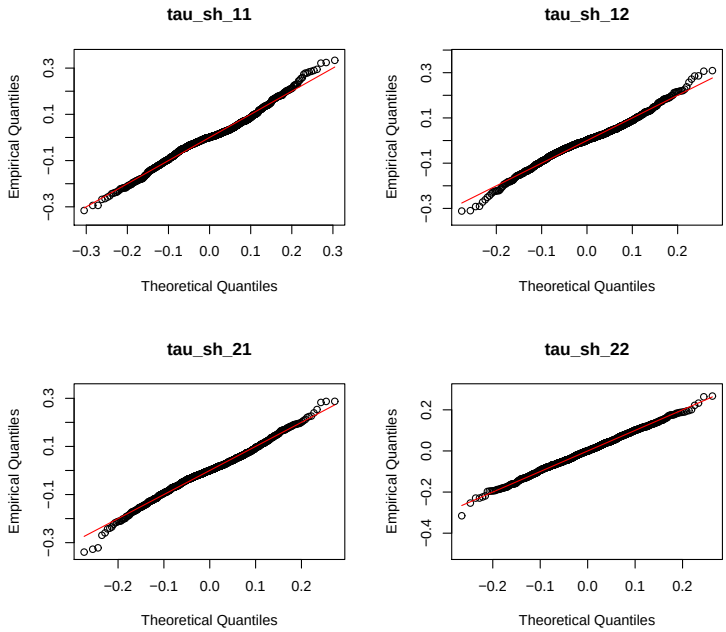


Figure 4.1: Quantile-Quantile (Q-Q) plots for comparing the distribution of the proposed shrinkage estimator with the normal distribution. The shrinkage estimates, $\hat{\boldsymbol{\tau}}_{shk} = (\hat{\tau}_{shk_{11}}, \hat{\tau}_{shk_{21}}, \hat{\tau}_{shk_{12}}, \hat{\tau}_{shk_{22}})^\top$, are obtained from the simulations of GEI in a 3×3 two-way table under H_0 of no interaction (N=1200 with repeated measures).

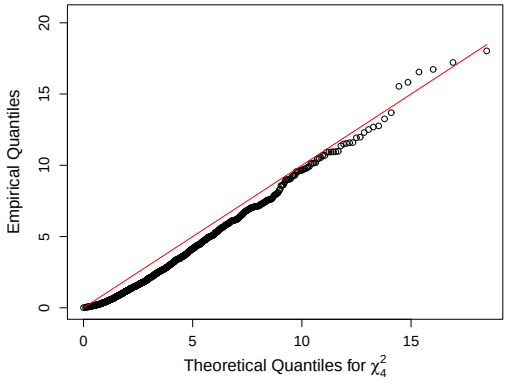


Figure 4.2: Quantile-Quantile (Q-Q) plot for comparing the distribution of the Wald statistics with the chi-squared distribution. The shrinkage estimates are obtained from the simulations of 3×3 GEI two-way table under H_0 of no interaction (N=1200 with repeated measures).

CHAPTER V

Likelihood-Based Test for Interactions in AMMI Models: Application to Gene-Environment Interactions in Multi-Ethnic Study of Atherosclerosis

5.1 Introduction

There has been a great deal of recent interest in identifying and delineating gene-environment interaction (GEI) effects on quantitative traits associated with common complex human diseases in prospective cohort studies (Fan et al., 2012). While most of the statistical literature has focused on GEI methods in case-control studies (Chatterjee et al., 2006; Mukherjee and Chatterjee, 2008; Mukherjee et al., 2008, 2012; VanderWeele et al., 2012), little attention has been given to efficient modeling and testing of interactions in longitudinal studies because the attempt to characterize complex interactions via longitudinal studies poses several statistical challenges. First, prohibitive sample sizes are required when traditional models for interaction analysis are used to detect modest interactions. Furthermore, cohort studies for GEI are typically characterized by substantially unequal sample sizes in G×E×Time configurations as a result of unbalanced allele-frequency, heterogeneous environmental exposure distributions in the population, and loss to follow up that are expected in a longitudinal study. This unbalanced data structure reduces statistical power for testing any kind of time varying pattern in the interaction parameter. In addition, there may be measurement error or nonlinear relations between genotypes, environmental exposures, and phenotypes, and it is challenging to detect any genetic modifying effect when

the model is not correctly specified. Moreover, the effect of GEI may be time or age dependent. This dynamic interaction between genetic and environmental factors over time further increases the complexity of the problem and may contribute to the difficulty of replicating GEI studies.

An interaction model including cross-product terms of gene (G) and environmental exposure (E) under the mixed model framework is typically used for testing gene-gene interaction (GGI) and GEI in longitudinal studies (Moreno-Macias et al., 2010). The interaction term in a product form is appropriate and straightforward for continuous or binary G and E. However, continuous exposure variables are often grouped into quantiles (e.g., tertiles, quartiles or quintiles) in epidemiologic practice to avoid issues with skewed distributions, outliers, and measurement error (Schaffrath Rosario et al., 2006; Siahpush et al., 2007). When G and E are treated as categorical variables, using a product form for GEI results in a saturated interaction structure. A saturated interaction structure consists of estimating a parameter for each configuration of G and E factors without any structural assumption for the interaction term. The number of parameters to be estimated and the degrees of freedom (df) for the interaction test, however, increase substantially with the number of categories of G or E. This may yield inefficient parameter estimates and may result in loss of power compared to a more parsimonious model. In previous chapters, we have proposed to model the interaction structure using parsimonious interaction models borrowed from the classical ANOVA literature, where the interaction structures depend on one or both of the main effects. However, given that the underlying interaction structure is usually unknown, the classical interaction models could have limited power for detecting interactions if the model is misspecified.

The class of additive main effects and multiplicative interaction (AMMI) models (Gauch Jr., 1992), frequently used in multi-location crop cultivar trials, may provide a solution to the problem of modeling interaction. AMMI models were previously

proposed as the "FANOVA" (factor analysis of variance) model by Gollob (1968) and were also studied in Mandel (1971). AMMI models entail a singular value decomposition (SVD) of the cell residual matrix after fitting the additive main effects, so the models do not have structural assumptions on the interaction term. Using AMMI models for GEI involves approximating the interaction term by one or a few multiplicative terms consisted of coefficients from genetic and environmental effects. By choosing a small number of leading multiplicative terms (or interaction factors), one is able to reduce the effective df of the resultant test. AMMI has been shown to perform well across a spectrum of interaction structures (Barhdadi and Dubé, 2010; Mukherjee et al., 2012). In chapter III, AMMI model has been shown to be a useful screening tool for detecting interaction effects specifically in the absence of main effects based on a parametric bootstrap approach.

The parameter estimation for AMMI models involves reduced rank approximation, which has been used to detect marginal genetic (Vounou et al., 2012) and exposure (Nettleton et al., 2007) associations to disease. Typically, GEI cohort studies have considerable heteroscedasticity amog genotype and/or exposure groups. Setting the weights inversely proportional to the variance can lead to a better estimation of the underlying GEI structure (e.g., by minimizing the weighted residual sum of squares). A variety of methods have been proposed for weighted lower rank approximation (LRA), and they primarily differ in the problem representation and the nonlinear optimization approach. Gabriel and Zamir (1979) introduced the "criss-cross regression", also called "alternating regression" in Croux et al. (2003), which is an iterative method to obtain LRA of the interaction matrix with least squares fit. Wentzell et al. (1997) extended the alternating regression and proposed a maximum likelihood principal component analysis algorithm that allows the inclusion of error covariance. However, as the dimension of the interaction matrix becomes large, convergence problems may occur. Similarly, Hwang and Takane (2004) proposed a multivariate reduced-rank

growth curve model with unbalanced data using the alternating maximum likelihood (AML) procedure. They attempted to find a reduced-rank representation for the whole data matrix constructed by individual observations instead of the interaction matrix. The main limitation of the criss-cross or alternating regression is that it may converge to local optima instead of a global optimum. Chen et al. (2008) discussed situations when a dead cycle happens in alternating regression. As such, the convergence criteria should include both convergence in the objective function and the parameter estimates. Srebro and Jaakkola (2003) considered the weighted LRA problem as a maximum-likelihood problem with missing values and implemented an Expectation-Maximization (EM) procedure. However, when the low rank matrix becomes undetectable (e.g., signal-to-noise ratio less than 1), EM often converges to a non-global minimum.

In addition to parameter estimation, testing for the multiplicative terms of AMMI models in non-replicated two-way table settings has been discussed in Mandel (1971). Johnson and Graybill (1972a) derived the maximum likelhood estimators and a likelihood ratio test specifically for the first multiplivative term of the AMMI model by partitioning the sum of squares. For replicated data, Gollob (1968) proposed a F-test for judging the significance of interaction factors through computing the sums of squares and mean squares for the interaction factors and the residual interaction. All of these researches assume that errors are normally distributed with a common variance. Piepho (1995) investigated the robustness of several F-tests (Gollob, 1968; Cornelius, 1980; Cornelius et al., 1992; Cornelius, 1993) to departures from the assumptions of normality and homogeneity of error variances. The most robust test, however, is equivalent to a saturated interaction test.

In this chapter, we propose to develop a likelihood-based test for AMMI models to detect GGI and GEI in longitudinal cohort studies with repeated outcome measures. The two-step regression estimation procedure for AMMI models in Chapter III does

not yield maximum likelihood estimates (MLEs). Here we aim to develop a likelihood-based estimation algorithm for AMMI models by applying LRA techniques and to establish the corresponding likelihood ratio test (LRT) for detecting interaction effects. In Section 5.3, we propose the ML estimation algorithm for AMMI models with the first interaction factor (denoted as AMMI1). In Section 5.4, we describe the LRT statistic for interaction and approximate the corresponding null distribution by a chi-square distribution. The MLEs and the proposed null distribution for approximating the LRT statistic are evaluated in a simulation study. In Section 5.5, we apply AMMI models to the search of interactions between genes related to body mass index (BMI) and several exposure variables (e.g., dietary intake, physical activity, psychosocial factors) using the Multi-Ethnic Study of Atherosclerosis (MESA) data. Specifically, to accommodate multiple exposures in the framework of categorical G and E, we create a health profile with "exposure categories" to summarize information of all exposure variables via various clustering and classification methods. In Section 5.6, we extend the model to allow for time-dependent changes in main and interaction effects over time. The chapter concludes with a discussion in Section 5.7.

5.2 Model

We use (4.1) to model the association between the phenotypic response of interest and genetic factors and environmental exposures. Please refer to Section 4.2 for notations. Here we focus on modeling the mean structure $\boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)$ using AMMI models. Write $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \boldsymbol{d}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\gamma}^\top)^\top$, where $\boldsymbol{\beta}$ are the parameters corresponding to main effects of gene and environment and $\{\boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\}$ represent departure from additivity. Let $\boldsymbol{d} = (d_1, d_2, \dots)^\top$ be a vector of scale parameters, and $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are interaction parameters for row (gene) and column (environment) effects, respectively. We first ignore covariates such that the design matrix $\boldsymbol{X}_k$ is mainly constructed by the indicator functions $I(\cdot)$ for row and column factors, for the t-th observation for the k-th

subject, the mean of an AMMI model (Gollob, 1968; Mandel, 1971) has the form

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = f(\boldsymbol{\beta}, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}_{kt}) = \beta_0 + \sum_{i=1}^{I} \beta_i^G I(G_k = i) + \sum_{j=1}^{J} \beta_j^E I(E_{kt} = j) + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{m=1}^{M} d_m \alpha_{im}\gamma_{jm} I(G_k = i, E_{kt} = j), \quad (5.1)$$

where $d_1 \geq d_2 \geq \cdots \geq d_M$. α_{mi} and γ_{mj} are distinct interaction parameters corresponding to the i-th row and j-th column for the m-th interaction factor, respectively. They are subject to constraints $\sum_{i=1}^{I} \alpha_{im}^2 = \sum_{j=1}^{J} \gamma_{jm}^2 = 1$ and $\sum_{i=1}^{I} \alpha_{im} = \sum_{j=1}^{J} \gamma_{jm} = 0$, as well as orthogonality restrictions $\sum_i \alpha_{im}\alpha_{im'} = \sum_i \gamma_{jm}\gamma_{jm'} = 0$ for $m \neq m'$. The first few terms in the SVD of the GEI matrix are believed to contain the signal of interaction, while the higher-order terms relate to noise due to measurement error. Given that at most three genotype groups on a SNP are considered in most candidate gene studies and an AMMI model with $M = \min(I - 1, J - 1) = 2$ would be equivalent to a fully saturated interaction model, we concentrate on AMMI models with only one multiplicative term

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = f(\boldsymbol{\beta}, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{x}_{kt}) = \beta_0 + \sum_{i=1}^{I} \beta_i^G I(G_k = i) + \sum_{j=1}^{J} \beta_j^E I(E_{kt} = j) + \sum_{i=1}^{I}\sum_{j=1}^{J} d_1 \alpha_i\gamma_j I(G_k = i, E_{kt} = j), \quad (5.2)$$

where we simplify the notations replacing $\alpha_i = \alpha_{i1}$ and $\gamma_j = \gamma_{j1}$. The γ_j corresponds to a hypothetical latent environmental variable that describes the largest amount of the environment interactions. Similarly, α_i describes the axis of genetic susceptibility accounting for the largest amount of genetic interactions. One could incorporate time-varying parameters into the above models to capture time-dependent changes in repeated measurements. Extension to including time-varying coefficients is discussed in Section 5.6.

5.3 Parameter Estimation

The log-likelihood corresponding to $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ is

$$\ell(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{V}) \propto -\frac{1}{2}\sum_{k=1}^{N}\log|\boldsymbol{V}_k(\boldsymbol{\theta})| - \frac{1}{2}\sum_{k=1}^{N}\big[(\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k))^\top \boldsymbol{V}_k(\boldsymbol{\theta})^{-1}(\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k))\big], \tag{5.3}$$

where $\boldsymbol{y}$ is the stacked n-dimensional response vector ($n = \sum_{k=1}^{N} n_k$). If the variance components of $\boldsymbol{y}$ is known, maximizing the likelihood is equivalent to minimizing the objective function with respect to $\boldsymbol{\eta}$

$$\min_{\boldsymbol{\eta}} \boldsymbol{Q}(\boldsymbol{\eta}|\boldsymbol{\theta}) = \sum_{k=1}^{N}\big[\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)\big]^\top \boldsymbol{V}_k(\boldsymbol{\theta})^{-1}\big[\boldsymbol{y}_k - \boldsymbol{f}(\boldsymbol{\eta}, \boldsymbol{X}_k)\big]. \tag{5.4}$$

The solution for $\boldsymbol{\eta}$, $\hat{\boldsymbol{\eta}}_{GLS}$, is the generalized least squares estimator, that is asymptotically normal and efficient under certain regularity conditions (Gumpertz and Pantula, 1992). We may express the multiplicative interaction term in (5.2) in a product form, $d_1\alpha_i\gamma_j = A_iB_j$, for example, $A_i = d_1^{1/2}\alpha_i$, $B_j = d_1^{1/2}\gamma_j$. Let $\boldsymbol{\eta} = (\boldsymbol{\beta}, \boldsymbol{A}^\top, \boldsymbol{B}^\top)^\top$, where $\boldsymbol{A} = (A_1, \ldots, A_I)^\top$ is an $I \times 1$ vector and $\boldsymbol{B} = (B_1, \ldots, B_J)^\top$ is a $J \times 1$ vector. The linearization method as described in Chapter IV, unfortunately, does not work for AMMI models. If one attempts to linearize an AMMI1 model,

$$\begin{aligned}
f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) &\approx f(\boldsymbol{\eta}^*, \boldsymbol{x}_{kt}) + \boldsymbol{D}_k^*(\boldsymbol{\eta} - \boldsymbol{\eta}^*) \\
&= f(\boldsymbol{\eta}^*, \boldsymbol{x}_{kt}) + (\beta_0 - \beta_0^*) + \sum_i\sum_j \Big[(\beta_i^G - \beta_i^{G*})I(G_k = i) + (\beta_j^E - \beta_j^{E*})I(E_{kt} = j) \\
&\qquad + B_j^*(A_i - A_i^*)I(G_k = i, E_{kt} = j) + A_i^*(B_j - B_j^*)I(G_k = i, E_{kt} = j)\Big] \\
&= f(\boldsymbol{\eta}^*, \boldsymbol{x}_{kt}) + (\beta_0 - \beta_0^*) + \sum_i\sum_j \Bigg\{\Big[(\beta_i^G + B_j^*A_i + \beta_j^E + A_i^*B_j) \\
&\qquad - (\beta_i^{G*} + \beta_j^{E*} + 2A_i^*B_j^*)\Big]I(G_k = i, E_{kt} = j)\Bigg\},
\end{aligned}$$

where $\boldsymbol{\eta}^* = (\boldsymbol{\beta}^*, \boldsymbol{A}^{*\top}, \boldsymbol{B}^{*\top})^\top$ is the approximation of $\hat{\boldsymbol{\eta}}$, and $\boldsymbol{D}_k^* = \boldsymbol{D}_k(\boldsymbol{\eta}^*)$ is the first derivative of $f(\boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$ evaluated at a given $\boldsymbol{\eta}^*$. It involves indicators of

row and column as well as current estimates of row and column interaction parameters $\boldsymbol{A}^*$ and $\boldsymbol{B}^*$. Given as such, the main effects and interaction effects cannot be simultaneously identified under the regression setting using linearization.

