ABSTRACT

FRANCK, CHRISTOPHER. Latent Group-Based Interaction Effects in Unreplicated Factorial Experiments. (Under the direction of Dr. Jason Osborne.)

Statistical models which are used to analyze unreplicated factorial experiments typically do not include interaction effects between the experimental treatments. This is because the standard tests which determine whether or not interaction is present do not have an appropriate error term. Many authors have developed models and statistical tests based on restricted forms of interaction to address non-additivity for these unreplicated experiments. Our strategy is to instead assume that levels of one of the factors belong to some smaller number of latent groups or environments. For example, consider a randomized complete block design (RCBD) in an agricultural experiment. If some of the blocks come from a poor environment there may not be treatment effects for these blocks. Perhaps the rest of the blocks come from a second environment, where treatment effects are pronounced. Our model is able to capture this type of non-additivity since we allow these groups to interact with the other treatments, but we assume that treatment effects are additive within group. We focus primarily on the detection of this grouping structure in a one versus two group case. We also consider point and interval estimation of the simple effect of the treatments in each group. The randomized complete block design (RCBD) is used to describe our methods which include both an all configurations approach as well as a Bayes Factor approach to diagnosis of this latent group-based interaction.

© Copyright 2010 by Christopher Franck

All Rights Reserved

Latent Group-Based Interaction Effects in Unreplicated Factorial Experiments

by
Christopher Franck

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2010

APPROVED BY:

Dr. Sujit Ghosh	Dr. David Dickey
Dr. Jeff Thompson	Dr. Jason Osborne Chair of Advisory Committee

DEDICATION

This work is dedicated with love to Helen Franck.

BIOGRAPHY

Chris Franck was born in Charlottesville, VA in 1982. He graduated from Charlottesville High School in 2001. He then attended Virginia Tech, graduating in 2005 with a B.S. in Statistics. He then attended North Carolina State University in 2005. He got married in the Summer of 2007. Chris successfully defended this PhD thesis during the Fall of 2010.

ACKNOWLEDGEMENTS

I would like to thank everyone who made this work possible. My advisor Jason Osborne has been a tremendous mentor and friend to me throughout my time in graduate school. David Dickey has always had a kind ear and a wise suggestion. Sujit Ghosh has challenged me to become a better statistician during every phase of my graduate education. Roger Woodard not only taught me how to teach but also helped me realize I love teaching. Pam Arroway has been a source of support whenever things did not go exactly as planned. Golde Holtzman set me on this path with his practical demonstration of the power of statistics. Bill Hunt has been strong mentor, advocate and friend.

I would not have been able to complete this work without the steadfast support of my loved ones. I could not have been any luckier to have Jim as my older brother and closest friend. I also have relied on the love and support of my parents Wes and Charlene since long before I began this process. And I must especially recognize my extraordinary wife Helen, who stuck by me during both the tough times and the good times during the last several years. Without their support this work would not have been possible.

TABLE OF CONTENTS

List of Tables . . . . . **vi**

List of Figures . . . . . **xii**

Chapter 1 Introduction **1**
1.1 Background 1
1.2 Non-additivity in Unreplicated Experiments 6
1.3 Component Number Selection for Finite Mixture Models 11

Chapter 2 All Configurations Approaches **15**
2.1 Types of group-based interaction 23
2.2 Statistical inference in the all configurations framework 26
2.3 Illustrative example - Sweetpotatoes 33
2.4 Illustrative example - Dog Lymphoma (CGH) 37
2.5 Maximum interaction F test for large data sets. 41
2.6 Simulation study 43

Chapter 3 Bayes Factor approach **60**
3.1 Approximation of the marginal distribution of the data. 62
3.2 Implementation of EM Algorithm for Two Component Model 68
3.3 Illustrative example 73
3.4 Simulation Study 75

Chapter 4 Concluding Remarks **85**
4.1 Future research. 87

References **89**

Appendices **91**
Appendix A Tables and Figures 92
A.1 Tables 92
A.2 Figures 101
Appendix B Computer Codes 120
B.1 SAS code 120
B.2 R code 139

LIST OF TABLES

Table 2.1 Degrees of freedom for model (1.2) 16
Table 2.2 Degrees of freedom for model (2.1) 16
Table 2.3 Form of each of the intervals under investigation. 31
Table 2.4 ANOVA table based on the replicated analysis 34
Table 2.5 Comparison of rejection rates for Tukey's test and ACMIF. Table counts refer to frequency of rejection of null hypothesis at $\alpha = 0.05$. Cell percentages are in parentheses. Non-zero off-diagonal counts suggest that nonadditivity is present which is unique to each method. 38
Table 2.6 Non-additivity labels and treatment vectors for each group. The labels describe the above treatment means for the rest of Chapter 3. 44
Table 2.7 Simulation results for additive data. First three columns describe experimental setup. Empirical rejection rates are given for both ACMIF and Tukey's test for non-additivity at $\alpha = 0.05$. Results are based on 1,000 MC data sets. The * denotes a run significantly different from actual error rate $\alpha = 0.05$ (unadjusted p-value=0.00207). Standard error for all proportions presented less than 0.0159. . . . 46
Table 2.8 MSE results for four types of estimators for additive data. First three columns describe experimental setup. Other columns report mean squared error for the various estimators. Results are based on 1,000 MC data sets. 47
Table 2.9 Squared bias results for four types of estimators for additive data. First three columns describe experimental setup. Other columns report squared bias for the various estimators. Results are based on 1,000 MC data sets. 48
Table 2.10 Variance results for four types of estimators for additive data. First three columns describe experimental setup. Other columns report variance for the various estimators. Note $variance + Bias^2 = MSE$. Results are based on 1,000 MC data sets. 49
Table 2.11 Coverage proportions for four types of intervals for additive data. First three columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Results are based on 1,000 MC data sets. Standard error for all proportions presented less than 0.0159. 51

Table 2.12 Simulation results for various latent group non-additive scenarios. Three treatments and seven blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF different from Tukey's test for all cases when estimates not equal (p-values < 0.0001). Standard error for all proportions presented less than 0.0159. 54
Table 2.13 MSE results for four types of estimators under various latent group non-additive scenarios. Three treatments and seven blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators. 55
Table 2.14 Squared bias results for four types of estimators under various latent group non-additive scenarios. Three treatments and seven blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators. 56
Table 2.15 Variance results for four types of estimators under various latent group non-additive scenarios. Three treatments and seven blocks were used. First two columns describe experimental setup. Other columns report variance of the various estimators. 57
Table 2.16 Coverage proportions for four types of intervals under various latent group non-additive scenarios. Three treatments and seven blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159. 58

Table 3.1 Data labels and corresponding treatment means for BF simulation study. 79
Table 3.2 Simulation distributional results: relative frequencies of BF approximation appear in columns 2-5. Median BF appears in column 6. Tukey's test rejection rates ($\alpha = 0.05$) in additive case appear in the last column. 80
Table 3.3 Simulation distributional results: relative frequencies of BF approximation appear in columns 2-5. Median BF appears in column 6. Tukey's test rejection rates ($\alpha = 0.05$) in accordion case appear in the last column. 81

Table A.1 Simulation results for various latent group non-additive scenarios. Two treatments and four blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF is different from Tukey's test for all cases (p-values < 0.0001). Standard error for all proportions presented less than 0.0159. 93
Table A.2 MSE results for four types of estimators under various latent group non-additive scenarios. Two treatments and four blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators. 93
Table A.3 Squared bias results for four types of estimators under various latent group non-additive scenarios. Two treatments and four blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators. 94
Table A.4 Variance results for four types of estimators under various latent group non-additive scenarios. Two treatments and four blocks were used. First two columns describe experimental setup. Other columns report variance of the various estimators. 94
Table A.5 Coverage proportions for four types of intervals under various latent group non-additive scenarios. Two treatments and four blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159. 95
Table A.6 Simulation results for various latent group non-additive scenarios. Two treatments and seven blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF different from Tukey's test for all cases when estimates not equal (p-values < 0.0001). Standard error for all proportions presented less than 0.0159. 95
Table A.7 MSE results for four types of estimators under various latent group non-additive scenarios. Two treatments and seven blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators. 96
Table A.8 Squared bias results for four types of estimators under various latent group non-additive scenarios. Two treatments and seven blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators. 96

Table A.9 Variance results for four types of estimators under various latent group non-additive scenarios. Two treatments and seven blocks were used. First two columns describe experimental setup. Other columns report variance of the various estimators. 97
Table A.10 Coverage proportions for four types of intervals under various latent group non-additive scenarios. Two treatments and seven blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159. 97
Table A.11 Simulation results for various latent group non-additive scenarios. Two treatments and eleven blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF different from Tukey's test for all cases when estimates not equal (p-values < 0.0001). Standard error for all proportions presented less than 0.0159. 98
Table A.12 MSE results for four types of estimators under various latent group non-additive scenarios. Two treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators. 98
Table A.13 Squared bias results for four types of estimators under various latent group non-additive scenarios. Two treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators. 99
Table A.14 Variance results for four types of estimators under various latent group non-additive scenarios. Two treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report variance of the various estimators. 99
Table A.15 Coverage proportions for four types of intervals under various latent group non-additive scenarios. Two treatments and eleven blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159. 100