To find the MLE of $\boldsymbol{\eta}$ for AMMI1 models, we employ the alternating maximum likelihood (AML) estimation in the spirit of Hwang and Takane (2004). The algorithm consists of two global steps: (i) fix variance parameters $\boldsymbol{\theta}$, minimize the objective function and obtain estimates of all coefficients until convergence is reached; (ii) Given $\boldsymbol{\eta}$, estimate variance component parameters. Specifically in (i), AML performs estimation of main effects and interaction effects sequentially. Provided a set of initial values of main effects $\boldsymbol{\beta}^*$, we first use the residuals after removing the main effects, $\boldsymbol{r}_k^* = \boldsymbol{y}_k - \beta_0^* - \beta_i^{G*}\boldsymbol{I}(G_k = i) - \beta_j^{E*}\boldsymbol{I}(E_k = j)$, to form an objective function

$$\boldsymbol{Q}(\boldsymbol{\eta}|\boldsymbol{\theta}) = \sum_{k=1}^{N} \big[\boldsymbol{r}_k^* - \boldsymbol{U}_k \mathrm{Vec}(\boldsymbol{A}\boldsymbol{B}^\top)\big]^\top \boldsymbol{V}_k(\boldsymbol{\theta})^{-1} \big[\boldsymbol{r}_k^* - \boldsymbol{U}_k \mathrm{Vec}(\boldsymbol{A}\boldsymbol{B}^\top)\big] \text{ for fixed } \boldsymbol{V}_k(\boldsymbol{\theta}),$$

where $\boldsymbol{U}_k$ is the indicator function matrix with dimension $n_k \times IJ$ corresponding to the configuration of G_k and E_k for the k-th subject. Then for the whole $n \times 1$ residual vector $\boldsymbol{r}$ and $n \times IJ$ matrix $\boldsymbol{U}$,

$$\begin{aligned}
\boldsymbol{Q}(\boldsymbol{\eta}|\boldsymbol{\theta}) &= \mathrm{tr}\big\{ \big[\boldsymbol{r} - \boldsymbol{U}\mathrm{Vec}(\boldsymbol{A}\boldsymbol{B}^\top)\big] \boldsymbol{V}(\boldsymbol{\theta})^{-1} \big[\boldsymbol{r} - \boldsymbol{U}\mathrm{Vec}(\boldsymbol{A}\boldsymbol{B}^\top)\big]^\top \big\} \\
&= \mathrm{tr}\big\{ \boldsymbol{C}[\boldsymbol{r} - \boldsymbol{U}\mathrm{Vec}(\boldsymbol{A}\boldsymbol{B}^\top)] \big\} \big\{ \boldsymbol{C}\big[\boldsymbol{r} - \boldsymbol{U}\mathrm{Vec}(\boldsymbol{A}\boldsymbol{B}^\top)\big] \big\}^\top \\
&= SS\big[\tilde{\boldsymbol{r}} - \tilde{\boldsymbol{U}}\mathrm{Vec}(\boldsymbol{A}\boldsymbol{B}^\top)\big], \qquad (5.5)
\end{aligned}$$

where $\tilde{\boldsymbol{r}} = \boldsymbol{C}\boldsymbol{r}, \tilde{\boldsymbol{U}} = \boldsymbol{C}\boldsymbol{U}$, $SS(\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{M}\boldsymbol{M}^\top)$, and $\boldsymbol{C}$ is the Cholesky decomposition of the $n \times n$ inverse variance-covariance matrix $\boldsymbol{V}^{-1}$. A closed form of least squares solution for $\boldsymbol{A}$ and $\boldsymbol{B}$ does not exist. We minimize the objective function by alternating least squares. Write (5.5) as

$$\begin{aligned}\boldsymbol{Q} &= SS\big[\tilde{\boldsymbol{r}} - \tilde{\boldsymbol{U}}\mathrm{Vec}(\boldsymbol{A}\boldsymbol{B}^\top)\big] \\ &= SS\big[\tilde{\boldsymbol{r}} - (\boldsymbol{A}\otimes\tilde{\boldsymbol{U}})\boldsymbol{B}\big] \qquad (5.6) \\ &= SS\big[\tilde{\boldsymbol{r}} - (\boldsymbol{B}\otimes\tilde{\boldsymbol{U}})\boldsymbol{A}\big]. \qquad (5.7)\end{aligned}$$

Then the solution is analogous to that for generalized least squares regression. We repeat the following two local steps: (a) update $\boldsymbol{B}$ for fixed $\boldsymbol{A}$: using (5.6), the least squares estimate of $\boldsymbol{B}$ is

$$\hat{\boldsymbol{B}} = [(\boldsymbol{A}\otimes\tilde{\boldsymbol{U}})^\top(\boldsymbol{A}\otimes\tilde{\boldsymbol{U}})]^{-1}(\boldsymbol{A}\otimes\tilde{\boldsymbol{U}})^\top\tilde{\boldsymbol{r}},$$

and (b) update $\boldsymbol{A}$ for fixed $\boldsymbol{B}$: using (5.7), the least squares estimate of $\boldsymbol{A}$ is

$$\hat{\boldsymbol{A}} = [(\boldsymbol{B}\otimes\tilde{\boldsymbol{U}})^\top(\boldsymbol{B}\otimes\tilde{\boldsymbol{U}})]^{-1}(\boldsymbol{B}\otimes\tilde{\boldsymbol{U}})^\top\tilde{\boldsymbol{r}},$$

till convergence is reached.

We adapt the estimation algorithm for $\boldsymbol{V}(\boldsymbol{\theta})$ in linear mixed models to this nonlinear setting of AMMI1 model. First, we take the derivative of l with respect to each θ_s of $\boldsymbol{\theta}$,

$$\frac{\partial\ell}{\partial\theta_s} = \frac{-1}{2}\left[\mathrm{tr}(\boldsymbol{V}(\boldsymbol{\theta})^{-1}\frac{\partial\boldsymbol{V}(\boldsymbol{\theta})}{\partial\theta_s}) - (\boldsymbol{Y}-\boldsymbol{f}(\boldsymbol{\eta}))^\top\boldsymbol{V}(\boldsymbol{\theta})^{-1}\frac{\partial\boldsymbol{V}(\boldsymbol{\theta})}{\partial\theta_s}\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{Y}-\boldsymbol{f}(\boldsymbol{\eta}))\right]. \qquad (5.8)$$

The ML equations for $\boldsymbol{V}$ are obtained by equating $\partial\ell/\partial\theta_s$ to 0. The equations for $\boldsymbol{V}(\boldsymbol{\theta})$ in (5.8) and the AML algorithm for $\boldsymbol{\eta}$ need to be solved simultaneously. Note that the regressors $(\alpha_i\gamma_j)$ for the fixed interaction effects are unobservable. To resolve the issue of nonlinearity in $d_1\alpha_i\gamma_j$ in ML equations, we replace one column for the fixed interaction term to account for the loss of df for estimating d_1, that is, in each iteration $\hat{\alpha}_i\hat{\gamma}_j$ is treated as the regressors for individuals in the i-th row and j-th column, as the alternating regression (Croux et al., 2003).

Initial values of main effect estimates and variance components can be obtained by a fully saturated regression model. As studied in Chen et al. (2008), two conditions need to be satisfied to achieve the final convergence: (1) the first condition requires convergence in both the row and the column interaction parameters $(\boldsymbol{A}, \boldsymbol{B})$; (2) the second condition requires convergence in the objective function. The procedure of parameter estimation for AMMI models is summarized in the following.

1. Given a set of initial values of main effects, obtain the residuals by removing the main effects to form an objective function.
2. Given an initial value of $\boldsymbol{B}$, update $\hat{\boldsymbol{A}}$ by minimizing the objective function.
3. Given the updated $\hat{\boldsymbol{A}}$, update $\hat{\boldsymbol{B}}$.
4. Given $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$, update the variance components and main effect estimates.
5. Repeat steps 1–4 until convergence is achieved.
6. Find $\hat{d}_1, \hat{\alpha}_i, \hat{\gamma}_j$ for $i = 1, \ldots, I, j = 1, \ldots, J$ by rescaling $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ obtained in 5 according to the orthonormal constraints.

Based on our simulation settings and data analysis examples, we found that this algorithm converges within about 50 iterations.

We performed a simulation study to evaluate the bias and mean squared error (MSE) of MLEs of AMMI1 models using the proposed AML estimation algorithm. Data were generated under AMMI1 models. Each dataset consisted of N=2000 subjects with repeated measures (see Section 4.6 for the simulation of repeated measurements). We used compound symmetric correlation structure for within-subject observations with error variance $\sigma^2 = 2, 4$ and within-subject correlation $\rho = 0.2, 0.5, 0.8$. Table 5.2 and Table 5.3 list the bias and MSE of the MLEs for AMMI1 models from 1000 simulations in a 3×3 and a 3×5 table, respectively. Based on the simulation results, the main effect parameter estimates appeared to be unbiased estimates. Even though the estimate for d_1 had positive bias and much larger MSE compared to main effect estimates (which may be due to the rescaling procedure), the estimated product

terms $(\hat{d}_1\hat{\alpha}_i\hat{\gamma}_j)$ appeared to be unbiased estimates.

5.4 Test for Interaction Effects with AMMI1 Models

The null hypothesis of no interaction is given by $H_0 : d_1 = 0$, and the alternative hypothesis is $H_a : d_1 \neq 0$. After obtaining the MLEs under H_0 and H_a, we have

$$\begin{aligned}
\ell(\hat{\boldsymbol{\eta}}_{ML,0}, \hat{\boldsymbol{\theta}}_0) = & -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_k \log|\hat{\boldsymbol{V}}_k(\hat{\boldsymbol{\theta}}_0)| \\
& -\frac{1}{2}\sum_k \Big\{[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}_0, \boldsymbol{X}_k)]^\top \hat{\boldsymbol{V}}_k(\hat{\boldsymbol{\theta}}_0)^{-1}[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}_0, \boldsymbol{X}_k)]\Big\} \\
\text{and } \ell(\hat{\boldsymbol{\eta}}_{ML}, \hat{\boldsymbol{\theta}}) = & -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_k \log|\hat{\boldsymbol{V}}_k(\hat{\boldsymbol{\theta}})| \\
& -\frac{1}{2}\sum_k \Big\{[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}, \boldsymbol{X}_k)]^\top \hat{\boldsymbol{V}}_k(\hat{\boldsymbol{\theta}})^{-1}[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}, \boldsymbol{X}_k)]\Big\},
\end{aligned}$$

respectively. The LRT statistic is given by

$$\begin{aligned}
\text{LRT} = & -2[\ell(\hat{\boldsymbol{\eta}}_{ML,0}, \hat{\boldsymbol{\theta}}_0) - \ell(\hat{\boldsymbol{\eta}}_{ML}, \hat{\boldsymbol{\theta}})] \\
= & \sum_k \log|\boldsymbol{V}_k(\hat{\boldsymbol{\theta}}_0)| - \sum_k \log|\boldsymbol{V}_k(\hat{\boldsymbol{\theta}})| + \\
& \sum_k \big\{[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}_0, \boldsymbol{X}_k)]^\top \boldsymbol{V}_k(\hat{\boldsymbol{\theta}}_0)^{-1}[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}_0, \boldsymbol{X}_k)]\big\} - \\
& \sum_k \big\{[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}, \boldsymbol{X}_k)]^\top \boldsymbol{V}_k(\hat{\boldsymbol{\theta}})^{-1}[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}, \boldsymbol{X}_k)]\big\} \\
= & \sum_k \log\frac{|\boldsymbol{V}_k(\hat{\boldsymbol{\theta}}_0)|}{|\boldsymbol{V}_k(\hat{\boldsymbol{\theta}}|} + RSS_0^* - RSS^*, \qquad (5.9) \\
\text{where } RSS_0^* = & \sum_k \big[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}_0, \boldsymbol{X}_k)]^\top \boldsymbol{V}_k(\hat{\boldsymbol{\theta}}_0)^{-1}[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}_0, \boldsymbol{X}_k)\big] \\
\text{and } RSS^* = & \sum_k \big[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}, \boldsymbol{X}_k)]^\top \boldsymbol{V}_k(\hat{\boldsymbol{\theta}})^{-1}[\boldsymbol{y}_k - \boldsymbol{f}(\hat{\boldsymbol{\eta}}, \boldsymbol{X}_k)\big].
\end{aligned}$$

The null distribution of the LRT statistic is not distributed in accordance with the standard chi-square distribution. One could apply the parametric bootstrap strategy discussed in Chapter III to derive its null distribution. However, the whole process may be quite computationally burdensome. In fact, we found that the LRT statistic

in (5.9) under H_0 is well approximated by a χ^2 with a fractional df ν. The density function of the χ^2 distribution with ν df is given by

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \;\; 0 < x < \infty,$$

where $\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt$ is the gamma function. Here we provide heuristic justification for the null distribution of the LRT statistic. The heuristic argument is supported by Monte-Carlo simulations. Through our empirical studies, we noted that similar results may hold for both cross-sectional and repeated measures cases.

In order to deduce the fractional df ν, we borrow the idea from the construction of a hypothetical ANOVA table in balanced designs. In the case of single observation per cell in a $I \times J$ table (Mandel, 1971), if the AMMI model in (5.1) can be represented by an ANOVA table containing sums of squares, numbers of df, and mean squares, the sum of squares corresponding to the m-th multiplicative interaction term can be expressed as

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{d}_m\hat{\alpha}_{mi}\hat{\gamma}_{mj})^2 = \hat{d}_m^2.$$

Under normality assumptions, $\hat{d}_m^2, m = 1, \ldots, M$, are distributed as the characteristic roots of a Wishart matrix. Under H_0 when interaction is absent, the mean square corresponding to each multiplicative interaction term in an AMMI model should be an estimate of random measurement error. Let σ^2 denote the variance for the random error, SS_{int} and df_{int} be the sum of squares and df due to interaction, respectively. For the interaction of an AMMI1 form, we have

$$E_{H_0}\left(\frac{SS_{int}}{df_{int}}\right) = E_{H_0}\left(\frac{\hat{d}_1^2}{df_{int}}\right) = E(MS_{error}) = \sigma^2 \;\Rightarrow\; df_{int} = \frac{E_{H_0}(\hat{d}_1^2)}{\sigma^2}.$$

We propose to estimate df_{int} using the empirical mean of $\hat{d}_1^2/\sigma^2$ by generating balanced data in a two-way table under H_0. Then we consider the estimated df_{int} as the estimate of the fractional df ν in our case.

Mandel (1971) evaluated the mean and variance of d_1^2/σ^2 by Monte Carlo tech-

niques for various values of I and J and provided Monte Carlo values of $E_{H_0}(\hat{d}_m^2/\sigma^2)$ for $m = 1, 2, 3$ under $H_0 : d_m = 0$. To use the results of Mandel to derive an estimate for ν for arbitrary numbers of rows and columns, we proceed as follows: (1) A large number of $I \times J$ matrices $\boldsymbol{Z} = ((z_{ij}))$ with $z_{ij} \sim \mathcal{N}(0, \sigma^2)$ are generated. (2) Constructing a basis set of interaction functions to remove row/column main effects, we estimate d_1 using the largest singular value by applying SVD to the residual (or interaction) matrix. (3) Repeat (1) and (2) for 10,000 simulations and compute the mean of $\hat{d}_1^2$, then $\hat{\nu} = \hat{E}(\hat{d}_1^2)/\sigma^2$. Finally, the null distribution of the LRT statistic in (5.9) is approximated by $\chi^2_{\hat{\nu}}$.

To assess the validity of the approximation of LRT by a $\chi^2_{\hat{\nu}}$, we examined the empirical null distribution of the LRT statistic in unbalanced two-way table settings with various variances and within-subject correlations. The empirical null distribution of LRT is compared to a $\chi^2_{\hat{\nu}}$ using both the first three moments and quantile-quantile plots. Table 5.1 lists the first three moments of the LRT under H_0 and the corresponding $\chi^2_{\hat{\nu}}$ under 3×3 and 3×5 table settings for N=2000 with repeated measures. The repeated measures are either uncorrelated (i.e., cross-sectional data, within-subject correlation $\rho = 0$) or correlated (within-subject correlation $\rho = 0.2, 0.7$). Under the same simulation settings, Figure 5.7 shows the quantile-quantile plots of the LRT statistics for AMMI1 model against $\chi^2_{\hat{\nu}}$. The results indicated that the null distribution of the LRT statistic is only dependent on the table dimension and remains unchanged with respect to different variances and within-subject correlations.