Table A.16 Simulation results for various latent group non-additive scenarios. Three treatments and four blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF different from Tukey's test for all cases (p-values < 0.0001). Standard error for all proportions presented less than 0.0159. 101
Table A.17 MSE results for four types of estimators under various latent group non-additive scenarios. Three treatments and four blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators. 102
Table A.18 Squared bias results for four types of estimators under various latent group non-additive scenarios. Three treatments and four blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators. 103
Table A.19 Variance results for four types of estimators under various latent group non-additive scenarios. Three treatments and four blocks were used. First two columns describe experimental setup. Other columns report variance the various estimators. 104
Table A.20 Coverage proportions for four types of intervals under various latent group non-additive scenarios. Three treatments and four blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159. 105
Table A.21 Simulation results for various latent group non-additive scenarios. Three treatments and eleven blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF different from Tukey's test for all cases when estimates not equal (p-values < 0.0001). Standard error for all proportions presented less than 0.0159. 106
Table A.22 MSE results for four types of estimators under various latent group non-additive scenarios. Three treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators. 107
Table A.23 Squared bias results for four types of estimators under various latent group non-additive scenarios. Three treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators. 108

Table A.24 Variance results for four types of estimators under various latent group non-additive scenarios. Three treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report variance of the various estimators. 109
Table A.25 Coverage proportions for four types of intervals under various latent group non-additive scenarios. Three treatments and eleven blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159. 110
Table A.26 Simulation distributional results for BF approximation and Tukey's test rejection rates in cancellatory case. 110
Table A.27 Simulation distributional results for BF approximation and Tukey's test rejection rates in inert versus active case. 111
Table A.28 Simulation distributional results for BF approximation and Tukey's test rejection rates in average versus active case. 111

LIST OF FIGURES

Figure 2.1 1000 MC Minimum p-values from additive data with four blocks and three treatments. The histogram of these p-values does not conform to the Beta(1,c) overlay because of the dependency between configurations. 21

Figure 2.2 Visual examples of some types of group-based interactions for two treatments. 25

Figure 2.3 Visual examples of some types of group-based interactions for two treatments. 27

Figure 2.4 Visual examples of some types of group-based interactions for three treatments. 28

Figure 2.5 Mean glucose at each combination of storage time and processing stage. 36

Figure 2.6 Histogram of Tukey p-values for each of the 5899 probes for chromosome of interest. The non-uniformity of this plot suggests that non-additivity may be present in these data. 39

Figure 2.7 Data from probe 5963 shows non-additivity according to the ACMIF test but not the Tukey test. Notice the similarity between this pattern and the cancellatory pattern in Section 2.1. 40

Figure 2.8 Data from probe 1280 shows non-additivity according to the ACMIF test but not the Tukey test. Notice the similarity between this pattern and the inert versus active pattern in Section 2.1. 41

Figure 2.9 Power plots for average versus active data for both methods. $a = 2$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis. 59

Figure 3.1 Interaction plot. 73

Figure 3.2 Interaction plot coded by most likely classification based on a two component model. 74

Figure 3.3 Histograms of univariate data in two component mixtures. First component is centered at zero and second component is centered at 60. 79

Figure 3.4 Histograms of $\log_{10}(BF)$ Bayes factors in additive data for $N = 300$ MC data sets and various levels of σ^2. 80

Figure 3.5 Histograms of $\log_{10}(BF)$ BF in accordion data for $N = 300$ MC data sets and various levels of σ^2. 81

Figure A.1 Plot of Tukey p-values vs. adjusted ACMIF p-values for lymphoma example 102

Figure A.2 Power plots for inert versus active data for both methods. $a = 2$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis. 112
Figure A.3 Power plots for cancellatory data for both methods. $a = 2$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis. 113
Figure A.4 Power plots for inert versus active data for both methods. $a = 3$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis. 114
Figure A.5 Power plots for average versus active data for both methods. $a = 3$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis. 115
Figure A.6 Power plots for cancellatory data for both methods. $a = 3$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis. 116
Figure A.7 Histograms of $log_{10}(BF)$ in cancellatory data for $N = 300$ MC data sets and various levels of σ^2. 117
Figure A.8 Histograms of $log_{10}(BF)$ in average versus active data for $N = 300$ MC data sets and various levels of σ^2. 118
Figure A.9 Histograms of $log_{10}(BF)$ inert versus active data for $N = 300$ MC data sets and various levels of σ^2. 119

Chapter 1

Introduction

1.1 Background

Two way unreplicated factorial designs are sufficient for the purpose of investigating the main effects of experimental treatments on some response. These designs are favorable when economical or physical constraints prohibit the researcher from replicating treatment combinations on multiple experimental units. One well known drawback of this approach is that the standard F test for interaction does not have an appropriate error term when the factorial combinations have only one replicate apiece. If interaction among the predictor variables exists then the additive model does not capture the true behavior of the response and any analysis based on it will yield biased results. Hence literature has arisen which addresses the problem of interaction in unreplicated experiments. The basic strategy for many current methods is to use a model which has some restricted form of interaction. An overview of many of these methods is included in our literature review. Instead of imposing some restriction on the form of the interaction parameters directly, we assume that levels of one of the factors arise from a smaller number of groups, and

the other factor may affect the response differently in each of these groups. We allow the groupings of one of the factor levels to interact with the other factor. For instance perhaps all varieties of some crop perform equivalently in poor soil but some of the varieties flourish more than others in fertile soil.

To illustrate the above idea using randomized complete block design (RCBD) terminology, imagine that we conduct such a study under the assumption that treatments do not interact with blocks. Upon collecting the data we begin to suspect that the blocks are possibly members of some smaller number of latent groups, or environments, and these groups interact with the treatments. We want to develop methods which can formally address whether or not such a latent structure is plausible, and if so, we want to conduct statistical inference with this structure in mind.

Unreplicated factorial experiments are conducted in many fields such as medicine. Imagine that several competing treatments are being tested on a number of patients, and some outcome of interest (such as blood pressure, cholesterol, etc.) is measured. Each patient receives each treatment in random order one time only. If there is no interaction present between patient and treatment and standard modeling assumptions hold, then this unreplicated design and an analysis based on the additive model are sufficient to characterize the behavior of the response. However, if the true nature of the treatments and patients are such that the effect of the treatments on the response is different for different patients, then interaction is present. Non-zero interaction effects present some difficulty since standard linear model theory cannot accommodate the full set of standard interaction parameters.

Suppose now that the patients in this population (and hence sample) are heterogenous in some unknown way. Perhaps the treatment affects different people differently based on some genetic trait. Some patients have this trait and the rest do not, so in this case there are two groups from which the patients arise. If the number of groups and patient membership within these groups are known *a priori* then standard linear modeling techniques will allow us to assess treatment effects, group effects, and interaction between these two factors. Additionally, the additive variability between patients within groups can be modeled. A full discussion of the merits and caveats of the group based model is included in Chapter 2.

Replication is a principle of sound experimental design for scientific experimentation. It seems unlikely that any single statistical method will arise that will be totally satisfactory to detect all types of interaction effects in unreplicated experiments. In lieu of a silver bullet method a number of methods have been proposed to detect and perform inference on restricted interaction parameter models. Many such models and methods have arisen. However the physical interpretation of many of the restricted interaction models in terms of real world phenomena is possibly non-intuitive for many cases. Methods based on the latent group restriction can be intuitive in situations where a grouping structure is sensible.

Typically neither the number of groups nor membership status is known in advance, otherwise the researcher would likely choose an appropriate design based on these features. Our challenge is to determine whether this grouping structure is present, and if so, to conduct statistical inference on the population in light of this structure.

Before elaborating on our methods we first review the basic problem and available literature more fully. Consider the balanced two-way analysis of variance (ANOVA) model for independent response variables y_{ijm} observed in a completely randomized experiment:

$$y_{ijm} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijm} \ . \tag{1.1}$$

where $i = 1, .., a > 1$ is an index for the levels of one of the two factors, $j = 1, ...b > 1$ indexes the levels of the other factor, and $m = 1, ..., n > 1$ is an index for the number of replicates for experimental units, $\epsilon_{ijm} \sim N(0, \sigma^2)$. Note α_i, β_j, and $(\alpha\beta)_{ij}$ may be either fixed or random effects depending on the nature of the study. In the fixed effect case we adopt the sum-to-zero parameterization (e.g., $\sum_{i=1}^{a} \alpha_i = 0$). For random effects we assume normality (e.g., $\alpha_i \overset{iid}{\sim} N(0, \gamma^2)$ for $i = 1, ..., a$. In this scenario all parameters are estimable and statistical inference can be performed using the two-way ANOVA technique. However, for the the special case where $m = 1$ the there is no error-variance estimate to use in the denominator for the usual F tests for model parameters.