5.5 MESA Data Analysis

Obesity is an important risk factor for many disorders, such as type 2 diabetes and cardiovascular disease (Hubert et al., 1983; Mokdad et al., 2003). Lifestyle patterns associated with dietary pattern, physical activity, and mental health have been known to contribute to increasing prevalence of obesity and overweight (Wadden et al., 2012;

Onyike et al., 2003). In addition, heritability studies have suggested a considerable genetic contribution to obesity risk (Stunkard et al., 1986; Maes et al., 1997). Body mass index (BMI) is a convenient, inexpensive measure of obesity. The identification of genetic modifying effects on the association between lifestyle patterns and BMI may lead to a better strategy of lifestyle intervention to reduce the risk of obesity.

The analysis dataset comes from the Multi-Ethnic Study of Atherosclerosis (MESA). The description of this longitudinal cohort study was described in Section 4.7.2. The outcome variable of interest was BMI, which was calculated as weight (kg)/height $(\text{m})^2$. The analysis included 6429 MESA participants who had both BMI data in at least one of MESA exams 1–4 and genotype data for the selected BMI-related SNPs. Table 5.4 provides baseline demographic information for the MESA study population in each self-reported ethnic group as well as in the combined sample. The primary analysis goal was to investigate GEI effects on the relationship between several exposures, including behavioral and psychosocial factors, and BMI. We analyzed the interactions separately for each individual exposure and for the exposure groups/categories generated by some clustering methods that summarize multiple exposures. We also used data from the MESA Neighborhood Study, an ancillary study to MESA, to conduct E×E interaction analysis. As such, the secondary analysis goal was to assess the effects of neighborhood environments on individual's dietary intake and physical activity and how neighborhood environments modify the associations between dietary intake and physical activity and BMI. Below we describe the genetic variables in Section 5.5.1, environmental exposure variables in Section 5.5.2, and neighborhood environment measures in Section 5.5.6, respectively. We applied the proposed AMMI1 model (5.2) to model these interactions and compared the results with a saturated interaction model.

5.5.1 Genes

Twenty-seven SNPs previously associated with BMI were selected for genotyping in all racial/ethnic groups based on prior GWAS findings (Speliotes et al., 2010). Further details of genotyping in MESA has been described previously (Bielinski et al., 2008). We adopted the best-guess approach that essentially uses the genotype with the highest imputed genotype probability. For initial exploratory analysis, we considered an additive model (i.e., allele counts). When demonstrating the use of AMMI models under the framework of categorical G and categorical E, we treated it as a nominal variable. That is, three genotype groups (wildtype, heterozygous, or homozygous) were considered for each SNP. Table 5.6 displays information on the 27 BMI SNPs that have been shown to reach genome-wide significance ($p<5\times10^{-8}$) levels in prior meta-analysis. In addition, the genetic risk score (GRS), calculated by summing BMI-increasing allele counts, was created as a summarized variable corresponding to the 27 SNPs. The GRS was also categorized into five categories using quintiles in order to illustrate the use of AMMI1 models and the comparison with saturated interaction models.

5.5.2 Environmental Exposures

The 11 exposure variables we considered for analysis are listed in Table 5.5. They involved variables in the following three domains: (1) dietary intake, (2) physical activity, and (3) psychosocial/mental health. The diet variables included total energy intake (kcal/day), percent calories from carbohydrate intake, percent calories from protein intake, percent calories from saturated fat intake, and percent calories from trans fat intake. The physical activity variables included total intentional exercise (MET-minute/week) and moderate and vigorous physical activity (MET-minute/week). The psychosocial (or mental health) variables included trait anxiety, trait anger, chronic burden, and depressive symptoms. In our analysis, total energy

intake, intentional exercise, physical activities, and the four psychosocial variables were log-transformed to approximate normality. See Appendix for details of each exposure measure. Given that not every variable was measured at all four exams by the study design, we replaced missing values by the last observed values (i.e., last observation carried forward). We examined Pearson's correlations among the exposure measures and BMI at baseline (Figures 5.8 and 5.9 in Appendix). As expected, higher energy intake, higher consumptions of carbohydrates and fats, more chronic burdens, and increased depressive symptoms were significantly positively correlated with higher values of BMI, while more intentional exercise and physical activities were associated with lower values of BMI.

5.5.3 Methods to Define Exposure Groups

Given that multiple exposures were considered for analysis, any existing GEI signal may be washed out by the adjustment of multiple testing if GEI test were repeated for each exposure variable. We combined information from the 11 exposure variables by utilizing three clustering and classification techniques (described below) to discover natural grouping patterns of overall health profile in the data. The overall health profile represented by the exposure groups/categories were then used for the analysis of GEI as a way of reducing the number of tests. For all clustering analyses, subject-level means corresponding to the repeated measurements of the exposure variables were used.

K-Means Clustering

K-means cluster analysis, one of the most popular clustering methods, is a dynamic nonparametric partitioning method and is suitable for quantitative-type variables (Jain, 2010). The k-means algorithm (1) randomly selects k centroids (k less than the number of data points) and assigns each data point to its closet centroid by minimizing the within-cluster sum of squares and (2) recalculates the centroids as

the average of all data points in a cluster and again assigns data points to their closest centroids. Step (2) is continued until the observations are not reassigned or the maximum number of iterations is reached. The number of clusters was chosen based on plotting the number of clusters and the corresponding total within-cluster sum of squares, and we chose six clusters. An efficient algorithm by Hartigan and Wong (1979) that minimizes the sum of squares of the observations to their assigned cluster centers was used. The means for six clusters determined by k-means are shown in Figure 5.2.

Latent Class Analysis (LCA)

Latent class analysis (LCA) is used to detect the presence of latent classes and to cluster a set of observed variables into groups based on their maximum likelihood class membership (Lazarsfeld and Henry, 1968). We used the **mclust** package in R that performs LCA on continuous data (Fraley, Raftery, Murphy, and Scrucca, Fraley et al.). The model parameters were estimated using maximum likelihood via the EM algorithm, and the best normal mixture model was chosen according to the maximum Bayesian information criterion (BIC) values among different covariance structures and different numbers of clusters. The various covariance restrictions result in a different combination of cluster shapes in each model. The constraints yield parsimonious models which facilitate a more flexible modeling strategy beyond assuming unequal covariance or equal covariance. The best model was reached with an eight-cluster solution, but six clusters appeared to be sufficient (Figure 5.10 in Appendix). The cluster means determined by **mclust** are shown in Figure 5.3.

Classification and Regression Trees (CART)

Classification and regression tree (CART) analysis is a machine-learning method that recursively partitions data into smaller groups that involves a categorical (for classification trees) or continuous (for regression trees) dependent variable and one or more independent variables (Breiman et al., 1984). At each split, data are partitioned

into two mutually exclusive groups based on a single independent variable. Then the splitting procedure is applied to each group separately. To have a reasonable number of splits, CART generates a sequence of sub-trees by growing a large tree and pruning it back. Specifically, it sequentially collapses nodes that result in the smallest change in purity. Then it uses cross-validation to select the optimal tree (i.e., the one with the lowest cross-validation misclassification rate). We implemented CART analysis via the **rpart** package in R to generate groups of overall health based on these behavioral and psychosocial factors. To prune the tree, we selected the complexity parameter associated with the smallest cross-validated error. The minimum number of observations in a node was set to be 500 before attempting a split and that a split must decrease the overall lack of fit by a factor of 0.004 (cost complexity factor) before being attempted. Percent calories from trans fats, chronic burden, and intentional exercise were shown to be significant predictors for classification. The resultant model separated the MESA participants into five groups (Figure 5.4).

5.5.4 Main Effects

We first investigated the genetic main effects and the exposure main effects on BMI using fixed effects models with unstructured correlation structure (which was chosen based on smallest AIC values) for within-subject correlation due to repeated measures data. Covariates considered included age at the time of data collection (centered at 65), age squared, gender, race/ethnicity, education, household income, and diagnosis of cancer. Race/ethnicity was classified as Caucasian, Chinese, African American, and Hispanic. Participants selected their highest education level from eight categories that were later collapsed into two categories: whether having a college degree or not. Participants selected their annual household income from 13 categories with different ranges of income ($0–$9,999, $10,000–$19,999, etc.) at MESA exams 1, 2, and 3. Continuous income in US dollars was assigned as the interval midpoint

of the selected category. Furthermore, we considered the adjustment for the first three principal components (PCs) to adequately remove confounding effect due to population stratification in analysis. The first three PCs together explained about 96% of the total observed variation according to the principal component analysis results of the MESA-SHARe data (see Appendix). Except for age and income, all other covariates were collected once at the baseline visit. For the analysis of genetic main effects, covariates included were age, age squared, gender, and the first three PCs following Speliotes et al. (2010). For the analysis of exposure main effects, covariates included age, age squared, gender, race, education, income, and diagnosis of cancer.

The last column of Table 5.6 shows the test results of genetic main effects on BMI (using additive models) in the MESA data. Out of 27 SNPs, only SNP #5 (rs2867125 near the *TMEM18* gene) and SNP #23 (rs7359397 near the *SH2B1, APOB48R*, and *SULT1A2* genes) were significantly associated with BMI considering the adjustment for multiple testing (adjusted p-value = 0.05/27 = 0.0019). Table 5.7 shows the estimated main effects of each environmental exposure variables on BMI. Except for trait anger and trait anxiety, all other variables were significantly associated with BMI. In particular, percent calories from trans fat intake appeared to have a profound impact on BMI. Mean BMI was estimated to increase by 1.74 (95% CI = [1.34, 2.14]) kg/m^2 for one percent increase in trans fat intake, adjusting for age, age squared, gender, race, education, income, and diagnosis of cancer.

In general, the overall health profile (exposure groups defined by k-means, LCA, and CART) was significantly associated with BMI (all p<0.0001). Table 5.8 shows the estimates of the cluster main effects (using k-means, LCA, and CART with subject-level means) on BMI, adjusted for age, age squared, gender, race, education, income, diagnosis of cancer, and the first three PCs.

5.5.5 Interaction Effects

SNP/GRS × E: One-at-a-time Analysis

As a data exploration and screening for possibly important SNPs, we investigated all pairwise interactions between the 27 BMI SNPs and the 11 exposure variables by treating both SNP data and exposure variables as continuous variables. Again, a linear fixed effects model was utilized with unstructured correlation structure. Covariates considered were age, age squared, gender, race, education, income, diagnosis of cancer, and the first three PCs. Table 5.9 shows the p-values for the $27 \times 11 = 297$ tests of GEIs. The test results of interactions between GRS and the exposure variables are displayed in the last row of Table 5.9. (We also considered generating GRS weighted by the published standardized effect sizes of BMI-increasing alleles. Since the results are similar to the unweighted GRS, the results for weighted GRS are not shown.) To reduce the number of subsequent tests, SNPs with a p-value corresponding to the GEI test less than 0.10 in Table 5.9 were preserved for further SNP×E interaction analysis using the aforementioned clustering methods for the 11 exposures. According to Table 5.9, SNPs #2, #6, #10, #16, #19, #22, and #23 were chosen for the following analysis. Due to a very small minor allele frequency, SNP #10 was excluded because one of the cells in the GEI two-way table had observations from only one participant.

SNP/GRS × Exposure Groups

To illustrate the use of AMMI1 model and saturated interaction model for the GEI structures, both G and E were treated as categorical variables. For G, we considered three genotype groups for each SNP. For E, we considered the overall health profile defined by exposure groups (obtained by k-means, LCA, and CART). Table 5.10 shows the test results of interactions between the six BMI SNPs (and GRS) and overall health profile (by using k-means, LCA and CART methods for grouping) using AMMI1 and saturated interaction models. When using k-means to generate

the overall health profile, significant modifying effects of SNP rs1558902 near the *FTO* gene and SNP rs7359397 near the *SH2B1, APOB48R*, and *SULT1A2* gene regions on the association between the health status and BMI were found using the AMMI1 model (p=0.029 and p=0.047, respectively). Significant modifying effects of SNP rs3817334 near *MTCH2, NDUFS3*, and *CUGBP1* and also SNP rs7359397 near the *SH2B1, APOB48R*, and *SULT1A2* gene regions were found using the saturated interaction model (p=0.020 and p=0.043, respectively). The GRS was found to have a significant modifying effect on the association between k-means clustering and BMI using both AMMI1 (p=0.022) and saturated interaction models (p=0.003). When using CART, SNPs rs543874 near the *SEC16B* gene and again SNP rs7359397 were found to have significant interaction effects using both AMMI model (p=0.018 and p=0.007, respectively) and saturated interaction model (p=0.006 and p=0.005, respectively). The GRS was found to have a significant modifying effect using the saturated interaction model (p=0.014) but not the AMMI1 model, indicating that the interactions between GRS and overall health profile (either grouped by k-means or CART) may require more than one interaction factors (e.g., AMMI2) to characterize the effects. Nevertheless, when taking multiple tests into account, these findings may not be regarded as statistically significant.

Remark: This is an ad hoc interaction analysis to illustrate the use of AMMI models with the exposure groups in a categorical framework. The uncertainty in clustering was not taken into account.

5.5.6 Neighborhood Environments

Neighborhood environments could have a profound impact on BMI. The availability of high-quality fruits and vegetables and low-fat foods may contribute to a healthy diet. Density of facilities for physical activity in a neighborhood, such as parks and recreational centers, may increase the likelihood that residents will be physically ac-

tive. A healthy diet with a good amount of physical activities may contribute to a normal range of BMI. In fact, the associations between neighborhood environments and BMI or obesity have been reported in several studies using the MESA data (Mujahid et al., 2008; Moore et al., 2013; Hirsch et al., 2014). Given as such, our secondary analysis goal was to investigate the interaction between neighborhood environments and overall health profile, that is, whether the effect of overall health profile on BMI was dependent on resources for physical activity and availability of healthy foods.

The neighborhood environment measures included in the MESA data were: 1-mile simple densities of supermarkets and fruits and vegetable markets, 1-mile recreational density, perceived healthy food availability, and perceived walkability. Simple densities of supermarkets and fruits and vegetable markets and recreational resources surrounding participant households were calculated using ArcGIS 9.3 for each follow-up year. Address information (time-dependent) was used to link study participants to density measures for the corresponding calendar year. Details of the questionnaires designed for perceived healthy food availability and walkability are described in Section 5.7. All neighborhood environment measures were collected at each MESA exam, and higher values indicated better neighborhood environments. We considered two summary measures for analysis, combined healthy food environment and combined physical activity environment, which were obtained by summing the standardized density and perceived availability variables. For the ease of interpretation, the two summary neighborhood variables were centered according to the median as well as rescaled such that a one unit increase equals the difference between the 10th and the 90th percentile of their baseline distribution in the sample.

Neighborhood Environments × Individual's Diet and Physical Activity

Given the significant associations between both dietary intake and physical activity and BMI (Table 5.7) and the significant marginal effects of physical neighborhood environments on BMI (Mujahid et al., 2008; Hirsch et al., 2014), we investigated the

interactions between the combined healthy food environment index and individual's diet and between the combined physical activity environment index and individual's physical activities on BMI. As before, we used an unstructured correlation structure to account for within-subject correlation of repeated measured BMI. All models were adjusted for age, gender, race, education, household income, and diagnosis of cancer.

The result indicated a significant interaction between the combined healthy food environment and saturated fat consumption (p=0.02). For people living within the neighborhood with a median healthy food environment index of the MESA population, one percent increase in saturated fat intake was associated with 0.19 (95% CI = [0.14, 0.23]) kg/m^2 increase in BMI. But this association was estimated to decrease by 0.03 (95% CI = [0.005, 0.05]) kg/m^2 for an increase from the 50th to the 90th percentile in the combined healthy food environment index. No other significant interactions were detected.

Neighborhood Characteristic Groups × Exposure Groups

To demonstrate the analysis under the categorical framework, we categorized the participants into four neighborhood characteristic groups according to the signs of their combined healthy food and physical activity neighborhood environment indices. The first group had both positive combined healthy food environment and combined physical activity environment indices, representing better healthy food and physical activity neighborhoods. The second group had a positive healthy food environment index but a non-positive physical activity environment index. The third group had a positive physical activity environment index but a non-positive healthy food environment index. The last group had both non-positive combined healthy food environment and combined physical activity environment indices, indicating worse healthy food and physical activity neighborhood environments.