The RCBD is a well known example of a two way unreplicated factorial design. The following additive model is commonly posited for RCBDs:

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \ . \tag{1.2}$$

where $i = 1, .., a; j = 1, ...b; \sum_{i=1}^{a} \alpha_i = 0; \beta_j \sim N(0, \sigma_B^2); \epsilon_{ij} \sim N(0, \sigma^2)$. Typically the treatment effects (denoted α_i) are considered fixed and the block effects (denoted β_j) are considered random in the RCBD experiment. We refer to (1.2) as the additive model throughout this thesis.

Choice of an unreplicated design does not preclude the possibility that the factors truly have non-zero interaction effects. Since ignoring important model terms leads to biased estimation, we are interested in investigating alternative (but necessarily more restrictive) forms of interaction effects to those described in model (1.1).

We will propose two methods to diagnose nonadditivity in unreplicated experiments. In this thesis we focus mainly on the one versus two group question although investigation of a greater number of groups is generally of interest. The first method considers all possible groupings of blocks into two groups and in doing so determines whether the optimal configuration exhibits significant group by treatment interaction while adjusting for the multiplicity of the large number of possible configurations. We present a model based on this configuration which can accommodate treatment effect, group effect, treatment by group interaction, and also additive block variability. Blocks are nested in groups in this case. Our review of restricted interaction parameter models illustrates historical developments for this problem while also demonstrating the difference between existing methods and the all configurations approach.

Our second strategy is to develop a Bayes factor which will determine if it is more plausible that the blocks belong to two latent groups or if all of the blocks belong in the same group. In the latter case the additive model is sufficient for inference. To implement this method we use finite mixture model theory and test whether or not a one component model or two component model is more plausible. Section 2.2 motivates the unique challenges associated with component number selection for finite models by giving an overview of some relevant literature.

We now review current models and methods for addressing non-additivity in unreplicated factorial experiments.

1.2 Non-additivity in Unreplicated Experiments

Perhaps the first formal investigation of non-additivity in unreplicated experiments is the one degree of freedom test for non-additivity proposed by Tukey (1949). In the original paper no model is explicitly defined, however it has been pointed out (Alin and Kurt, 2006) that the test is based on the following model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \nu\alpha_i\beta_j + \epsilon_{ij} \ . \tag{1.3}$$

It is clear that this model includes a single ν parameter in addition to those in (1.2). A test of non-additivity can then be defined in terms of hypotheses $H_0 : \nu = 0$ versus $H_1 : \nu \neq 0$. The ANOVA table for this test is similar to the ANOVA table for the additive model but with a single degree of freedom entry used for ν. As a result there is one less degree of freedom for the error term. This test is easy to implement with standard statistical software, however the form of the model is restrictive and the interpretation of the interaction parameter in terms of physical phenomena is not immediately intuitive. While Tukey's suggestion of transforming non-additive data into a more additive scale using a power or log transformation is one approach to addressing interaction, a more direct approach is to model interaction explicitly. If interaction is part of the nature of the variables, then analysis on a scale which masks this interaction seems to ignore a

fundamental feature of the data. The model (1.3) will henceforth be referred to as the 'Tukey Model.' Tukey (1955) explains how to apply this method to larger experiments.

Mandel (1961) developed a generalization of Tukey's model.

$$y_{ij} = \mu + \alpha_i + \beta_j + \theta_j \alpha_i + \epsilon_{ij} \ . \tag{1.4}$$

where $i = 1, ..., a > 2$; $j = 1, ..., b$. Mandel's interaction parameters are essentially block-specific constants multiplied with the treatment main effects. This model has been called the "rows-linear model" by Alin and Kurt (2006). Instead of adding one interaction parameter as in the Tukey model, Mandel allows $(b-1)$ parameters for interaction. Note that if $a = 2$ then there are not enough degrees of freedom to accommodate this form of interaction after accounting for block, treatment, and error terms.

To see the linear expression, define $Q_j = (\theta_j + 1)$ and let $\mu_j = \mu + \beta_j$. Now for any given level of j, Mandel's model can be expressed as:

$$y_{ij} = \mu_j + \alpha_i Q_j + \epsilon_{ij} \ . \tag{1.5}$$

That is, each observation can be characterized as a block-specific intercept plus a treatment effect times a block-specific constant plus some random error.

Ten years after publishing the previous model, Mandel (1971) put forth another model for non-additive effects:

$$y_{ij} = \mu + \alpha_i + \beta_j + \eta_{ij} \ . \tag{1.6}$$

where $\eta_{ij} = \sum_{q=1}^{Q} \theta_q U_{qi} \nu_{qj.}$, $Q \leq min(a-1, b-1)$. Both the value of Q and the estimates of parameters $\theta_q, U_{qi}, \nu_{qj.}$ are determined according to features of the spectral decomposition of $\boldsymbol{DD^T}$, where $\boldsymbol{D}$ is the matrix of residuals from the additive model defined by $d_{ij} = y_{ij} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j$. The hat terms are the usual least squares estimators.

The motivation of this model is that α_i and β_j account for all additive variability in the response, and the term η_{ij} is some combination of interaction effects plus random variability (to distinguish from ϵ_{ij} which denotes only random error). The parameters $\theta_q, U_{qi}, \nu_{qj.}$ are estimated using the matrix $\boldsymbol{DD^T}$, which has $min(a, b)$ real eigenvalues with orthogonal eigenvectors. If the full decomposition is performed there will be no terms left for residual error which leads to the restriction $Q \leq min(a-1, b-1)$. Mandel recommends choosing Q so that only a few multiplicative interaction terms are included, combining the rest of the decomposition terms into residual error. Note that the estimates $\hat{\theta}_q$ for $q = 1, ..., Q$ are the eigenvalues of $\boldsymbol{DD^T}$. If there are no positive eigenvalues, then (1.6) reduces to the additive case. So testing the hypothesis $\theta_q = 0 \ \forall q = 1, ..., Q$ is analogous to testing for the presence of non-additivity by determining if any of the eigenvalues of $\boldsymbol{DD^T}$ differ significantly from zero. This eigenvalue decomposition strategy is similar to the Factor Analysis of Variance (FANOVA) technique illustrated by Gollob (1968).

The Tukey one degree of freedom test for non-additivity is designed for an experiment with a two way layout. In 1993 Zafar-Yab (1993) extended this idea to a p way layout. This method allocates $2^p - p - 1$ degrees of freedom for non-additivity, and each of these degrees of freedom is associated with a sum of squares for the linear component of interactions between each of the factors. For example in a $p = 3$ factor experiment, $2^p - p - 1 = 4$ degrees of freedom and sums of squares terms exist for the linear$\times$linear

components of $A\times B$, $A\times C$, $B\times C$, and the linear$\times$linear$\times$linear components of $A\times B\times C$. The model for the $p = 3$ case is as follows:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \nu_k + \delta_1\alpha_i\beta_j + \delta_2\alpha_i\nu_k + \delta_3\beta_j\nu_k + \delta_4\alpha_i\beta_j\nu_k \ . \tag{1.7}$$

where $\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = \sum_{k=1}^{d}\nu_k = \sum_{i=1}^{a}\sum_{j=1}^{b}\alpha_i\beta_j = \sum_{i=1}^{a}\sum_{k=1}^{d}\alpha_i\nu_k = \sum_{j=1}^{b}\sum_{k=1}^{d}\nu_k\beta_j = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{d}\alpha_i\beta_j\nu_k = 0$. Notice that if $p = 2$ this is identical to the Tukey Model.

Pardo and Pardo (2005) address the analogous case of non-additivity in log-linear models for count data. This addresses multinomial, product multinomial, and Poisson sampling schemes. They develop test statistics to diagnose non-additivity using ϕ-divergences, which include Kullback-Leibler divergence as a special case.