We investigated the interaction effects between the neighborhood characteristic groups and overall health profile groups on BMI. A multilevel mixed-effects model

was utilized with measurement occasions nested within persons and persons nested within similar or different neighborhoods (according to their census tracts). The AMMI1 model is given by

$$Y_{nkt} = \beta_0 + \beta_i^N I(\text{NBHD}_{nk} = i) + \beta_j^H I(\text{Health}_k = j) + d_1 \alpha_i \gamma_j I(\text{NBHD}_{nk} = i, \text{Health}_k = j)$$
$$+ \boldsymbol{\beta}^{C\top} \boldsymbol{Z}_{nkt}^C + a_n + b_k + e_{nkt}, \;\; a_n \sim N(0, \sigma_a^2), b_k \sim N(0, \sigma_b^2), \;\; e_{nkt} \sim N(0, \sigma_e^2),$$

where Y_{nkt} is the BMI value measured at time t for individual k in neighborhood n, NHBH_{nk} is the neighborhood characteristics ($i = 1, \ldots, 4$) for individual k in neighborhood n, Health_k is the overall health profile (using k-means, LCA, or CART) for individual k, $\boldsymbol{Z}_{nkt}^C$ is a 9×1 vector for covariates including age, age^2, gender, race, education, income, and cancer diagnosis. a_n is a random neighborhood component, b_k is a random subject-level intercept, and e_{nkt} is the random measurement error.

The results indicated a LRT statistic of 18.24 (p=0.023) when using an AMMI1 model and a LRT statistic of 28.87 (p=0.004) when using a saturated interaction model for the interaction between exposure groups defined by CART (Figure 5.4) and neighborhood characteristic groups. The two-way interaction forms a 5×4 table, resulting in df=8.35 for the interaction test using an AMMI1 model and df=12 using a saturated interaction model. On the other hand, the interaction wa s not significant for exposure groups defined by k-means (p=0.46 and p=0.19 for AMMI1 and saturated interaction structures, respectively) or LCA (p=0.50 and p=0.67 for AMMI1 and saturated interaction structures, respectively). Figure 5.5 shows the estimated effects of exposure groups defined by CART on BMI for each healthy food and physical activity neighborhood environment group. Both healthy food and physical activity environments appeared to have a significant interaction effect on the association between exposure groups and BMI (or vise versa). Specifically, less intake of trans fats, more intentional exercise, and fewer chronic burdens did not seem to affect BMI under worse healthy food and physical activity neighborhood environments (both combined

indices were negative), as compared to better neighborhood environments (one of the combined indices were not negative).

5.6 Time-Varying Interaction

In addition to testing the presence of GEI, one may want to know whether the GEI is time or age dependent and if so, how to characterize the temporal trend of GEI. In this section, we explored the possibility of allowing time-varying GEI under the AMMI1 model and considered to further analyze the interaction between SNP #23 (rs7359397) and exposure groups defined by CART (Table 5.10). Specifically, we examined how the GEI matrix of an AMMI1 form may change across age.

To evaluate the possible change in GEI across age, we first define five age intervals (40–49, 50–59, 60–69, 70–79, 80–89) for the MESA data and categorized observations into the five age groups according to participant's age at the time of measurement. Subsequently, we fit an AMMI1 model using the proposed AML estimation algorithm in Section 5.3 but allowing both main effects and interaction effects to change over time by incorporating indicator functions for age groups in the model. Let A_{kt} denote age group for the k-th subject at the t-th observation, $A_{kt} = l = 1, \ldots, L$, where we defined $L = 5$ age groups. Following the notations in (5.2), the mean AMMI1 model (dropping the covariates) is given by

$$\begin{aligned} f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = & \sum_l \beta_l^A I(A_{kt} = l) + \sum_i \sum_l \beta_{il}^G I(G_k = i, A_{kt} = l) + \\ & \sum_j \sum_l \beta_{jl}^E I(E_{kt} = j, A_{kt} = l) + \\ & \sum_i \sum_j \sum_l d_{1l} \alpha_{il} \gamma_{jl} I(G_k = i, E_k = j, A_{kt} = l), \end{aligned} \tag{5.10}$$

where β_l^A is the intercept, β_{il}^G is the genetic main effect of the i-th genotype, and β_{jl}^E is the exposure main effect effect of the j-th exposure category, specifically for the l-th age group. The interaction parameters $(d_{1l}, \alpha_{il}, \gamma_{jl})$ are also specific to the

l-th age group. Model (5.10) ultimately yields five sets of age-specific main-effect and interaction-effect parameters as a three-way G×E×Age interaction model. The model used an unstructured within-subject correlation structure and adjusted for the same set of covariates as described previously. To prevent building an over-specified model and to remain consistent with the analyses in other MESA studies, any possible time-varying effects of covariates on BMI were not considered here. To compare the likelihoods of models with time-varying versus the models with time-invariant effects and determine whether the interaction is significantly dependent on age, we also fit model (5.2) without time-varying effects and the following model with only time-varying main effects (covariates were dropped)

$$f(\boldsymbol{\eta}, \boldsymbol{x}_{kt}) = \sum_l \beta_l^A I(A_{kt} = l) + \sum_i \sum_l \beta_{il}^G I(G_k = i, A_{kt} = l) + \sum_j \sum_l \beta_{jl}^E I(E_{kt} = j, A_{kt} = l) + \sum_i \sum_j d_1 \alpha_i \gamma_j I(G_k = i, E_k = j). \quad (5.11)$$

Table 5.11 shows the log-likelihoods of models (5.2), (5.10), and (5.11) and the corresponding estimates for the interaction parameters. The likelihood increased as the model incorporated time-varying coefficients. Based on the estimates of d_1 from the AMMI1 model (5.10), the magnitude of interaction effects appeared to be age dependent. We observed that $\hat{d}_1$ was largest for age less than 50 and age greater than 80, thereby implying that the effect of interaction between SNP rs7359397 and the exposure groups determined by the criteria of CART – intentional exercise, chronic burden, and trans fat intake – may be particularly manifest in these two age groups.

We also investigated age contributions to the interaction using the same approach described in the data analysis section in Chapter III. Here we replaced $\hat{\alpha}_i$ and $\hat{\gamma}_j$ with the estimates from model (5.2) via the proposed AML estimation algorithm. Figure 5.6 shows contributions of the five age intervals to the first interaction factor. The plot indicates that the modifying effect of health profile (through exposure groups

determined by CART) on the effect of SNP rs7359397 on BMI was largest bewteen age 80 and 89. This was somewhat consistent with the findings of the time-varying interaction model (5.10). Moreover, to evaluate the potential contribution of the second interaction factor, we fit the data with a three-way saturated G×E×Age interaction model and decomposed the estimated interaction matrix for each age group via SVD. The results indicate that the contribution of the second interaction factor to the overall interaction was the largest (27%) for the age interval of 60–69 as opposed to the age intervals of 40–49 and 80–89 (around 10%). As such, a saturated interaction (AMMI2) or another interaction model, rather than AMMI1, may be required to better describe the interaction structure for age 50–79 in this particular GEI example.

5.7 Discussion

We have developed the estimation algorithm and proposed a likelihood-based test for AMMI models to detect GGI and GEI in longitudinal cohort studies. The problem of testing interactions of the AMMI form in unbalanced repeated measures data is complex because the statistic under the null does not have a standard distribution. We approximated the null distribution of the LRT statistic by a χ^2 with fractional df. We applied AMMI models to test for interactions between BMI SNPs and several exposure variables using the MESA data. To reduce the number of tests due to multiple exposures, we used various clustering and classification methods to summarize information of all exposure variables to create an overall health profile with exposure groups. We also considered to extend the AMMI1 model to allow for time-varying effects by categorizing the MESA data into a reasonable number of age groups.

In this chapter, we mainly considered the modeling and test for the first interaction factor in AMMI models. Testing the significance of each multiplicative term and selecting an optimal number of multiplicative terms for a given longitudinal

dataset are natural follow-up questions. Many researchers have studied this problem primarily under the balanced setting in crop cultivar trials. See Cornelius et al. (2001) for a review of several earlier papers concerning sequential testing procedures for AMMI components. Cross validation is another option (Gauch Jr, 1988; Dias and Krzanowski, 2003, 2006). Forkman and Piepho (2014) recently proposed two parametric bootstrap methods for determining the number of multiplicative terms in AMMI models. A comprehensive comparison of model-building strategies using the MESA data would be worthwhile but is beyond the scope of this chapter. In addition, the simulations for assessing the heuristic null distribution of the LRT statistic for AMMI1 were done only for 3×3 and 3×5 table settings according to the data analysis example. It was conjectured that results for tables with other dimension would be similar. This conjecture needs to be examined by further extensive simulation studies.

Inference concerning time-varying interaction was not provided in the present work. Future study is warranted to evaluate time-varying interaction effects appropriately and efficiently. One can use a smooth function to account for the temporal trend of genetic and exposure main effects as well as GEI effects that allows the parameters to change over time. Functional principal component analysis, an effective dimension reduction method, is a potential approach to investigate time-varying GEI. The residual vector (after removing the main effects and covariate effects) can be considered as a random curve, and the time-varying interaction may be investigated by observing the trajectory of functional principal component score with respect to age or time. More research in the area of longitudinal GEI is neede. With an adequate and powerful statistical tool and the availability of rich longitudinal data, we may be able to identify distinctive GEI effects at different stages of life and ultimately have a better understanding of the sophisticated relationship among gene, environment, and complex diseases.

Table 5.1: Moments of the likelihood ratio test statistic for AMMI1 model corresponding to 3×3 and 3×5 table settings from 1000 simulations and moments of χ^2_ν

Table	σ^2	Moment	$\rho = 0$	$\rho = 0.2$	$\rho = 0.7$	χ^{2*}_ν
3×3	4	Mean	3.55	3.53	3.58	3.55
		Variance	6.56	6.64	6.90	7.10
		Skewness	1.63	1.77	1.77	1.50
	10	Mean	3.53	3.68	3.53	3.55
		Variance	7.01	7.87	7.34	7.10
		Skewness	1.59	1.61	1.89	1.50
3×5	4	Mean	6.26	6.26	6.30	6.34
		Variance	11.47	11.39	11.38	12.68
		Skewness	1.21	1.27	1.33	1.12
	10	Mean	6.58	6.26	6.27	6.34
		Variance	12.31	12.17	12.23	12.68
		Skewness	1.30	1.18	1.18	1.12

*ν is estimated through 10000 simulations described in Section 5.4.

Table 5.2: Bias and mean squared error (MSE) of the maximum likelihood estimates of AMMI1 model parameters using the proposed AML estimation procedure (3×3 tables, 1000 simulations)

		$\sigma^2 = 2$			$\sigma^2 = 4$		
Parameter	True	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
Bias ($\times 10^{-3}$)							
μ	12.00	2.4	2.8	3.2	3.4	4.0	4.6
R_1	-0.27	-2.5	-3.3	-3.9	-3.7	-4.6	-5.4
R_2	-0.54	-0.2	-0.2	-0.2	-0.2	-0.2	-0.5
C_1	-0.27	0.9	1.0	1.0	1.3	1.4	1.5
C_2	-0.54	-0.9	-1.0	-1.2	-1.3	-1.5	-2.0
α_1	-0.41	0.2	1.1	1.7	1.9	2.9	4.2
γ_1	-0.71	4.9	7.6	10.1	9.6	16.0	22.2
d_1	1.00	6.9	11.2	15.5	14.8	24.4	33.5
MSE							
μ	12.00	0.90	1.39	1.87	1.79	2.78	3.74
R_1	-0.27	1.85	2.89	3.90	3.72	5.79	7.83
R_2	-0.54	1.41	2.19	2.93	2.84	4.40	5.89
C_1	-0.27	1.53	2.36	3.17	3.03	4.71	6.32
C_2	-0.54	1.59	2.47	3.31	3.17	4.96	6.64
α_1	-0.41	2.71	4.28	5.82	5.54	8.72	11.79
γ_1	-0.71	2.11	3.30	4.48	4.24	8.36	12.79
d_1	1.00	6.21	9.64	12.95	12.38	19.20	25.72

Results of $\hat{R}_3, \hat{C}_3, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\gamma}_2$, and $\hat{\gamma}_3$ are not presented as these estimates can be obtained by constraints: $\sum_i R_i = \sum_j C_j = \sum_i \alpha_i = \sum_j \gamma_j = 0$ and $\sum_i \alpha_i^2 = \sum_j \gamma_j^2 = 1$.

Table 5.3: Bias and mean squared error (MSE) of the maximum likelihood estimates of AMMI1 model parameters using the proposed AML estimation procedure (3×5 tables, 1000 simulations)

		$\sigma^2 = 2$			$\sigma^2 = 4$		
Parameter	True	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
Bias ($\times 10^{-3}$)							
μ	12.00	2.3	2.8	3.2	3.3	4.0	4.6
R_1	-0.27	-0.4	-0.3	-0.5	-0.6	-0.6	-0.6
R_2	-0.54	0.9	1.3	1.3	1.2	1.6	1.9
C_1	-0.62	2.1	2.8	3.1	2.7	3.7	4.3
C_2	0.34	0.3	0.4	0.6	0.6	1.0	1.1
C_3	-0.56	-3.4	-4.5	-5.2	-5.0	-6.5	-7.5
C_4	-0.22	0.0	-0.1	-0.3	-0.1	-0.1	-0.4
α_1	-0.71	3.5	5.8	8.5	7.8	14.1	21.3
γ_1	-0.27	8.0	12.2	15.0	14.7	21.6	29.6
γ_2	-0.27	6.4	9.6	13.7	13.2	20.0	27.1
γ_3	-0.54	8.7	14.2	19.9	18.6	31.7	44.6
d_1	1.00	31.0	46.3	61.0	59.2	88.6	117.3
MSE							
μ	12.00	0.75	1.16	1.55	1.48	2.30	3.10
R_1	-0.27	1.77	2.74	3.70	3.54	5.52	7.43
R_2	-0.54	1.22	1.89	2.53	2.45	3.79	5.07
C_1	-0.62	2.72	4.22	5.69	5.45	8.50	11.49
C_2	0.34	3.04	4.72	6.34	6.09	9.44	12.72
C_3	-0.56	3.14	4.87	6.54	6.26	9.77	13.13
C_4	-0.22	2.84	4.39	5.92	5.69	8.87	11.92
α_1	-0.71	1.65	2.69	3.82	3.59	6.58	10.01
γ_1	-0.27	9.77	15.15	20.62	19.72	31.01	42.43
γ_2	-0.27	10.15	15.76	21.05	20.18	31.15	41.47
γ_3	-0.54	7.24	11.16	14.88	14.23	22.75	31.14
d_1	1.00	14.70	22.27	29.31	28.55	42.59	55.12

Results of $\hat{R}_3, \hat{C}_5, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\gamma}_4$, and $\hat{\gamma}_5$ are not presented as these estimates can be obtained by constraints: $\sum_i R_i = \sum_j C_j = \sum_i \alpha_i = \sum_j \gamma_j = 0$ and $\sum_i \alpha_i^2 = \sum_j \gamma_j^2 = 1$.

Table 5.4: Baseline characteristics of the study participants in MESA

	CAU (N=2527)		CHN (N=775)		AFA (N=1677)		HIS (N=1450)		All (N=6429)	
	N	%	N	%	N	%	N	%	N	%
Gender (female)	1286	52	389	51	842	55	722	51	3239	52
Education										
Less than high school degree	118	5	190	25	167	11	629	45	1104	18
High school degree or above, without bachelor degree	1108	45	278	36	837	54	646	46	2869	46
Bachelor degree or above	1235	50	301	39	538	35	139	10	2213	36
Total annual gross family income										
Less than $20,000	270	11	323	42	335	22	562	40	1490	24
$20,000 or more but less than $50,000	794	32	228	30	633	41	602	43	2257	37
$50,000 or more	1397	57	218	28	574	37	250	18	2439	39
Diagnosis of cancer	318	13	19	2	103	6	71	5	511	8
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age (year)	66.6	10.3	64.7	10.4	64.2	9.9	62.4	10.3	64.3	10.3
Body mass index (kg/m^2)	28.4	5.4	23.7	3.2	30.4	6.1	29.5	5.2	28.8	5.7
Neighborhood measures										
Density of favorable food stores	1.2	2.6	1.7	1.2	3.0	4.5	4.2	5.1	2.8	4.2
Density of recreational facilities	3.0	5.9	3.0	2.6	3.8	7.3	4.4	6.5	3.7	6.2
Perceived healthy food availability	3.3	0.4	3.7	0.2	3.3	0.5	3.5	0.3	3.4	0.4
Perceived walkability	3.9	0.3	3.8	0.2	3.8	0.2	3.8	0.2	3.8	0.2
Combined healthy food environment	-0.9	1.3	0.1	0.5	-0.6	1.7	0.3	1.5	-0.3	1.5
Combined physical activity environment	-0.4	1.4	-0.8	0.8	-0.7	1.3	-0.5	1.2	-0.6	1.3
GRS (count of BMI-increasing alleles)	23.5	3.2	23.4	3.3	23.8	2.9	23.3	3.2	23.7	3.1

CAU = Caucasian, CHN = Chinese, AFA = African American, HIS = Hispanic, GRS = genetic risk score

Table 5.5: Baseline summary of environmental exposure variables for the study participants in MESA

	CAU (N=2527)		CHN (N=775)		AFA (N=1677)		HIS (N=1450)		All (N=6429)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total energy intake (kcal/day)	1514	755	1055	473	1599	955	1605	856	1505	836
Percent calories from carbohydrate intake	51.7	9.0	56.2	7.7	53.3	9.1	55.1	7.9	53.9	8.6
Percent calories from protein intake	15.3	3.3	17.9	3.3	15.3	3.2	16.0	3.2	15.9	3.4
Percent calories from saturated fat intake	11.2	3.4	7.8	2.4	10.2	2.9	10.4	3.1	10.2	3.2
Percent calories from trans fat intake	0.9	0.3	0.5	0.2	1.0	0.4	0.7	0.3	0.8	0.4
Total intentional exercise (MET-min/wk)	1467	2509	1023	1383	1607	2615	1187	1973	1349	2252
Physical activity* (MET-min/wk)	6268	6217	3595	3734	6237	6546	5629	6184	5695	6099
Psychosocial factors										
Trait anger	14.5	3.3	14.7	3.5	14.2	3.6	15.0	4.4	14.6	3.9
Trait anxiety	17.0	4.7	16.2	4.7	15.7	4.6	16.2	4.8	16.3	4.7
Chronic burden	1.3	1.2	0.8	1.1	1.4	1.3	1.2	1.2	1.2	1.2
Depressive symptoms (CESD)	8.4	7.3	6.2	6.8	8.5	8.1	10.0	9.2	8.7	8.2

CAU = Caucasian, CHN = Chinese, AFA = African American, HIS = Hispanic

CESD = Center for Epidemiologic Studies Depression Scale

* Moderate and vigorous physical activity

Table 5.6: Information of the 27 SNPs that are associated with BMI at genome-wide significance ($p < 5 \times 10^{-8}$) levels and the test results of their main effects on BMI (using additive models), adjusted for age, age^2, gender, and the first three global principal components in the MESA data

SNP	SNP ID	Nearby Genes	BMI-increasing Allele	Other	Freq of BMI-increasing Allele	Standardized Effect Size	Relative Effect Size	p-value
1	rs2815752	NEGR1	A	G	61%	0.000481	0.850291	0.748
2	rs543874	SEC16B	G	A	19%	0.000819	1.447792	0.016
3	rs1514175	TNNI3K	A	G	43%	0.000195	0.344712	0.724
4	rs1555543	PTBP2	C	A	59%	0.000140	0.247486	0.774
5	rs2867125	TMEM18	C	T	83%	0.001750	3.093573	0.001
6	rs713586	RBJ,ADCY3,POMC	C	T	47%	0.000641	1.133132	0.002
7	rs9816226	ETV5	T	A	82%	0.000336	0.593966	0.037
8	rs13078807	CADM2	G	A	20%	0.000144	0.254557	0.326
9	rs10938397	GNPDA2	G	A	43%	0.000912	1.612194	0.746
10	rs13107325	SLC39A8	T	C	7%	0.000282	0.498507	0.504
11	rs2112347	FLJ35779,HMGCR	T	G	63%	0.000228	0.403048	0.031
12	rs987237	TFAP2B	G	A	18%	0.000310	0.548004	0.093
13	rs206936	NUDT3,HMGA1	G	A	21%	0.000091	0.161043	0.075
14	rs10968576	LRRN6C	G	A	31%	0.000234	0.413655	0.292
15	rs10767664	BDNF	A	T	78%	0.000750	1.325817	0.020
16	rs3817334	MTCH2,NDUFS3,CUGBP1	T	C	41%	0.000157	0.277538	0.623
17	rs7138803	FAIM2	A	G	38%	0.000375	0.662909	0.069
18	rs4771122	MTIF3,GTF3A	G	A	24%	0.000156	0.275770	0.081
19	rs10150332	NRXN3	C	T	21%	0.000295	0.521488	0.233
20	rs11847697	PRKD1	T	C	4%	0.000130	0.229808	0.337
21	rs2241423	MAP2K5,LBXCOR1	G	A	78%	0.000353	0.624018	0.052
22	rs1558902	FTO	A	T	42%	0.004080	7.212445	0.038
23	rs7359397	SH2B1,APOB48R,SULT1A2	T	C	40%	0.000581	1.027066	1.9×10^{-4}
24	rs12444979	GPRC5B,IQCK	C	T	87%	0.000404	0.714173	0.058
25	rs571312	MC4R	A	C	24%	0.001140	2.015242	0.083
26	rs29941	KCTD15	G	A	67%	0.000074	0.129930	0.111
27	rs3810291	TMEM160,ZC3H4	A	G	67%	0.000216	0.381835	0.042
Genetic risk score (count of BMI-increasing alleles)								6.7×10^{-5}

Table 5.7: Estimated main effects of exposure and neighborhood environment variables on BMI adjusted for age, age^2, gender, race, education, income, and diagnosis of cancer in the MESA data

Variable	Estimate	SE	p-value
Total Energy Intake (kcal/day)	6.4×10^{-4}	8.5×10^{-5}	<0.0001
Percent Calories from Carbohydrate	-0.060	0.007	<0.0001
Percent Calories from Protein	0.053	0.020	0.007
Percent Calories from Saturated Fat	0.180	0.021	<0.0001
Percent Calories from Trans Fat	1.737	0.205	<0.0001
Total Intentional Exercise (MET-min/wk)	-4.4×10^{-5}	6.4×10^{-6}	<0.0001
Physical Activity* (MET-min/wk)	-1.4×10^{-5}	2.7×10^{-6}	<0.0001
Trait Anger	0.005	0.006	0.410
Trait Anxiety	-0.004	0.005	0.383
Chronic Burden	0.032	0.015	0.029
Depressive Symptoms (CESD)	-0.004	0.002	0.029
Density of Favorable Food Stores	-0.047	0.010	<0.0001
Recreational Resources Density	-0.011	0.003	0.001
Perceived Healthy Foods Availability	0.054	0.050	0.284
Perceived Walkability	-0.136	0.090	0.132
Combined Healthy Food Environment	-0.020	0.020	0.313
Combined Physical Activity Environment	-0.063	0.021	0.002

* Moderate and vigorous physical activity

Table 5.8: Estimates of the exposure cluster main effects on BMI adjusted for age, age^2, gender, race, education, income, diagnosis of cancer, and the first three principal components in the MESA data

Effect	Estimate	SE	p-value
K-means: Group B vs. Group A	1.982	0.300	<0.0001
K-means: Group C vs. Group A	0.434	0.297	0.143
K-means: Group D vs. Group A	-0.189	0.288	0.512
K-means: Group E vs. Group A	0.944	0.307	0.002
K-means: Group F vs. Group A	0.671	0.281	0.017
LCA: Group B vs. Group A	0.056	0.211	0.792
LCA: Group C vs. Group A	1.214	0.259	<0.0001
LCA: Group D vs. Group A	0.825	0.203	<0.0001
LCA: Group E vs. Group A	0.021	0.214	0.922
LCA: Group F vs. Group A	1.560	0.236	<0.0001
CART: Group 2 vs. Group 1	1.205	0.215	<0.0001
CART: Group 3 vs. Group 1	1.107	0.179	<0.0001
CART: Group 4 vs. Group 1	1.995	0.219	<0.0001
CART: Group 5 vs. Group 1	3.252	0.248	<0.0001

See Figure 5.2, Figure 5.3, and Figure 5.4 for the characteristics of each group classified by k-means, latent class analysis (LCA), and classification and regression tree (CART), respectively.

Table 5.9: P-values from the tests of interactions between 27 SNPs and 11 exposure variables, adjusted for age, age^2, gender, race, education, income, diagnosis of cancer, and the first three principal components in the MESA data

			Dietary Variable				Physical Activity (PA)			Psychosocial Factors		
SNP	SNP ID	Total Energy	Carb.	Protein	Sat. Fats	Trans Fats	Intentional Exercise	Mod./Vig. PA	Anger	Anxiety	Chronic Burden	CESD
1	rs2815752	0.083	0.787	0.617	0.207	0.709	0.771	0.182	0.649	0.845	0.766	0.387
2	rs543874	0.357	0.179	0.997	0.465	0.680	0.398	0.839	0.707	0.789	0.005	0.710
3	rs1514175	0.058	0.980	0.622	0.500	0.607	0.646	0.846	0.021	0.594	0.910	0.361
4	rs1555543	0.320	0.280	0.042	0.120	0.285	0.963	0.356	0.401	0.833	0.234	0.157
5	rs2867125	0.952	0.620	0.057	0.374	0.622	0.864	0.486	0.406	0.312	0.438	0.449
6	rs713586	0.225	0.942	0.686	0.799	0.575	0.607	0.363	0.045	0.118	0.002	0.814
7	rs9816226	0.237	0.971	0.628	0.884	0.468	0.944	0.993	0.916	0.462	0.039	0.574
8	rs13078807	0.838	0.755	0.390	0.823	0.570	0.330	0.301	0.703	0.316	0.592	0.680
9	rs10938397	0.098	0.632	0.823	0.946	0.645	0.499	0.708	0.198	0.971	0.243	0.434
10	rs13107325	0.013	0.685	0.298	0.898	0.003	0.104	0.924	0.228	0.217	0.695	0.752
11	rs2112347	0.696	0.926	0.068	0.185	0.304	0.158	0.715	0.578	0.526	0.107	0.491
12	rs987237	0.985	0.634	0.344	0.383	0.832	0.159	0.640	0.773	0.570	0.547	0.030
13	rs206936	0.231	0.836	0.622	0.544	0.681	0.108	0.176	0.341	0.974	0.068	0.635
14	rs10968576	0.290	0.224	0.197	0.020	0.760	0.874	0.779	0.746	0.269	0.582	0.795
15	rs10767664	0.721	0.662	0.959	0.354	0.085	0.370	0.449	0.046	0.078	0.359	0.986
16	rs3817334	0.952	0.073	0.805	0.040	0.551	0.644	0.850	0.790	0.806	0.001	0.096
17	rs7138803	0.863	0.173	0.597	0.813	0.728	0.652	0.637	0.468	0.315	0.262	0.091
18	rs4771122	0.084	0.640	0.525	0.358	0.771	0.075	0.433	0.591	0.793	0.461	0.582
19	rs10150332	0.571	0.016	0.505	0.737	0.490	0.342	0.436	0.008	0.001	0.653	0.191
20	rs11847697	0.022	0.885	0.137	0.115	0.107	0.030	0.343	0.557	0.940	0.028	0.800
21	rs2241423	0.564	0.113	0.879	0.750	0.212	0.873	0.262	0.960	0.555	0.430	0.303
22	rs1558902	0.152	0.278	0.387	0.253	0.197	0.298	0.318	0.154	0.449	0.002	0.289
23	rs7359397	0.976	0.923	0.475	0.269	0.814	0.003	0.574	0.381	0.544	0.724	0.168
24	rs12444979	0.019	0.819	0.458	0.304	0.110	0.598	0.143	0.750	0.771	0.313	0.695
25	rs571312	0.970	0.014	0.821	0.020	0.457	0.244	0.658	0.607	0.531	0.709	0.811
26	rs29941	0.029	0.372	0.345	0.454	0.341	0.345	0.467	0.149	0.238	0.149	0.227
27	rs3810291	0.645	0.524	0.123	0.195	0.360	0.052	0.617	0.948	0.895	0.926	0.741
GRS		0.076	0.232	0.743	0.692	0.746	0.782	0.751	0.210	0.032	0.968	0.324

Table 5.10: P-values for the tests of interaction between overall health clustering (via k-means, LCA, and CART) and each of the 27 BMI-related SNPs using AMMI1 and saturated interaction (SAT) models in the MESA data

SNP	Chromosome	SNP ID	Gene	K-means		LCA		CART	
				AMMI1	SAT	AMMI1	SAT	AMMI1	SAT
2	1	rs543874	SEC16B	0.628	0.516	0.590	0.552	0.018	0.006
6	2	rs713586	RBJ, ADCY3, POMC	0.761	0.836	0.278	0.155	0.200	0.247
16	11	rs3817334	MTCH2, NDUFS3, CUGBP1	0.051	0.020	0.128	0.228	0.359	0.476
19	14	rs10150332	NRXN3	0.999	1.000	0.822	0.831	0.829	0.907
22	16	rs1558902	FTO	0.029	0.056	0.762	0.841	0.456	0.612
23	16	rs7359397	SH2B1, APOB48R, SULT1A2	0.047	0.043	0.346	0.307	0.007	0.005
GRS (count of BMI-increasing alleles)*				0.022	0.003	0.220	0.132	0.136	0.014

*The genetic risk score (GRS) was categorized into 5 groups based on quintiles.

Covariates were age, age^2, gender, race, education, income, diagnosis of cancer, and the first three principal components.

Table 5.11: Interaction parameter estimates using time-invariant and time-varying AMMI1 models for the GEI between SNP rs7359397 and overall health represented by the exposure groups using CART in the MESA data

				Genetic Interaction			Environment Interaction				
Model	Log-likelihood	Age	$\hat{d}_1$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$
(5.2)	-42734.94	40–89	0.95	-0.49	0.81	-0.33	-0.19	0.65	0.22	0.03	-0.70
(5.11)	-42715.33	40–89	0.95	-0.50	0.81	-0.31	-0.19	0.64	0.22	0.03	-0.71
(5.10)	-42625.89	40–49	2.66	-0.78	0.19	0.59	0.48	0.27	0.24	-0.22	-0.77
		50–59	1.40	-0.66	0.75	-0.08	-0.07	0.55	0.39	-0.15	-0.72
		60–69	0.78	-0.61	0.78	-0.17	-0.29	0.58	0.00	0.37	-0.67
		70–79	0.95	-0.81	0.46	0.35	-0.15	0.75	-0.20	0.19	-0.59
		80–89	2.47	-0.23	0.79	-0.56	-0.27	0.31	-0.25	0.72	-0.50

The model adjusted for age, age^2, gender, race, education, income, diagnosis of cancer, and the first three principal components.

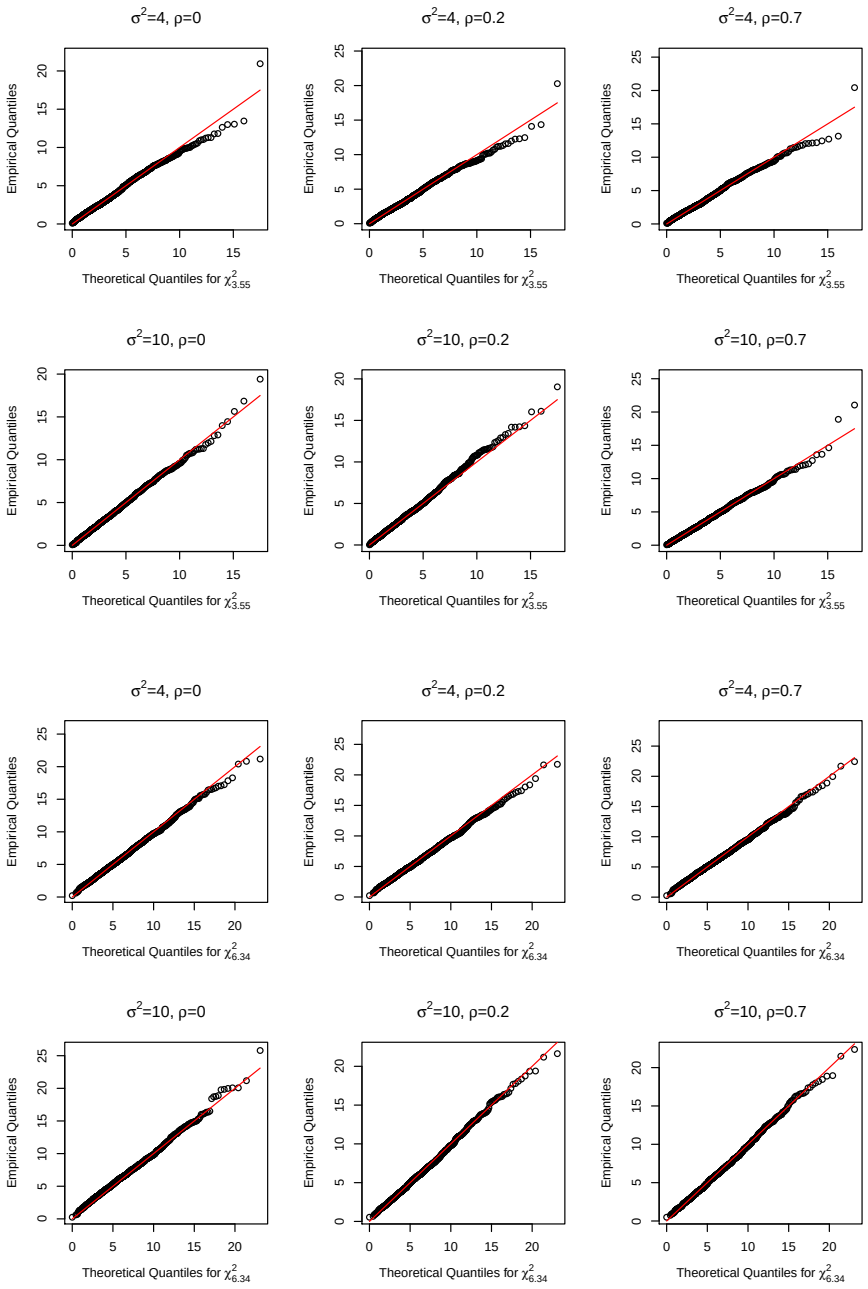


Figure 5.1: Quantile-Quantile (Q-Q) plot for comparing the distribution of the LRT statistics for AMMI1 with the coorresponding χ^2 distribution under H_0 : $d_1 = 0$. Data were simulated under (a) 3×3 and (b) 3×5 GEI two-way tables (N=2000 with repeated measures).