Johnson and Graybill (1972a) introduced a model to assess interaction in unreplicated experiments. Their model is as follows:

$$y_{ij} = \mu + \alpha_i + \beta_j + \lambda\tau_i\gamma_j + \epsilon_{ij} \ . \tag{1.8}$$

where $\sum_i^a \alpha_i = \sum_j^b \beta_j = \sum_i^a \tau_i = \sum_j^b \gamma_j = 0$, $\sum_i^a \tau_i^2 = \sum_j^b \gamma_j^2 = 1$, and $\epsilon_{ijm} \sim N(0, \sigma^2)$. Notice the second assumption need not constrain the structure of the model since λ can absorb any needed normalizing constants. The paper develops likelihood theory for this model and a likelihood ratio (LR) test of $\lambda = 0$ is also presented. In this model the main effect for treatment and block have the same maximum likelihood estimators (MLEs) as in the additive model. The MLEs for parameters τ_i and γ_j are the components of the

normalized eigenvector corresponding to the largest eigenvalues of the outer and inner products of the $a \times b$ matrix of residuals respectively. The parameter λ is estimated by the largest eigenvalue of the inner product of the $a \times b$ matrix of residuals. Johnson and Graybill develop a corresponding LR test (LRT) of $H_0 : \lambda = 0$ versus $H_1 : \lambda \neq 0$ based on their model, and suggest that it may be better than Tukey's and related tests in cases where the true form of the interaction is not a function of the treatment and block effects. This model is a specific case of (1.6) with $Q = 1$.

Estimation of error variance in non-additive models with unreplicated data is not straightforward. Nonadditivity between treatment and block effects will lead to inflated error variance estimation when the additive model(1.2) is used to estimate σ^2. One way to approach this problem is given by Johnson and Graybill (1972b). The authors decompose the uniformly minimum variance unbiased estimate (UMVUE) for σ^2 into a number of squared two by two contrasts. The authors then argue that the effects of the factors on the response is additive if and only if the mean of each of the two-by-two contrasts is zero. Some distributional theory is provided for various functions of the eigenvalues from residual matrices, which ultimately lead to a procedure for determining if these contrasts are significantly larger than we would expect under additivity. Each of these contrasts is compared to a cutoff C_α, and the probability that one of these contrasts exceeds C_α under additivity is smaller than the preassigned α level. The proposed estimator of variance is then based on only the two by two contrasts which do not differ significantly from zero. If no contrasts exceed C_α then their estimate of variance is equivalent to the UMVUE in the additive model. Hirotsu (1982) uses a similar method to determine the pattern of interaction effects in two-way layouts for both replicated and unreplicated experiments.

Barker et al. (2009) develop the orthogonal interaction model (OIM) for unreplicated experiments. Rather than restricting the form of the interaction to some type of multiplicative setup, interactions and main effects are constrained to be orthogonal. This model may be used in two-way factorial designs whether or not replication is present. Starting with the original two-way model (1.1), the additional constraints $\sum_{i=1}^{a} \alpha_i(\alpha\beta)_{ij} = 0 \forall i$, and $\sum_{j=1}^{b} \beta_j(\alpha\beta)_{ij} = 0 \forall j$ are imposed which implies that main effects and interactions are orthogonal. Barker develops a remeasurement process in order to assign degrees of freedom for the purpose of testing for effects.

The common strategy in the above cases is to develop a restricted form of interaction model. The various models retain main effects and borrow degrees of freedom from the error term. Over time the models have escalated in complexity. In practice one may not know which of the available methods most accurately captures interaction present in a given data set. In the literature, competing methods are generally compared in terms of statistical power for various forms of non-additivity.

1.3 Component Number Selection for Finite Mixture Models

Our proposed use of Bayes Factor to detect group based additivity is equivalent to choosing the number of components in the context of finite mixture models. Selecting the correct number of components in a finite mixture is a model selection problem. The goal is to choose the minimum number of components needed to sufficiently describe the data. This problem is different from other model selection problems because mixture models have a somewhat unusual parameter space associated with them. Consider a two

component multivariate mixture:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \omega f_1(\boldsymbol{y}|\boldsymbol{\theta}_1) + (1-\omega) f_2(\boldsymbol{y}|\boldsymbol{\theta}_2) \ . \tag{1.9}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $0 \leq \omega \leq 1$. If $\omega = \{0, 1\}$ or $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, then the above expression reduces to a one component model. It is clear that ω can be defined on the boundary of its parameter space, which is problematic for a variety of standard statistical procedures. For instance the standard asymptotic properties of the likelihood ratio test (LRT) statistic require that the parameter space contain an open set, and the true parameter be an interior point of this set (see e.g. Casella and Berger (2002)). Since this condition does not hold the LRT statistic does not enjoy its usual asymptotic properties under the null hypothesis. In the univariate case Ghosh and Sen (1984) derived the asymptotic distribution of the LRT statistic in the $g = 1$ versus $g = 2$ case under a separation condition $|\mu_2 - \mu_1| > \epsilon > 0$, where μ_1 and μ_2 are the location parameters for each component in the mixture.

A more recent paper by Garel (1996) allows the relaxation of this condition in the univariate case. This paper also focuses on the $d = 1$ versus $d = 2$ problem due to the adequacy of available theoretical results in this case. Garel notes that testing a difference of more than one component requires nonstationary random fields for which there is a comparative dearth of available results. This technical paper highlights the difficult theory needed for this problem.

Akaike's (1974) information criterion (AIC) and the Bayesian information criterion (BIC) (Schwartz, 1978) are two popular methods used for model selection. Each criterion mea-

sures the likelihood achieved by a candidate model with a penalty attached for the number of parameters in that model. Both criteria have been used by researchers for the component number selection problem, but Titterington et al. (1985) point out that the regularity conditions for these criteria are also violated since the parameter ω can exist on the edge of the parameter space. Hence, the asymptotic justification for these two criteria is invalid, though both have been used for our purpose in scientific literature.

Bayesian methodologies have been developed to address this issue as well. The deviance information criterion (DIC) was first developed by Spiegelhalter et al. (2002). DIC was designed in a general framework to compare hierarchical models when the number of parameters is unknown. Celeux et al. (2006) focus the potential of the DIC on estimating the number of components in mixture models. Since various model specifications may be of interest in this context, Celeux et al. (2006) explore 8 different parameterizations of DIC in the paper.

Another Bayesian approach to choosing the component number is the reversible jump Markov chain Monte Carlo (RJMCMC) method (Richardson and Green, 1997). RJMCMC is a fully Bayesian method which models component number, mixing proportions, and component parameters simultaneously. The method iterates through six move types. The first four of these moves involve updating component parameters and mixing weights, while the final two steps allow splitting or merging of components and adding or removing components. The word "jump" is used to describe the method's ability to change component number, and it is "reversible" since it can both increase and decrease the number of components. The method is broadly described by the authors as a "random sweep Metropolis-Hastings method adapted for general state spaces."

These and several other methods have been developed to address the component number estimation problem in finite mixture models. Despite the number of methods proposed, no single gold standard approach has emerged. This is likely due to many reasons including the unusual parameters spaces of the models, difficulty in theoretical development, and lack of strong empirical results across the methods. Additionally many methods have not seen literature development in the multivariate case that is of interest here. In light of the current state of the literature our focus is to develop a form of Bayes Factor(BF) and assess its empirical performance in our scenario.

Chapter 2

All Configurations Approaches

Our methods for assessing non additivity diverge from many existing techniques. We focus on determining if the levels of one of the predictors fall into some smaller number of latent groups. If this latent grouping structure exists and is known in advance the following model would be appropriate:

$$y_{ij(k)} = \mu + \alpha_i + \xi_k + (\alpha\xi)_{ik} + \beta_{j(k)} + \epsilon_{ij(k)} \ . \tag{2.1}$$

where $i = 1, .., a; j = 1, ...b; k = 1, ..., g; g < b$. The treatment effects are denoted α_i, the group effects are denoted ξ_k, the treatment by group interaction parameters are denoted $(\alpha\xi)_{ik}$, the block effects are denoted $\beta_{j(k)}$, and the error terms are denoted $\epsilon_{ij(k)}$. The following constraints $\sum_{i=1}^{a} \alpha_i = \sum_{k=1}^{g} \xi_k = \sum_{i=1}^{a} (\alpha\xi)_{ik} = \sum_{k=1}^{g} (\alpha\xi)_{ik} = 0$ hold. Also $\beta_{j(k)} \sim N(0, \sigma_B^2)$; $\epsilon_{ij(k)} \sim N(0, \sigma^2)$. Model (2.1) accommodates the full set of treatment by group interaction parameters instead of some restricted form of treatment by block interaction. Note that we treat the group effects as fixed but they may be considered random as appropriate. Also note the block effects $\beta_{j(k)}$ are now nested in group. Hence

Table 2.1: Degrees of freedom for model (1.2)

Source	df
Treatment	a-1
Block	b-1
Error	(a-1)(b-1)
Total	ab - 1

Table 2.2: Degrees of freedom for model (2.1)

Source	df
Treatment	a-1
Group	g-1
Treatment×Group	(a-1)(g-1)
Block(Group)	b-g
Error	(a-1)(b-g)
Total	ab - 1

we maintain the traditional notion that blocks are heterogeneous but add the assumption that they belong to some smaller (but possibly unknown) number of groups. Within each group the treatment effects are constant, but the treatment effects may be different in the different groups. We can now assess the interaction between these groups and the treatment variable. If treatment by group interaction is present then we are interested in assessing the simple effect of the treatments separately within each group.