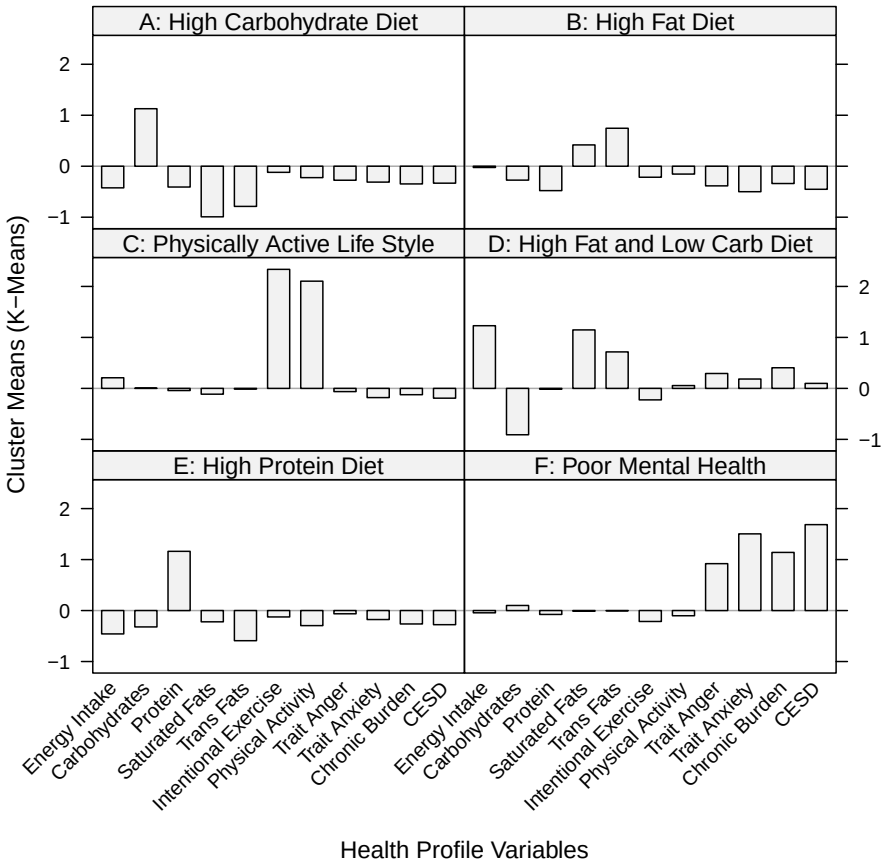


Figure 5.2: Cluster means of the 11 health profile variables (standardized) using k-means in the MESA data

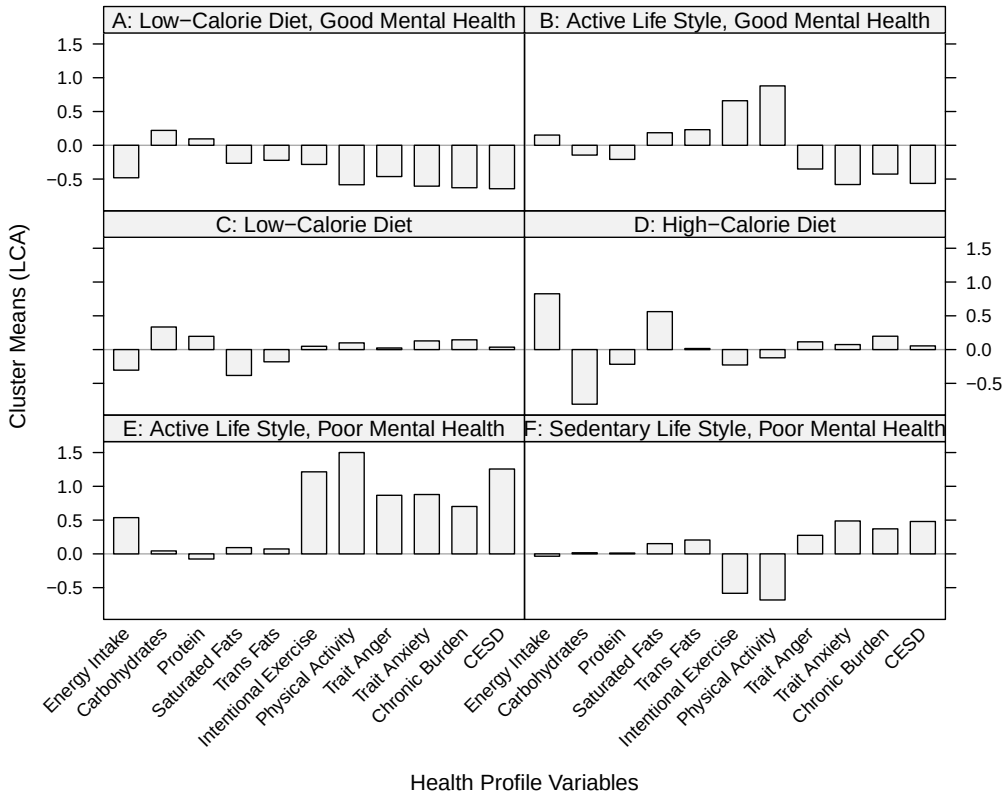


Figure 5.3: Cluster means of the 11 health profile variables (standardized) from the results of LCA in the MESA data. Except for percent calories from carbohydrates, percent calories from protein, and percent calories from saturated and trans fats, all others variables were log-transformed to achieve normality.

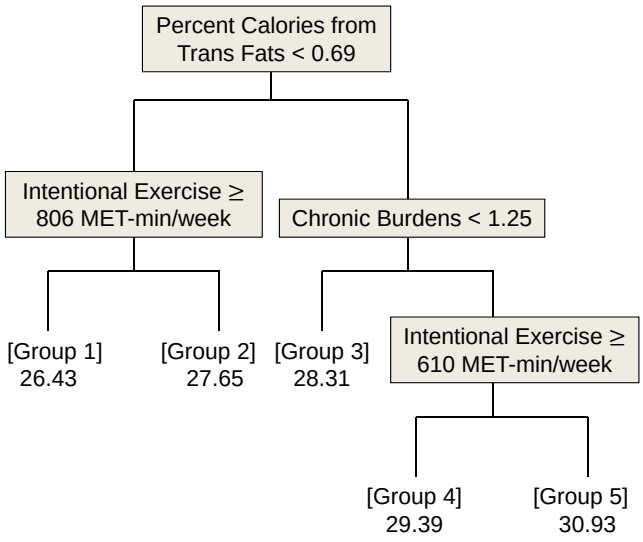


Figure 5.4: Grouping criteria of classification and regression tree (CART) analysis results. Means of BMI for the five groups are shown.

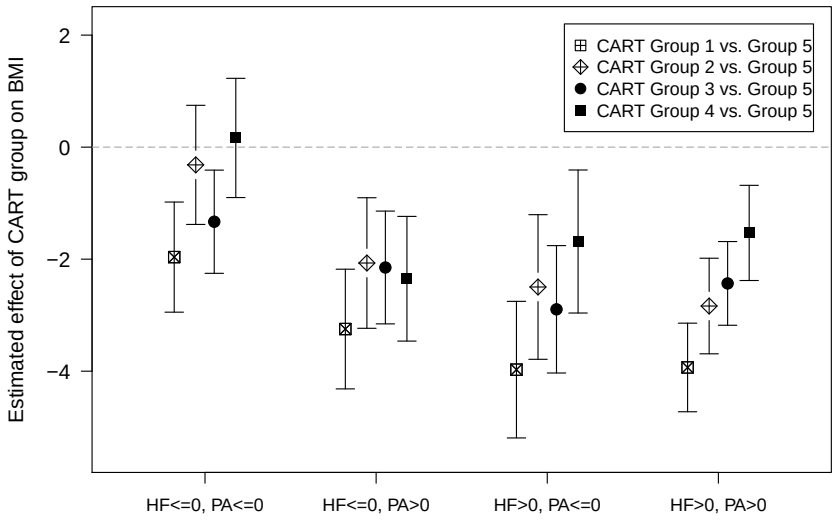


Figure 5.5: Estimated effect of CART group on BMI and the corresponding confidence intervals for four neighborhood environment groups based on combined healthy food (HF) and combined physical activity (PA) environment indices

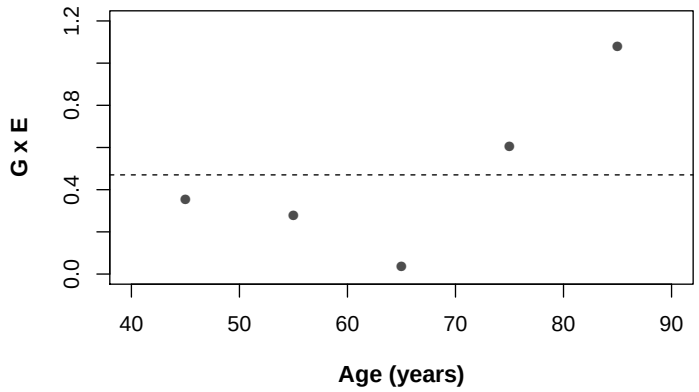


Figure 5.6: Age-specific contributions to the *first* interaction factor in SNP rs7359397 × exposure groups (determined by CART) based on the MESA data

5.8 Appendix

5.8.1 Variables in the MESA Data Analysis

- **Body mass index (BMI) [exams 1, 2, 3 ,4]** was calculated as weight (in kilograms) divided by height squared (in meters^2). Weight and height were measured using a balanced beam scale and a vertical ruler, respectively, with participants wearing light clothing and no shoes. Height was recorded to the nearest 0.5 cm and weight to the nearest 0.5 lb. Out of 6429 MESA participants, 393 (6%) had only one BMI measurement, 342 (5%) had two BMI measurements, 502 (8%) had three BMI measurements, and 5192 (81%) had four BMI measurements.

- **Total energy intakes (kcal/day), percent calories from carbohydrate intake, percent calories from protein intake, percent calories from saturated fat intake, and percent calories from trans fat intake [exam 1]** were measured through dietary intake from a self-administered, 120-item, modified-Block style food frequency questionnaire (Nettleton et al., 2006). Participants recorded the serving size (small, medium, or large) and frequency of consumption (average times per day, week, or month) of specific beverages and foods. Daily frequency responses were weighted according to reported serving sizes (small: frequency $\times$ 0.5; medium: frequency $\times$ 1.0; large: frequency $\times$ 1.5) and consequent servings/day were categorized into 47 food groups and the corresponding calorie intakes were calculated.

- **Moderate and vigorous physical activity (MET-minute/week)[exams 1, 2, 3]** was obtained from MESA Typical Week Physical Activity Survey, which is a self-report questionnaire developed for the MESA. The questionnaire was designed to identify the time spent in and frequency of various physical activities during a typical week in the past month to capture typical activity

patterns in one's daily life. The survey has 28 items in 9 categories: household chores, lawn/yard/garden/farm, care of children/adults, transportation, walking (not at work), dancing and sport activities, conditioning activities, leisure activities, and occupational and volunteer activities. The questions differentiated between light-, moderate-, and vigorous-intensity activities. Participants reported the average number of days per week and time per day engaged in these activities. Minutes of activity were summed for each discrete activity type and multiplied by metabolic equivalent (MET) level to compute the total MET-minutes per week. Moderate and vigorous physical activity was derived by the sum of moderate and vigorous MET-minutes/week (Bertoni et al., 2009).

- **Total intentional exercise (MET-minute/week)[exams 1, 2, 3]** was also derived from MESA Typical Week Physical Activity Survey. It was the sum of walking for exercise, sports/dancing, and conditioning MET-minutes/week (Bertoni et al., 2009). The average daily time spent on doing intentional exercise among the participants was 210 MET-minutes (equivalent to approximately two hours of running per week), and the median daily time was 120 MET-minutes (equivalent to approximately 1.2 hours of running per week).

- **The trait anger and trait anxiety scales (Spielberger Trait Anger and Anxiety Inventory) [exams 1, 3]**: the trait anger scale was designed to assess an individuals disposition to feeling angry, and trait scales were chosen over state scales to better capture the relations that occur over longer periods. The trait anxiety scale captures differences between people in their disposition to respond to stressful situations with varying amounts of state anxiety. Possible ranges are 10 to 40 for trait anger and 10 to 40 for trait anxiety. Higher scores indicated higher levels of anger and anxiety, respectively.

- **The chronic burden scale [exams 1, 3]:** Respondents were asked to indicate

whether they had experienced any ongoing difficulties in five domains (personal health, health of close others, relationship, work, and finances) (Bromberger and Matthews, 1996). For each of the five domains, respondents were classified as having chronic burden if they had experienced the circumstance for six or more months and it was moderately or very stressful. The number of domains in which chronic burden was experienced was the estimate of overall chronic burden. The range of the scale was 0 to 5.

- **Depressive symptoms [exams 1, 3, 4]** were evaluated by the Center for Epidemiologic Studies Depression (CESD) Scale (Radloff, 1977). The range of the scale was 0 to 60. Higher scores indicating greater symptoms.

- **Perceived healthy food availability and walkability [exams 1, 2, 3, 4]** were obtained by integrating responses from questionnaires administered to MESA participants during MESA exams 2 and 3 (2002-2005) and MESA exam 5 (2010-2011) and two auxiliary surveys, the community surveys (CS), administered in 2004 and 2010 to random samples of other residents of MESA neighborhoods. In both surveys, respondents were asked to consider their "neighborhood" as the area within 20-minute walk or 1 mile from their home. Questions regarding healthy food availability and walkability were answered using a 5-point Likert scale (see Table 5.12 for the questions). To maximize the use of available data, measures were aggregated at the census tract level by pooling all available MESA and CS respondents in each tract using conditional empirical Bayes estimates (Mujahid et al., 2008) adjusted for respondents age, sex, source, and site. For example, a mean walkability for neighborhood k is a weighted average of the estimated crude mean walking environment for neighborhood k and the estimated grand neighborhood walking environment mean. The range of walkability is from 1 to 5. These estimates of neighborhood conditions were

linked to each study participant based on the census tract of residence. Higher values represent better neighborhood conditions.

5.8.2 Principal Component Analysis of the MESA data

Population stratification and admixture can result in nonhomogeneous distribution of allele frequencies for some genes in the study population. Consequently, it can cause type I error inflation and/or loss of power in genetic association studies. They are known to affect studies that collect data from multi-ethnic samples and even those that target a single ethnicity and structured association tests (Pritchard and Donnelly, 2001). Recent studies have shown that principal component analysis (PCA) can provide control variables that yield type I error control similar to what was previously observed with structured association tests (Patterson et al., 2006; Price et al., 2006; Zhang et al., 2003). More importantly, this method does not require that a panel of ancestry informative markers and is readily available as is the case with the structured association tests approach. Here, we describe our approach to compute principal components (PCs) in the MESA-SHARe study.

The PCs were computed on 8227 individuals. 2590 self-report as African-Americans, 2174 as Hispanics-Americans, 2686 as European-Americans and 777 as Chinese. The PCs were computed separately in each self-reported ethnic group as well as in the combined sample. Only the results of the combined analysis are relevant to this chapter thus are reported here. Chromosome-specific PCs were first computed to help reduce the computational burden, and these PCs were later combined to provide the final set of eigenvalues and eigenvectors.

The combined analysis results show that the first three PCs should be considered for further analysis according to the scree plot (elbow rule). The first PC explained about 78%, the second explained about 16%, and the third accounted for less than 1% of the observed variation. These three PCs together explained about 96% of the

total observed variation. The projection in the space represented by the first two PCs in the combined analysis revealed two clines: variations between European and African and variations between European and Chinese. The Hispanics appeared to be the most heterogeneous group with some clustering with one of the other three ethnicities in MESA and others displaying various levels of admixture, implying that the third PC in the combined analyses seemed to explain a forth ancestral population.