Notice that the above model (2.1) can be implemented for any $g < b$. The available degrees of freedom for this model are summarized in Table 2.2.

Table 2.2 illustrates that model (2.1) has the same number of total degrees of freedom as the additive model summarized in Table (2.1). Using the least squares criterion it is

apparent that the group model partitions the block and error sum of squares from the additive model into the group, treatment by group, block(group), and error terms above. Notice that the degree of freedom entries for these terms sum to $a(b-1)$ which equals the sum of the block and error df from the ANOVA table from the additive model (1.2).

One important feature of the group-based model analysis is that the treatment by group interaction test is equivalent to the F test for comparing nested models. In this case we are comparing the group-based model to the additive model. Regard model (2.1) as the full model and (1.2) as the reduced model and define the F statistic in the usual way:

$$F = \frac{(SS(E)_{Reduced} - SS(E)_{Full})/(dfE_{Full} - dfE_{Reduced})}{MS(E)_{Full}} \ . \tag{2.2}$$

where

$$F \overset{H_0}{\sim} F_{dfE_{Full}-dfE_{Reduced},dfE_{Full}} \ .$$

Model (2.1) partitions the error sum of squares from the additive model into two parts: $SS(E)_{Reduced} = SS(E)_{Full} + SS(Treatment \times Group)$ so the difference in sum of squares in the numerator of (2.2) $(SS(E)_{Reduced} - SS(E)_{Full}) = SS(Treatment \times Group)$ where $(Treatment \times Group)$ stands for treatment by group interaction. Also notice the treatment by group term has $(a-1)(g-1)$ degrees of freedom, and the difference in degrees of freedom in the numerator of (2.2) $(dfE_{Full} - dfE_{Reduced}) = (a-1)(b-1) - (a-1)(b-g) = (a-1)(g-1)$, which is also the number of degrees of freedom from the treatment by group term in model (2.1). Finally, the proper denominator for the treatment by group term is $MS(E)_{Full}$. Hence a test of treatment by group interaction in the group model is equivalent to the standard full versus reduced model F test for comparing nested linear models.

Before implementing the above analysis we want to detect whether or not the described grouping structure plausibly exists. As such we are interested in methods which can help us decide whether the simpler model (1.2) is sufficient, or whether we should favor model (2.1). This question is complicated by the fact that block membership in the groups is unknown and must be treated as a latent variable. Hence two questions arise. First, is it plausible that such a grouping structure exists? If so, how does one conduct inference in light of these features?

One of the methods we have investigated for this purpose is called the all configurations maximum interaction F test (ACMIF). The "all configurations" label implies that we will consider every single possible grouping of blocks for some number of groups. We will use the maximum of all of the interaction F tests among all possible groupings as a criterion. The goal is to determine whether or not a grouping structure plausibly exists. We wish to develop a test which will incorrectly conclude that there is a latent grouping structure when in fact there is none with some preassigned level of significance. Once we have established a level α test we are interested in comparing the power of this method with some other available methods for various alternatives. We continue to use the letter g to denote the number of groups which are present, and we focus on the following hypotheses:

$H_0 : g = 1$, i.e. blocks all belong to the same group. Additive model is sufficient.
$H_1 : g = 2$, i.e. blocks belong to two groups. Effects of the other factor on response vary across these groups.

While it is certainly of interest to investigate the possibility of $g > 2$, for the purpose of this manuscript we restrict our attention to this simpler case. Focus on this simpler

case is consistent with many current papers in the literature such as Garel (1996). In the $g = 2$ case, there are $c = 2^{(b-1)} - 1$ possible groupings of blocks such that neither group is empty. We refer to the set of possible groupings as "configurations."

In the $g = 2$ case we want to count all the unique ways in which the blocks may be put into two non-empty groups. There are 2^b ways of placing blocks into two groups. Notice this is the standard ordered sampling with replacement formula. We wish to eliminate the configurations where all blocks are contained in one (or the other) group. This leaves us with $2^b - 2$ such groupings. Notice that half of these configurations are duplicate since the group label is arbitrary. For example the case where the first block is in the first group and the other blocks are in the second group is identical to the case where the first block is in the second group and all of the other blocks are in the first group. The grouping of which blocks are together is what is important and the group label is irrelevant. So we must divide the above expression by two which leaves the formula $c = 2^{b-1} - 1$ possible configurations.

The all configurations framework has unique features and challenges. For instance, statistics which arise from the configurations are not independent. In the context of linear models the null distributions of many statistics are known. However, the joint distribution of many of these same statistics over all configurations is not immediately clear. This precludes the use of several convenient strategies available for independent data, including convenient order statistic results for independent samples.

As an example of the type of complication this introduces consider the following. The p-value that corresponds to the maximum interaction F test among all configurations

would be an intuitive statistic to examine. After all, we are naively choosing to investigate every possible configuration of blocks and so the configuration with the lowest p-value would seem to correspond to the configuration which holds the most evidence for the existence of $g = 2$ latent groups. If the c such p-values from all configurations were independent, then the null distribution of the minimum would be $Beta(1, c)$. It would be straightforward to develop a critical value for a level α test based on this null distribution and render a decision by comparing the minimum interaction p-value for a given data set to this cutoff. Unfortunately simulations indicate that the distribution of the minimum p-value is not well approximated by a $Beta(1, c)$ distribution. General distributional results for this minimum p-value given the dependency between the configurations are more difficult to derive than in the independent case and have remained elusive.

Figure 2.1 provides a visual illustration of the impact of dependency between configurations on the minimum interaction p-value. The histogram was generated by plotting the minimum p-value for one thousand additive Monte Carlo (MC) data sets, each of which had four blocks and three treatments. In the $b = 4$ case there are $c = 7$ possible configurations of blocks into $g = 2$ groups. Since the data is truly additive, the interaction p-values from each of the seven configurations are identically (but not independently) distributed according to a $uniform(0, 1)$ distribution. Were these p-values independent then the minimum such p-value would arise from a $Beta(1, c)$ distribution, and the histogram would approximate this distribution. The $Beta(1, c)$ distribution is portrayed by the overlay and it is clear that the dependency between the configurations leads to a poor approximation of the $Beta(1, c)$ null distribution in the independence case. Hence using the independent order statistics result would lead to poor statistical inference and is not recommended in this case.

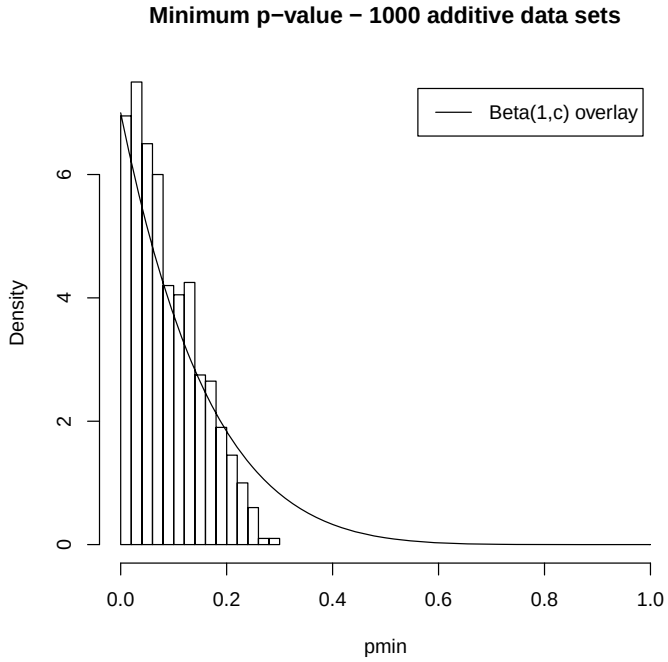

Figure 2.1: 1000 MC Minimum p-values from additive data with four blocks and three treatments. The histogram of these p-values does not conform to the Beta(1,c) overlay because of the dependency between configurations.

Our most promising all configurations strategy is to investigate the minimum of all interaction p-values and perform a Bonferroni adjustment based on the c configurations. For even modest sized experiments the issue of multiplicity arises since the number of configurations c increases rapidly as a function of the number of blocks. If each of the c interaction p-values is judged against α then the probability of avoiding all type I errors is bigger than $1-\alpha$ since many p-values are under simultaneous assessment.

The Bonferroni adjustment is made by splitting the α region into c equally sized parts. To motivate the Bonferroni adjustment, for any events A_v, $P(\bigcap_{v=1}^{c} A_v) \leq \sum_{v=1}^{c} P(A_v)$. Let us define A_v as the event where we reject the null hypothesis associated with the vth treatment by group interaction F test when there is no latent grouping structure. By comparing p_v to $\frac{\alpha}{c}$ we have bounded the probability of A_v above by $\frac{\alpha}{c}$ for $v = 1, ..., c$. Hence $P(\bigcap_{v=1}^{c} A_v) \leq \sum_{v=1}^{c} P(A_v) = \sum_{v=1}^{c} \frac{\alpha}{c} = \alpha$. The probability of wrongly rejecting any of the Treatment$\times$Group interaction F tests at the $\frac{\alpha}{c}$ cutoff is less than α for $v = 1, ..., c$. Alternatively one can multiply a raw p-value by c and compare this Bonferroni adjusted p-value with α directly. We will assess power under various alternatives in Section 2.6. Unless specifically stated otherwise, we refer to the Bonferroni adjusted maximum interaction F criteria as ACMIF from now on.

By replacing α with $\frac{\alpha}{c}$ when calculating the cutoff from the null distribution we ensure that $P(\text{rejecting any test}|H_0 \text{ true}) < \alpha$ even though these p-values are not independent. In many contexts, such as mean separation in the ANOVA framework, the Bonferroni adjustment is regarded as one of the more conservative multiplicity adjustments. Our simulation study in Section 2.6 compares power of ACMIF to Tukey's one degree of free-

dom test for several latent group-based interaction scenarios.

To apply the Bonferroni adjustment to the ACMIF test let $F_1, ..., F_c$ represent the c Treatment$\times$Group interaction effects corresponding to all possible configurations, and let $F_{(1)}, ..., F_{(c)}$ represent the corresponding order statistics such that $F_{(1)} < F_{(2)} < ... < F_{(c)}$. These F statistics are each on $(a-1)(g-1)$ numerator degrees of freedom and $(a-1)(b-g)$ denominator degrees of freedom. Each of these F statistics yields a p-value. Denote these ordered p-values $p_{(1)}, .., p_{(c)}$. Then $F_{(c)}$ corresponds to $p_{(1)}$. If the null hypothesis is true then there is no environmental structure whatsoever in the data. In this case $F_1, ..., F_c \overset{id}{\sim} F_{(a-1)(g-1),(a-1)(b-g)}$. Notice these are identically distributed but not independent random variables.

Once all of the p-values are computed the strategy is to compare $p_{(1)}$ to $\frac{\alpha}{c}$. If $p_{(1)} < \frac{\alpha}{c}$ then reject $H_0 : g = 1$. Otherwise fail to reject H_0. This testing procedure maintains a type I error rate α due to the Bonferroni correction.

2.1 Types of group-based interaction

Many of the existing methods to detect nonadditivity in unreplicated experiments assume a form of interaction based on an accompanying model. Clearly the power of any of these tests would seem to be higher for data which exhibits non-additivity of the specified form, and possibly lower for non-additivity of a different form.

There are an infinite number of ways two factors can interact just as surely as there

are an infinite number of ways to draw a line on a Cartesian plane. There are also a number of ways to subdivide this broad class of possibilities. For instance Rao (1998) uses interaction plots to characterize interaction in 2×2 experiments as either quantitative if the treatment effects at a given level of the other factor have the same sign slope, or qualitative if the treatment effects at the levels of the other factor have opposite sign slopes. In our case the other factor is the group. Figure 2.2 shows these two types of interaction. Since most of the current methods to detect non-additivity focus on specific subclasses of interaction it is desirable to classify various non-additive forms by the type of method which can detect them. Methods which assume a restricted interaction parameterization perform best for data that exhibits interaction consistent with the restrictions. For our methods latent group based interactions are the non-additive form of interest. We are interested in investigating any type of non-additivity that can be explained by the existence of latent groups which contain the blocks. With this in mind we now turn the discussion to some forms of non-additivity which can be characterized by our group based interactions. While the types of latent group based interaction discussed herein are not a comprehensive list, they demonstrate the breadth of scenarios which fall under the class of latent group interactions.

Figure 2.3 illustrates some of these forms when there are two treatment groups. Each panel includes two profiles which correspond to the two groups. The top left quadrant of the plot illustrates the null hypothesis of additive data: There is only $g = 1$ group here so any configuration of blocks into two groups will yield two groups that behave identically on average. The top right panel illustrates a condition where one group respond to the treatments (call this behavior "active"), while the other group does not respond to either treatment and is "inert." This scenario may occur in the aforementioned clinical trial

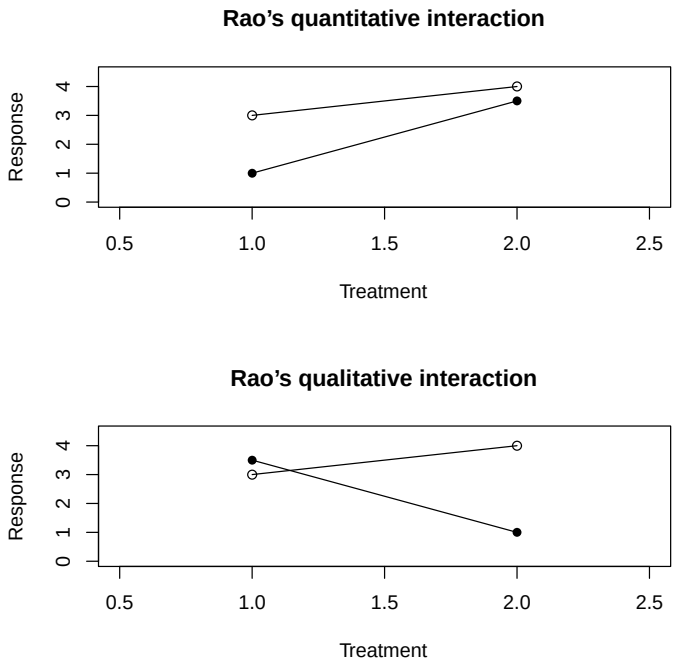

Figure 2.2: Visual examples of some types of group-based interactions for two treatments.

if one group of patients respond to the treatments, while another group does not. The groupings may be due to some unknown genetic difference between patients. The lower left panel illustrates a similar situation where the inert group is replaced by an "average" group. Even though these most recent panels are similar in some ways, we will see that detection of this type of crossing profile is more problematic for some methods than others. The lower right panel illustrates a form of non-additivity we call "cancellatory". In this setting the optimal treatments for the two groups are opposite. This situation may arise in learning tasks where the two groups are composed of people with different learning styles.

The forms of interaction for the $a = 2$ case are also of interest for $a > 2$. As the number of treatment levels grow other forms of group based interaction become possible. Figure (2.4) illustrates some interesting possible forms of interaction in the $a = 3$ treatment case. The "one optimal" and "two optimal" types may occur if some type of intervention is successful on some subset of a population. The "accordion" pattern is an extreme case of crisscrossing profiles. The "different optimal" pattern may occur if different subsets of the population have an increased response for different treatments.

2.2 Statistical inference in the all configurations framework

Now that we have proposed an all configurations method to detect latent group interactions and discussed some possible forms of latent group interaction, our attention turns to conducting statistical inference based on the result of the ACMIF analysis. We will

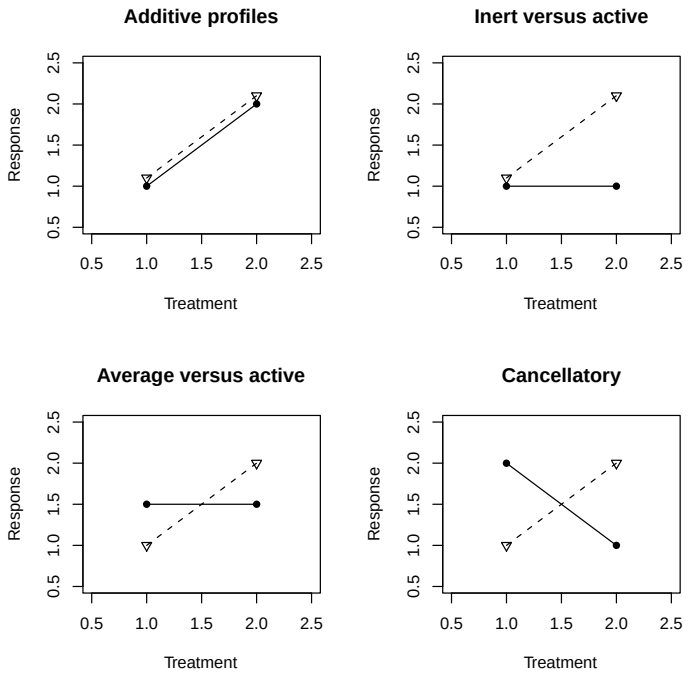

Figure 2.3: Visual examples of some types of group-based interactions for two treatments.