Table 5.12: Questions for healthy food availability and walkability in MESA

Walking Environment:
1. My neighborhood offers many opportunities to be physically active.
2. Local sports clubs and other facilities in my neighborhood offer many opportunities to get exercise.
3. It is pleasant to walk in my neighborhood.
4. The trees in my neighborhood provide enough shade.
5. In my neighborhood it is easy to walk places.
6. I often see other people walking in my neighborhood.
7. I often see other people exercising (e.g., jogging, bicycling, and playing sports) in my neighborhood.
8. My neighborhood has heavy traffic.
9. There are busy roads to cross when out for walks in my neighborhood.
10. In my neighborhood it is easy to walk places.
Availability of Healthy Foods:
1. A large selection of fresh fruits and vegetables is available in my neighborhood.
2. The fresh fruits and vegetables in my neighborhood are of high quality.
3. A large selection of low-fat products is available in my neighborhood.
4. There are many opportunities to purchase fast foods in my neighborhood.

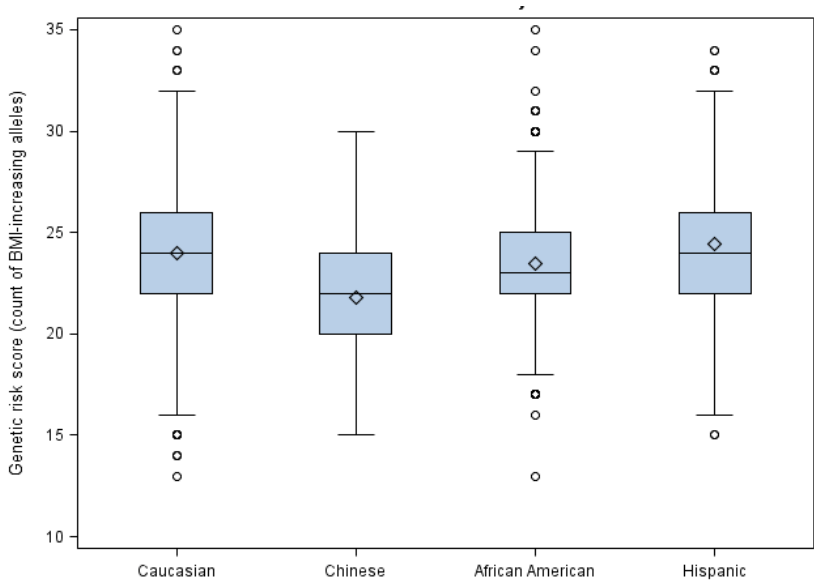


Figure 5.7: Boxplot of genetic risk scores (count of BMI-increasing alleles) using 27 SNPs for the four race groups from the MESA data

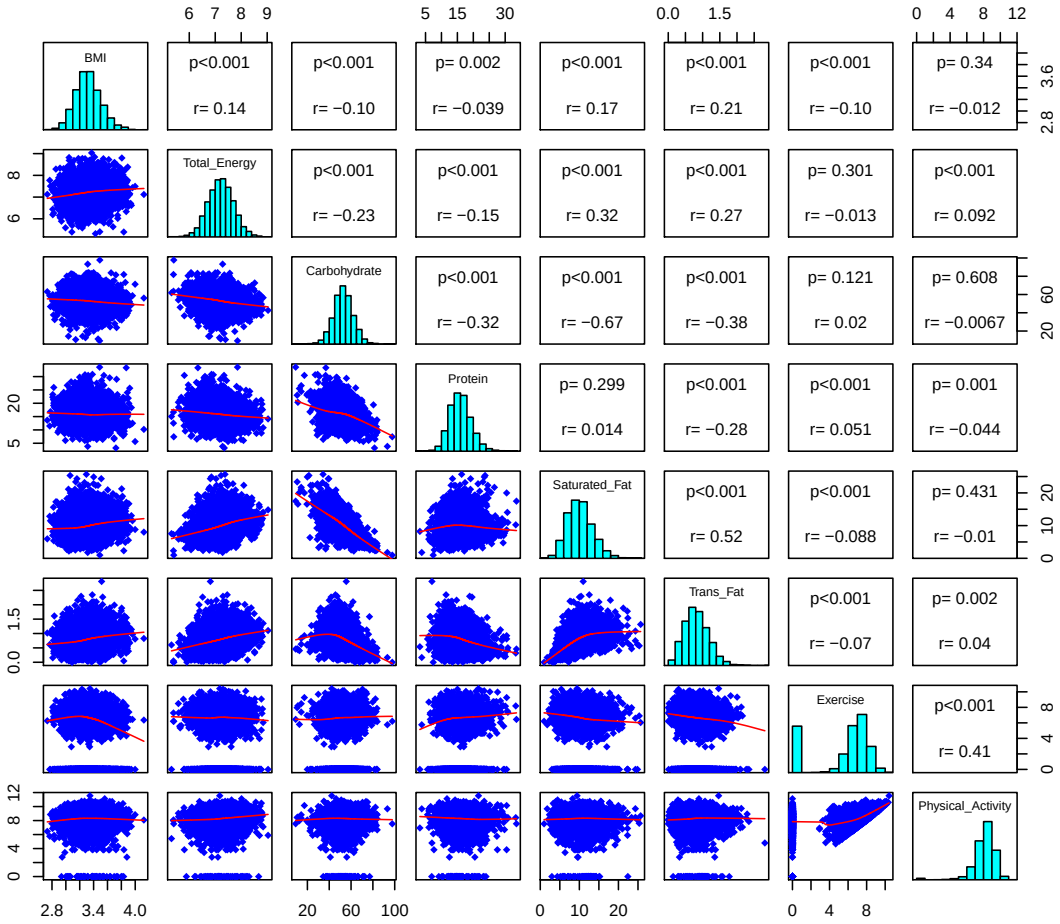


Figure 5.8: Baseline Pearson's correlation coefficients and the corresponding p-values among the five dietary variables (total energy intake was log-transformed), intentional exercise (log-transformed), moderate and vigorous physical activities (log-transformed), and BMI (log-transformed) in the MESA data

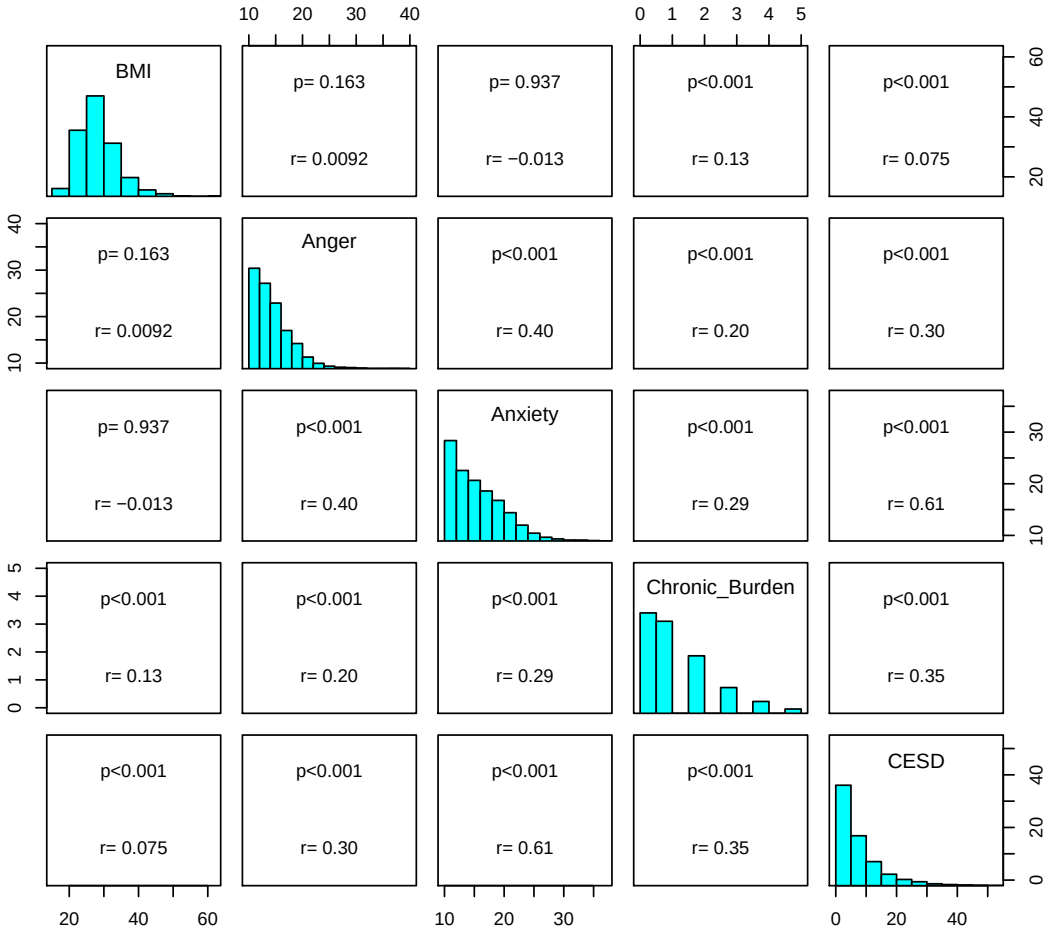


Figure 5.9: Baseline Spearman's correlation coefficients and the corresponding p-values among the four psychosocial factors and BMI in the MESA data

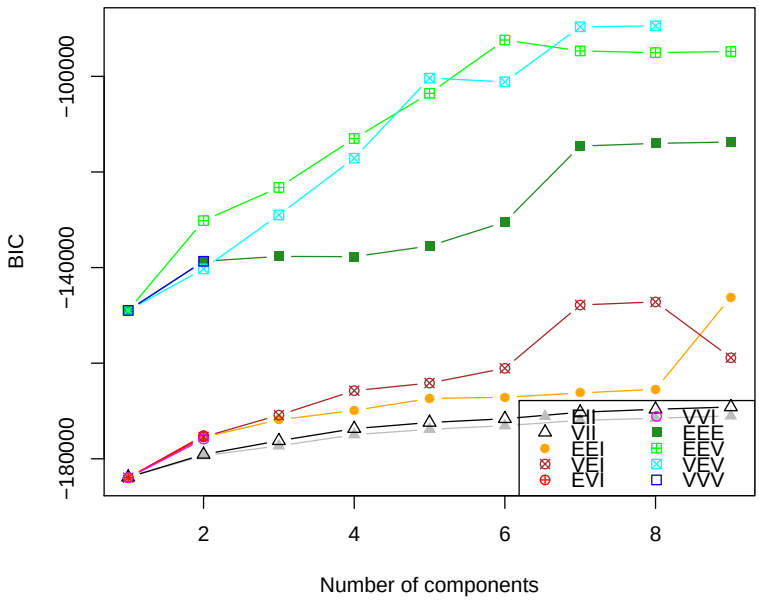


Figure 5.10: Bayesian information criterion (BIC) values of different number of clusters and different covariance structures from latent class analysis (LCA) results. Each letter in the name of a model corresponds to the constraint placed on the volume (a constant), shape (a diagonal matrix with entries proportional to the eigenvalues) and orientation/direction (an orthogonal matrix of eigenvectors) of the cluster covariance matrices respectively. The constraint can be equal (E), variable (V) or identity (I).

CHAPTER VI

Conclusions and Future Work

This chapter reviews the dissertation, summarizes its contributions, and discusses directions for future study. This dissertation has concentrated on the problem of adapting classical models for non-additivity proposed in the ANOVA literature, originally designed for single observations per cell, to modeling of GEI/GGI in longitudinal cohort studies. There are, however, still many extensions of this research that deserve further investigation.

6.1 Summary of This Dissertation

In Chapter II, we first provided an overview of the classical interaction models that use a sparse representation of interaction structure to save df for the interaction term. Then we explored the interaction structures by simply reducing repeated measures data to summary level cell means in a two-way layout. Even though the unbalanced repeated measures data structures were not taken into account, the cell-mean based method can still provide an exploratory analysis of interaction structures.

In Chapter III, we modified the cell-mean based method to properly account for the correlation across repeated measurements. Moreover, we developed a unique parametric bootstrap procedure for testing GEI and GGI in the form of Tukey's, Mendel's, and AMMI models. Specifically, the proposed two-stage estimation approach was performed in a regression setting that can be incorporated into any standard mixed model estimation tools. Both proposed methods accounted for the unbalanced and

longitudinal nature of the outcome data and were developed based on likelihood ratio test. These classical interaction and AMMI models were compared to a fully saturated interaction model. Our simulation studies showed that AMMI1 appeared to be a robust and flexible model in detecting interaction effects across a spectrum of interaction structures. More importantly, it was relatively powerful in detecting certain epistasis structures when main effects were absent. Tukey's and Mandel's models, unfortunately, could fail to detect interactions when the structure was misspecified.

In Chapter IV, we proposed the estimation algorithm for Tukey's 1-df model and showed that it may be a useful model for GEI when both G and E have strong main effects. We also proposed an adaptive shrinkage estimator that combines estimates from Tukey's one-df model and a saturated interaction model for GEI. The shrinkage estimator shrinks the MLEs under a general, saturated interaction structure toward Tukey's one-df model estimator that allows for data-adaptive relaxation of the structural assumption in Tukey's product form. Our unique simulation setting of multiple GEI tests represented the search for GEI over many candidate SNPs with different interaction patterns. The results indicated that the test based on the shrinkage estimator can be considered as a robust and unified approach for interaction detection.

In Chapter V, we developed the estimation algorithm specifically for AMMI1 models and proposed a corresponding likelihood-based test for GEI in longitudinal cohort studies. The null distribution of the test statistic was approximated by a chi-square with an estimated fractional df. We applied AMMI models to tests for interactions between BMI SNPs and several exposure variables (e.g., diet, exercise, psychosocial factors) using the data in MESA. To illustrate the analysis in a two-way table under the framework of categorical G and E, we summarized information from multiple exposures to create an overall health profile using various clustering and classification methods. In addition, we considered to extend the model to allow for time-varying effects by categorizing the MESA data into several age groups.

6.2 Future Work

Statistical methods specifically designed for GEI/GGI in longitudinal cohort studies are still very limited. There are several interesting research questions pertinent to this dissertation that deserve future work. First, the problem of selecting an optimal number of interaction factors under the AMMI model is an important topic that needs further research to obtain appropriate inference. Second, new strategies are needed to efficiently assess time-varying interaction effects using data from longitudinal cohort studies in order to better understand the long-term implications of GEI in public health. Third, given that we have only focused on normally distributed random effects and errors, future study could investigate the application of these classical interaction models to outcome data following other common distributions. Fourth, instead of considering the interaction as fixed effect, one could treat it as random effect or assuming that G is fixed and E is random (or vice versa), which will lead to a totally different interpretation. Lastly, given that the classical interaction models and AMMI models were developed under the balanced design setting for crop cultivar performance trials, the impact of missingness on interaction tests using these models has not been considered. However, missing data or dropouts during follow-up visits is an inevitable issue in human studies and deserves future investigation. In conclusion, the development of a comprehensive modeling approach and a powerful testing device tailored to GEI/GGI in longitudinal settings is necessary to provide a thorough understanding of the contributions of genes and environmental exposures to complex diseases.

BIBLIOGRAPHY

BIBLIOGRAPHY

Barhdadi, A. and Dubé, M. (2010). Testing for gene-gene interaction with AMMI models. *Statistical Applications in Genetics and Molecular Biology* **9,** 1–27.

Bates, D. and Watts, D. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.

Bell, B., Rose, C., and Damon, A. (1966). The veterans administration longitudinal study of healthy aging. *Gerontologist* **6,** 179–184.

Bertoni, A. G., Whitt-Glover, M. C., Chung, H., Le, K. Y., Barr, R. G., Mahesh, M., Jenny, N. S., Burke, G. L., and Jacobs, D. R. (2009). The association between physical activity and subclinical atherosclerosis the multi-ethnic study of atherosclerosis. *American Journal of Epidemiology* **169,** 444–454.

Bielinski, S. J., Pankow, J. S., Li, N., Hsu, F.-C., Adar, S. D., Jenny, N. S., Bowden, D. W., Wasserman, B. A., and Arnett, D. (2008). ICAM1 and VCAM1 polymorphisms, coronary artery calcium, and circulating levels of soluble ICAM-1: the multi-ethnic study of atherosclerosis (MESA). *Atherosclerosis* **201,** 339–344.

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Roux, A. V. D., Folsom, A. R., Greenland, P., JacobsJr, D. R., Kronmal, R., Liu, K., et al. (2002). Multi-ethnic study of atherosclerosis: objectives and design. *American Journal of Epidemiology* **156,** 871–881.

Boik, R. (1989). Reduced-rank models for interaction in unequally replicated two-way classifications. *Journal of Multivariate Analysis* **28,** 69–87.

Bookman, E., McAllister, K., Gillanders, E., Wanke, K., Balshaw, D., Rutter, J., Reedy, J., Shaughnessy, D., Agurs-Collins, T., Paltoo, D., et al. (2011). Gene-environment interplay in common complex diseases: forging an integrative model–recommendations from an NIH workshop. *Genetic Epidemiology* **35,** 217–225.

Bradman, A., Eskenazi, B., Sutton, P., Athanasoulis, M., and Goldman, L. R. (2001). Iron deficiency associated with higher blood lead in children living in contaminated environments. *Environmental Health Perspectives* **109,** 1079–1084.