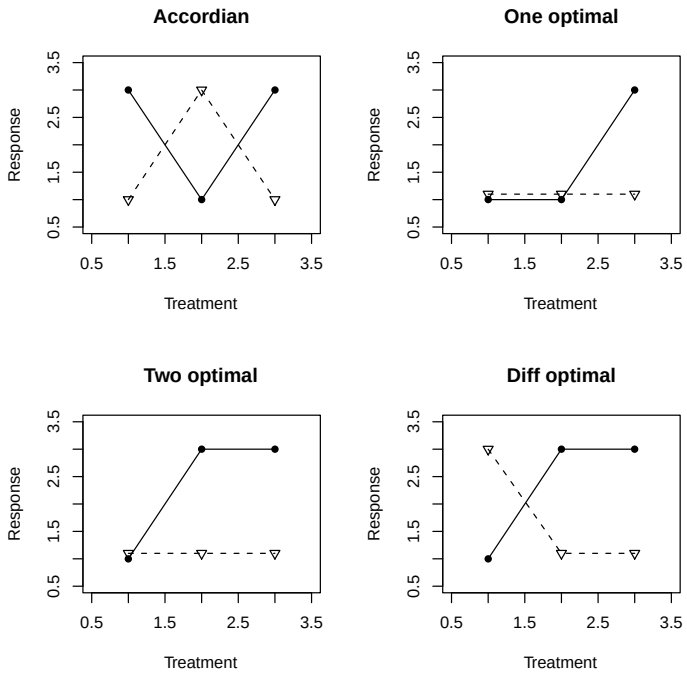

Figure 2.4: Visual examples of some types of group-based interactions for three treatments.

favor either the notion that there is no latent group structure (i.e. $g = 1$), or we will come to believe that there are two latent groups ($g = 2$) which contain the blocks, and this group effect plausibly interacts with the treatment effect. Development of an estimator under model uncertainty is the goal.

Suppose we are interested in assessing the difference in treatment levels i and i'. If we know the null hypothesis $H_0 : g = 1$ is true we estimate the difference in these treatment levels using the additive model (1.2). Denote this estimator $\tilde{\gamma}$ and note

$$\tilde{\gamma} = \bar{\boldsymbol{y}}_{i'\cdot} - \bar{\boldsymbol{y}}_{i\cdot} \ . \tag{2.3}$$

Now suppose we know that the alternative hypothesis $H_1 : g = 2$ is true. If this is the case then we would like to adjust our inference techniques to reflect our belief that latent groups exist and interact with the treatments. In the presence of treatment by group interaction we are interested in assessing the difference in means across treatment levels i and i' for a specific group k. Some authors refer to this contrast of means across levels of one factor while a second factor is held fixed as a simple factorial effect. Let

$$\gamma = \alpha_{i'} - \alpha_i + (\alpha\xi)_{i'k} - (\alpha\xi)_{ik} \ . \tag{2.4}$$

Estimation of the parameters in (2.4) is not trivial. Even with *a priori* knowledge that H_1 is true, the membership of the blocks in the groups remains a latent variable. In this manuscript we will explore use of the configuration prescribed by the ACMIF procedure. We will use the configuration of blocks which corresponds to $F_{(c)}$. In other words, we will take the grouping of blocks which provides the most evidence in favor of H_1 and use this configuration to estimate γ. We will use breves to denote estimation of parameters

in this fashion. For example $\breve{\gamma}$ is the least squares estimator of the following expression:

$$\gamma = \alpha_{i'} - \alpha_i + (\alpha\xi)_{i'k} - (\alpha\xi)_{ik} \ . \tag{2.5}$$

If we know which hypothesis is true in advance then we have proposed a strategy to estimate the difference in treatment means between treatment i and i'. In practice we need to perform a hypothesis test in order to determine whether H_0 or H_1 is preferable. We will suggest an estimator that uses the results from the ACMIF procedure to estimate γ:

$$\gamma^* = \begin{cases} \tilde{\gamma} & \text{if } p_{(1)} > \frac{\alpha}{c} \\ \breve{\gamma} & \text{else} \end{cases} \ . \tag{2.6}$$

Note γ^* takes the value of the appropriate estimator of γ from either the additive model or the group-based model depending on the decision we arrive at using the ACMIF. If we conclude that $g = 1$ then γ^* takes the value prescribed by the additive model (1.2). If we reject the null hypothesis $H_0 : \ g = 1$ then γ^* is an estimator based on model (2.1).

The final estimator we will investigate is the least conventional. We wish to consider an estimator which is a weighted average of $\tilde{\gamma}$ and $\breve{\gamma}$. The weight should be chosen using some measure of evidence of one model versus the other. Consider

$$\dot{\gamma} = p_{(1)}\tilde{\gamma} + (1 - p_{(1)})\breve{\gamma} \ . \tag{2.7}$$

We wish to emphasize that for $\dot{\gamma}$ we are not justifying the use of $p_{(1)}$ based on its probabilistic interpretation. A p-value is the probability of a test statistic arising which is as extreme or more extreme than the observed statistic if the null hypothesis (and modeling

Table 2.3: Form of each of the intervals under investigation.

Interval	Formula
$\tilde{\gamma}$ interval	$\tilde{\gamma} \pm t_{(a-1)(b-1)} \times SE(\tilde{\gamma})$
$\breve{\gamma}$ interval	$\breve{\gamma} \pm t_{(a-1)(b-g)} \times SE(\breve{\gamma})$
γ^* interval	$\tilde{\gamma}$ or $\breve{\gamma}$ interval as chosen by ACMIF
$\gamma^* \pm max(MOE)$	$\gamma^* \pm max\left(MOE(\tilde{\gamma}), MOE(\breve{\gamma})\right)$
$\dot{\gamma}$ interval	$\dot{\gamma} \pm t_{(a-1)(b-g)} UB(SE(\dot{\gamma}))$

assumptions) are true. In the frequentist paradigm a p-value should never be interpreted as the probability that the null hypothesis is true. Our justification for using $p_{(1)}$ is that it will be small if the alternative hypothesis is true but larger if the null hypothesis is true. In this way we use it as a convenient weight. We will comment on the performance of $\tilde{\gamma}$, $\breve{\gamma}$, γ^*,and $\dot{\gamma}$ in terms of mean squared error (MSE) in Section 2.6.

In addition to the point estimators $\tilde{\gamma}$, $\breve{\gamma}$, γ^*, and $\dot{\gamma}$ we are interested in assessing the performance of various confidence intervals based on these estimators. In our simulation study we will investigate the coverage of four types of intervals: One interval from the additive model, an interval based on γ^* using either the additive interval or group interval as chosen by γ^*, an interval based on γ^* using the wider margin of error from either $\tilde{\gamma}$ or $\breve{\gamma}$, and an interval based on $\dot{\gamma}$.

Table 2.3 shows a number of intervals based on the various estimators under consideration. Let MOE stand for margin of error, which in this case is the product of the appropriate t distribution quantile multiplied by the standard error of an estimator of the treatment effect. The first two rows are standard Wald-type intervals based on the additive and group models respectively. The γ^* interval is simply whichever

of the first two intervals is chosen by the ACMIF method. The $\gamma^* \pm max(MOE)$ interval uses the γ^* estimator but chooses the wider margin of error between additive and group intervals. The idea here is to account for model uncertainty due to the estimated number of groups based on ACMIF by using a wider interval. Let $UB = \sqrt{p_{(1)}^2 \times \widehat{Var}(\tilde{\gamma}) + (1 - p_{(1)}^2) \times \widehat{Var}(\breve{\gamma}) + 2p_{(1)}(1 - p_{(1)})\widehat{Var}(\tilde{\gamma})\widehat{Var}(\breve{\gamma})}$. UB is an upper bound for the standard error of $\dot{\gamma}$ since in general $Cov(\tilde{\gamma}, \breve{\gamma}) \leq SE(\tilde{\gamma}) * SE(\breve{\gamma})$. Notice that between using an upper bound for the standard error of $\dot{\gamma}$ and the more conservative $t_{(a-1)(b-g)}$ critical value we may expect the length of this latter interval to be greater than the others. We will compare the performance of several of these intervals in Section 2.6.

Other methods to estimate block membership in groups include clustering and mixture model methods. Instead of choosing the most significant grouping we could also explore standard clustering techniques. It would seem that any technique for assigning blocks to groups will be as successful as its ability to incorporate the group by treatment interaction into the clustering algorithm. In the mixture model framework we could regard the random vector of data $\boldsymbol{y}_j$ arising from block j as a draw from a two component multivariate normal mixture mixture model. The two components in this mixture model represent the two groups which contain the blocks. We could then model the probability that a given block j arises from the kth component using the following formula

$$P(\text{block } j \text{ belongs to group } k) = \frac{\omega f_i(\boldsymbol{y}_j|\boldsymbol{\theta}_k)}{\omega f_1(\boldsymbol{y}_j|\boldsymbol{\theta}_1) + (1-\omega) f_2(\boldsymbol{y}_j|\boldsymbol{\theta}_2)} \ . \tag{2.8}$$

where $0 < \omega < 1$ is a mixing weight, $\omega > (1 - \omega)$, f_i represents the component density for group i which is multivariate normal in this case, and $\boldsymbol{\theta}_i$ is the vector of parameters associated with the ith component. In the multivariate normal case this includes a mean

vector and covariance matrix. Mixture model notation will be more formally defined and discussed in Chapter 3. Assigning each $\boldsymbol{y}_j$ into its most probable component according to (2.8) is equivalent to the Bayes rule for assigning i.i.d. data into components (McLachlan and Peel, 2000). One way to approximate this rule is to replace the parameters ω and $\boldsymbol{\theta}_k$ for $k = 1, 2$ in (2.8) with suitable estimates. A 'most probable' assignment can be made by placing each block in whichever group has a higher estimated probability of membership.