Bradu, D. and Gabriel, K. (1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics* **20,** 47–68.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). Classification and Regression Trees. Technical report, Wadsworth.

Brem, R., Storey, J., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436,** 701–703.

Bromberger, J. T. and Matthews, K. A. (1996). A longitudinal study of the effects of pessimism, trait anxiety, and life stress on depressive symptoms in middle-aged women. *Psychology and Aging* **11,** 207–213.

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

Chen, C., He, X., and Wei, Y. (2008). Lower rank approximation of matrices based on fast and robust alternating regression. *Journal of Computational and Graphical Statistics* **17,** 186–200.

Chen, Y., Chatterjee, N., and Carroll, R. (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association* **104,** 220–233.

Chung, J. and Wessling-Resnick, M. (2003). Molecular mechanisms and regulation of iron transport. *Critical Reviews in Clinical Laboratory Sciences* **40,** 151–182.

Cornelius, P. (1980). Functions approximating Mandel's tables for the means and standard deviations of the first three roots of a Wishart matrix. *Technometrics* **22,** 613–616.

Cornelius, P. (1993). Statistical tests and retention of terms in the additive main effects and multiplicative interaction model for cultivar trials. *Crop Science* **33,** 1186–1193.

Cornelius, P., Seyedsadr, M., and Crossa, J. (1992). Using the shifted multiplicative model to search for separability in crop cultivar trials. *Theoretical and Applied Genetics* **84,** 161–172.

Cornelius, P. L., Crossa, J., Seyedsadr, M. S., Liu, G., and Viele, K. (2001). Contributions to multiplicative model analysis of genotype-environment data. In *Statistical Consulting Section, American Statistical Association, Joint Statistical Meetings, August*, volume 7.

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests for goodness-of-fit of a nonlinear regression model. *Journal of Multivariate Analysis* **91,** 35–52.

Crossa, J., Gauch, H., and Zobel, R. (1990). Additive main effects and multiplicative interaction analysis of two international maize cultivar trials. *Crop Science* **30,** 493–500.

Croux, C., Filzmoser, P., Pison, G., and Rousseeuw, P. (2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing* **13,** 23–36.

Cruickshanks, K., Wiley, T., Tweed, T., Klein, B., Klein, R., Mares-Perlman, J., and Nondahl, D. (1998). Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin. *American Journal of Epidemiology* **148,** 879–886.

Culverhouse, R., Klein, T., and Shannon, W. (2004). Detecting epistatic interactions contributing to quantitative traits. *Genetic Epidemiology* **27,** 141–152.

Dias, C. and Krzanowski, W. (2003). Model selection and cross validation in additive main effect and multiplicative interaction models. *Crop Science* **43,** 865–873.

Dias, C. T. d. S. and Krzanowski, W. J. (2006). Choosing components in the additive main effect and multiplicative interaction (AMMI) models. *Scientia Agricola* **63,** 169–175.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1,** 211–218.

Evans, D., Marchini, J., Morris, A., and Cardon, L. (2006). Two-stage two-locus models in genome-wide association. *PLoS Genetics* **2,** e157.

Fan, R., Albert, P., and Schisterman, E. (2012). A discussion of gene–gene and gene–environment interactions and longitudinal genetic analysis of complex traits. *Statistics in Medicine* **31,** 2565–2568.

Fan, R., Zhang, Y., Albert, P. S., Liu, A., Wang, Y., and Xiong, M. (2012). Longitudinal association analysis of quantitative traits. *Genetic Epidemiology* **36,** 856–869.

Fisher, R. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh.

Forkman, J. and Piepho, H.-P. (2014). Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. *Biometrics* doi: 10.1111/biom.12162.

Fraley, C., Raftery, A., Murphy, T., and Scrucca, L. mclust Version 4 for R: Normal mixture modeling for model-based clustering, classiffication, and density estimation. Technical report, Department of Statistics, University of Washington.

Franklin, S. S., Khan, S. A., Wong, N. D., Larson, M. G., and Levy, D. (1999). Is pulse pressure useful in predicting risk for coronary heart disease? The Framingham Heart Study. *Circulation* **100,** 354–360.

Freeman, G. et al. (1973). Statistical methods for the analysis of genotype-environment interactions. *Heredity* **31,** 339–354.

Gabriel, K. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58,** 453–467.

Gabriel, K. and Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* **21,** 489–498.

Gallant, A. (2009). *Nonlinear Statistical Models*. Wiley, New York.

Gao, X., Starmer, J., and Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology* **32,** 361–369.

Gauch, H. and Zobel, R. (1988). Predictive and postdictive success of statistical analyses of yield trials. *Theoretical and Applied Genetics* **76,** 1–10.

Gauch Jr, H. (1988). Model selection and validation for yield trials with interaction. *Biometrics* **44,** 705–715.

Gauch Jr., H. (1992). *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*. Elsevier, Amsterdam, The Netherlands.

Gollob, H. (1968). A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika* **33,** 73–115.

Gumpertz, M. L. and Pantula, S. G. (1992). Nonlinear regression with variance components. *Journal of the American Statistical Association* **87,** 201–209.

Hanumara, R. and Thompson Jr, W. (1968). Percentage points of the extreme roots of a wishart matrix. *Biometrika* **55,** 505–512.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of Royal Statistical Society: Series C (Applied Statistics)* **28,** 100–108.

Hirsch, J., Diez-Roux, A., Moore, K., Evenson, K., and Rodriguez, D. (2014). Change in walking and body mass index following residential relocation: The Multi-Ethnic Study of Atherosclerosis. *American Journal of Public Health* **104,** e49–e56.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85,** 809–822.

Hubert, H. B., Feinleib, M., McNamara, P. M., and Castelli, W. P. (1983). Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the framingham heart study. *Circulation* **67,** 968–977.

Hwang, H. and Takane, Y. (2004). A multivariate reduced-rank growth curve model with unbalanced data. *Psychometrika* **69,** 65–79.

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

Johnson, D. and Graybill, F. (1972a). An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association* **67,** 862–868.

Johnson, D. and Graybill, F. (1972b). Estimation of σ^2 in a two-way classification model with interaction. *Journal of the American Statistical Association* **67,** 388–394.

Jung, J., Sun, B., Kwon, D., Koller, D., and Foroud, T. (2009). Allelic-based gene-gene interaction associated with quantitative traits. *Genetic Epidemiology* **33,** 332–343.

Khoury, M. and Wacholder, S. (2009). Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies–challenges and opportunities. *American Journal of Epidemiology* **169,** 227–230.

Knutson, M. and Wessling-Resnick, M. (2003). Iron metabolism in the reticuloendothelial system. *Critical Reviews in Biochemistry and Molecular Biology* **38,** 61–88.

Ko, Y., Chudhuri, P., Park, S., Vokonas, P., and Mukherjee, B. (2013). Novel likelihood ratio tests for screening gene-gene and gene-environment interactions with unbalanced repeated-measures data. *Genetic Epidemiology* **37,** 581–591.

Kooperberg, C. and LeBlanc, M. (2008). Increasing the power of identifying gene $\times$ gene interactions in genome-wide association studies. *Genetic Epidemiology* **32,** 255–263.

Kraft, P., Yen, Y., Stram, D., Morrison, J., and Gauderman, W. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity* **63,** 111–119.

Kwong, W. T., Friello, P., and Semba, R. D. (2004). Interactions between iron deficiency and lead poisoning: epidemiology and pathogenesis. *Science of The Total Environment* **330,** 21–37.

Lazarsfeld, P. and Henry, N. (1968). *Latent Structure Analysis.* Houghton Mifflin, Boston.

Lee, E. Statistical analysis software for multiplicative interaction models. Unpublished Doctoral Dissertation. Kansas State University.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed model sby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61,** 381–400.

Lindstrom, M. and Bates, D. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46,** 673–687.

Maes, H. H., Neale, M. C., and Eaves, L. J. (1997). Genetic and environmental factors in relative body weight and human adiposity. *Behavior Genetics* **27,** 325–351.

Maity, A., Carroll, R., Mammen, E., and Chatterjee, N. (2009). Testing in semiparametric models with interaction, with applications to gene–environment interactions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71,** 75–96.

Malzahn, D., Schillert, A., Müller, M., and Bickeböller, H. (2010). The longitudinal nonparametric test as a new tool to explore gene-gene and gene-time effects in cohorts. *Genetic Epidemiology* **34,** 469–478.

Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association* **56,** 878–888.

Mandel, J. (1971). A new analysis of variance model for non-additive data. *Technometrics* **13,** 1–18.

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

Milliken, G. and Johnson, D. (1989). *Analysis of Messy Data, Volume II: Nonreplicated Experiments*. Van Nostrand Reinhold, New York.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., and Marks, J. S. (2003). Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *Journal of the American Medical Association* **289,** 76–79.

Moore, K., Diez-Roux, A., Auchincloss, A., Evenson, K., Kaufman, J., Mujahid, M., and Williams, K. (2013). Home and work neighbourhood environments in relation to body mass index: the Multi-Ethnic Study of Atherosclerosis (MESA). *Journal of Epidemiology and Community Health* **67,** 846–853.

Mordukhovich, I., Wilker, E., Suh, H., Wright, R., Sparrow, D., Vokonas, P., and Schwartz, J. (2009). Black carbon exposure, oxidative stress genes, and blood pressure in a repeated-measures study. *Environmental Health Perspectives* **117,** 1767–1772.

Moreno-Macias, H., Romieu, I., London, S. J., and Laird, N. M. (2010). Gene-environment interaction tests for family studies with quantitative phenotypes: A review and extension to longitudinal measures. *Human Genomics* **4,** 302–326.

Mujahid, M.S.and Diez-Roux, A., Shen, M., Gowda, D., Sanchez, B., Shea, S., Jacobs Jr., D., and Jackson, S. (2008). Relation between neighborhood environments and obesity in the Multi-Ethnic Study of Atherosclerosis. *American Journal of Epidemiology* **167,** 1349–1357.

Mukherjee, B., Ahn, J., Gruber, S. B., and Chatterjee, N. (2012). Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *American Journal of Epidemiology* **175,** 177–190.

Mukherjee, B., Ahn, J., Gruber, S. B., Rennert, G., Moreno, V., and Chatterjee, N. (2008). Tests for gene-environment interaction from case-control data: a novel study of type i error, power and designs. *Genetic Epidemiology* **32,** 615–626.

Mukherjee, B. and Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case–control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64,** 685–694.

Mukherjee, B., Ko, Y., VanderWeele, T., Roy, A., Park, S., and Chen, J. (2012). Principal interactions analysis for repeated measures data: application to gene–gene and gene–environment interactions. *Statistics in Medicine* **31,** 2531–2551.

Murcray, C., Lewinger, J., and Gauderman, W. (2009). Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology* **169,** 219–226.

Nettleton, J. A., Steffen, L. M., Mayer-Davis, E. J., Jenny, N. S., Jiang, R., Herrington, D. M., and Jacobs, D. R. (2006). Dietary patterns are associated with biochemical markers of inflammation and endothelial activation in the Multi-Ethnic Study of Atherosclerosis (MESA). *American Journal of Clinical Nutrition* **83,** 1369–1379.

Nettleton, J. A., Steffen, L. M., Schulze, M. B., Jenny, N. S., Barr, R. G., Bertoni, A. G., and Jacobs, D. R. (2007). Associations between markers of subclinical atherosclerosis and dietary patterns derived by principal components analysis and reduced rank regression in the Multi-Ethnic Study of Atherosclerosis (MESA). *American Journal of Clinical Nutrition* **85,** 1615–1625.

Nocedal, J. and Wright, S. (1999). *Numerical Optimization.* Springer Verlag, New York.

Oman, S. (1991). Multiplicative effects in mixed model analysis of variance. *Biometrika* **78,** 729–739.

Onyike, C. U., Crum, R. M., Lee, H. B., Lyketsos, C. G., and Eaton, W. W. (2003). Is obesity associated with major depression? Results from the third National Health and Nutrition Examination Survey. *American Journal of Epidemiology* **158,** 1139–1147.

Park, S., Elmarsafawy, S., Mukherjee, B., Spiro III, A., Vokonas, P., Nie, H., Weisskopf, M., Schwartz, J., and Hu, H. (2010). Cumulative lead exposure and age-related hearing loss: The VA Normative Aging Study. *Hearing Research* **269,** 48–55.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58,** 545–554.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* **2,** e190.

Perlstein, T., Weuve, J., Schwartz, J., Sparrow, D., Wright, R., Litonjua, A., Nie, H., and Hu, H. (2007). Cumulative community-level lead exposure and pulse pressure: the Normative Aging Study. *Environmental Health Perspectives* **115,** 1696.

Piepho, H. (1994). On tests for interaction in a nonreplicated two-way layout. *Australian & New Zealand Journal of Statistics* **36,** 363–369.

Piepho, H. (1995). Robustness of statistical tests for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trials. *Theoretical and Applied Genetics* **90,** 438–443.

Piepho, H. (1997). Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics* **53,** 761–766.

Piepho, H.-P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics* **97,** 195–201.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38,** 904–909.

Pritchard, J. K. and Donnelly, P. (2001). Case–control studies of association in structured or admixed populations. *Theoretical Population Biology* **60,** 227–237.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Radloff, L. S. (1977). The CES-D scale a self-report depression scale for research in the general population. *Applied Psychological Measurement* **1,** 385–401.

Schaffrath Rosario, A., Wellmann, J., Heid, I. M., and Wichmann, H.-E. (2006). Radon epidemiology: continuous and categorical trend estimators when the exposure distribution is skewed and outliers may be present. *Journal of Toxicology and Environmental Health* **69,** 681–700.

Siahpush, S. H., Vaughan, T. L., Lampe, J. N., Freeman, R., Lewis, S., Odze, R. D., Blount, P. L., Ayub, K., Rabinovitch, P. S., Reid, B. J., et al. (2007). Longitudinal study of insulin-like growth factor, insulin-like growth factor binding protein-3, and their polymorphisms: risk of neoplastic progression in Barrett's esophagus. *Cancer Epidemiology Biomarkers & Prevention* **16,** 2387–2395.

Smith, A., Cullis, B., and Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* **57,** 1138–1147.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42,** 937–948.

Srebro, N. and Jaakkola, T. (2003). Weighted low-rank approximations. In *International Conference on Machine Leaning*, volume 20, page 720.

Stunkard, A. J., Foch, T. T., and Hrubec, Z. (1986). A twin study of human obesity. *Journal of the American Medical Association* **256,** 51–54.

Tukey, J. (1949). One degree of freedom for non-additivity. *Biometrics* **5,** 232–242.

Tukey, J. (1962). The future of data analysis. *Annals of Mathematical Statistics* **33,** 1–67.

VanderWeele, T. J., Mukherjee, B., and Chen, J. (2012). Sensitivity analysis for interactions under unmeasured confounding. *Statistics in Medicine* **31,** 2552–2564.

Vonesh, E. and Carter, R. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics* **48,** 1–17.

Vonesh, E., Wang, H., and Majumdar, D. (2001). Generalized least squares, Taylor series linearization and Fisher's scoring in multivariate nonlinear regression. *Journal of the American Statistical Association* **96,** 282–291.

Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D., and Montana, G. (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage* **60,** 700–716.

Wadden, T. A., Webb, V. L., Moran, C. H., and Bailer, B. A. (2012). Lifestyle modification for obesity new developments in diet, physical activity, and behavior therapy. *Circulation* **125,** 1157–1170.

Wang, Y., Huang, C., Fang, Y., Yang, Q., and Li, R. (2012). Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **61,** 1–24.

Wentzell, P. D., Andrews, D. T., Hamilton, D. C., Faber, K., and Kowalski, B. R. (1997). Maximum likelihood principal component analysis. *Journal of Chemometrics* **11,** 339–366.

Wong, M., Day, N., Luan, J., Chan, K., and Wareham, N. (2003). The detection of gene–environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *International Journal of Epidemiology* **32,** 51–57.

Xu, S. (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63,** 513–521.

Zhang, A., Park, S., Wright, R., Weisskopf, M., Mukherjee, B., Nie, H., Sparrow, D., and Hu, H. (2010). *HFE H63D* polymorphism as a modifier of the effect of cumulative lead exposure on pulse pressure: the Normative Aging Study. *Environmental Health Perspectives* **118,** 1261–1266.

Zhang, H. (1997). Multivariate adaptive splines for analysis of longitudinal data. *Journal of Computational and Graphical Statistics* **6,** 74–91.

Zhang, H. (2004). Mixed effects multivariate adaptive splines model for the analysis of longitudinal and growth curve data. *Statistical Methods in Medical Research* **13,** 63–82.

Zhang, S., Zhu, X., and Zhao, H. (2003). On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic Epidemiology* **24,** 44–56.

Zobel, R., Wright, M., and Gauch, H. (1988). Statistical analysis of a yield trial. *Agronomy Journal* **80,** 388–393.