One drawback of the mixture model approach described here is that mixture models typically require a fairly large number of observations in order for the model fit to be adequate. The ACMIF method was designed to be effective with a class of experiments that is probably too small to benefit from mixture theory. Nonetheless, a most probable configuration based on mixture model assignment remains as an interesting avenue for future research. We comment on the empirical performance of the γ^* estimator in section 2.6

2.3 Illustrative example - Sweetpotatoes

Truong et al. (2010) provide a relevant food science experiment. The experiment was conducted in the following manner: At each of the three time points (four, seven and eleven months) two lugs of potatoes were randomly selected from storage. A lug is a plastic bin which is filled with 40 potatoes from storage. Potatoes from each lug were randomly assigned to two groups for a total yield of four groups of potatoes. Preliminary analysis did not show evidence of variability between these four groups due to lugs. At

Table 2.4: ANOVA table based on the replicated analysis

Source	df	SS	MS	F	p-value
Time	2	16.6	8.3	40.8	< 0.0001
Stage	3	23.1	7.7	37.8	< 0.0001
Time*Stage	6	6.3	1.0	5.1	0.0007
Error	36	7.3	0.2	-	-
Total	47	53.3	-	-	-

each of the processing steps all 40 potatoes were processed, and a fixed amount of potato was taken from this total amount of material at each step for glucose testing. The first step is the raw potatoes. For the second processing step each group was cut into strips, blanched and dehydrated. Glucose measurements were taken on a portion of the dehydrated potato strips from each group. Then the remaining strips were parcooked and another portion of the strips from each group was measured for glucose. Parcook stands for "partial cook." Frozen french fries from the grocery store are in the parcooked stage. Finally the remaining strips of each group were fried to finish the processing. Glucose measurements were immediately taken on the finished French fries. This is an example of a replicated crossed two way factorial experiment. In this case it is possible to estimate the full complement of interaction parameters. The ANOVA table based on model (1.1) is portrayed in Table 2.4.

Based on the standard two way model there is strong evidence that interaction is present between Time and Stage judging by the small p-value (0.0007).

In order to illustrate our method we now regard the average glucose at each combination of time and stage as a single observation in an unreplicated experiment. Our reasons for

this approach are manifold. First, since this data set truly has replication and we have used the two way model (1.1) to establish interaction using standard ANOVA techniques, we can see if our method is able to detect this interaction based on the more crude cell mean data. Second, the most significant grouping established by our method lends credible interpretation to the experimental phenomenon: We will see that the most significant configuration puts raw potatoes in one group and all three processed stages in the other. Third and perhaps most importantly, this example illustrates how our method may be useful in situations where the data arise from an experiment where subsampling was implemented in place of true replication. Statisticians who collaborate with researchers outside the field of statistics sometimes encounter experiments where subsampling was mistakenly used in place of true replication. In a case such as this, one can regard cell means of subsampled units as single observations in an unreplicated experiment and apply the techniques described herein

For this particular dataset we are not placing blocks into groups but rather treatments. We will refer to all possible groupings of treatments (instead of blocks) as configurations for the duration of this example. The data are illustrated in Figure 2.5. To conduct this analysis we run model (2.1) on every possible configuration of the treatments into $g = 2$ groups such that no group is empty. In this case there are $c = 7$ configurations. Four of those configurations place a single treatment in one group and all other treatments in the other. The other three configurations consist of all possible groupings of two versus two treatments. Based on the plot it seems that the three processed stages have an additive relationship with storage time in terms of mean glucose. However there is some concern visually that the raw potatoes do not have the same type of performance in terms of additivity with storage time with respect to mean glucose. Despite this visual evidence

of possible interaction the Tukey single degree of freedom test for non-additivity does not provide strong evidence of an interaction effect in this case (p-value = 0.16).

The all configurations maximum interaction F based on the $c = 7$ runs of model (2.1) is 60.8 on two numerator degrees of freedom and four denominator degrees of freedom. This leads to an unadjusted p-value = 0.0010. The Bonferroni adjusted p-value based on the $c = 7$ configurations is 0.0070, so at $\alpha = 0.05$ we reject the null hypothesis of no grouping structure and instead favor the alternative. The most significant configuration places raw potatoes into one group and the other three post-processed types of potatoes into the other group. Our naive method of looking at all possible configurations of treatments into groups has yielded a grouping which corroborates the visual evidence from Figure (2.5) and also makes sense given the context of the experiment.

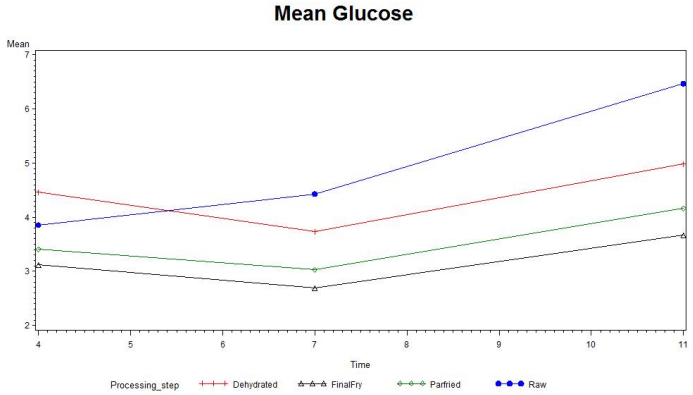

Figure 2.5: Mean glucose at each combination of storage time and processing stage.

2.4 Illustrative example - Dog Lymphoma (CGH)

Comparative genomic hybridization (CGH) is a method designed to assess copy number variations (CNV) in an individual. An example of an occurrence of CNV is when more (or fewer) than two copies of an autosomal region of the genome are found in cells of diploid organisms. One application of CGH is to determine if the amount of DNA in normal cells is different than the amount of DNA in tumor cells. This technique is also a genome wide screening procedure (Shinawi and Wai Cheung, 2008). In this illustrative example normal samples from dogs are compared to tumor samples. The response variable for this technique is a light intensity reading which is used as a proxy for DNA increases or losses in tumor cells relative to normal cells. The technique uses arrays which are variable among themselves and are considered blocks for our application. Probes correspond to a gene located on the chromosome which is positively related to lymphoma.

For this illustrative example each of the probes is regarded as an individual data set to which a test for non-additivity is applied. The block term is the pairing of normal and test samples. There are $b = 6$ blocks per probe. The $a = 2$ treatment levels are the normal sample and the test sample.

For a particular chromosome of interest there are 5899 candidate probes for which we test for non-additivity. Figure 2.6 shows the histogram of p-values from Tukey's one degree of freedom test for non-additivity for this data. Tukey's test was applied to each probe and the non-uniformity in the histogram suggests that non-additivity may be present in some of these probes. The identity of the chromosome under discussion is being withheld at the request of our collaborators.

Table 2.5: Comparison of rejection rates for Tukey's test and ACMIF. Table counts refer to frequency of rejection of null hypothesis at $\alpha = 0.05$. Cell percentages are in parentheses. Non-zero off-diagonal counts suggest that nonadditivity is present which is unique to each method.

	Tukey Fail to reject H_0	Tukey Reject H_0	Total
ACMIF Fail to reject H_0	3919 (66.43)	1296 (21.97)	5215 (88.40)
ACMIF Reject H_0	410 (6.95)	274 (4.64)	684 (11.60)
Total	4329 (73.37)	1570 (26.61)	5899 (100.00)

Multiplicity is always a concern when conducting many tests simultaneously. In this example we are testing almost six thousand probes with two methods for about twelve thousand tests total. In practice the researcher would almost certainly want to implement a multiplicity adjustment for a case such as this. However, our purpose here is merely to investigate and compare the performance of ACMIF and Tukey for these data. Hence we make no multiplicity adjustment based on the twelve thousand tests but instead report the unadjusted significance counts for the purpose of comparison.

Table 2.5 compares the rejection rates of Tukey's test and ACMIF among all of the probes for the chromosome of interest. The null hypothesis for Tukey's test is $H_0^{Tukey} : \nu = 0$ in model (1.2) which is equivalent to $H_0 : g = 1$ in the all configurations approach. For this table $\alpha = 0.05$. The first entry in each cell is the count and the entry in parentheses is the proportion of counts in that cell. About sixty-six percent of the probes show no significant non-additivity for either method. Roughly twenty-two percent of the data sets

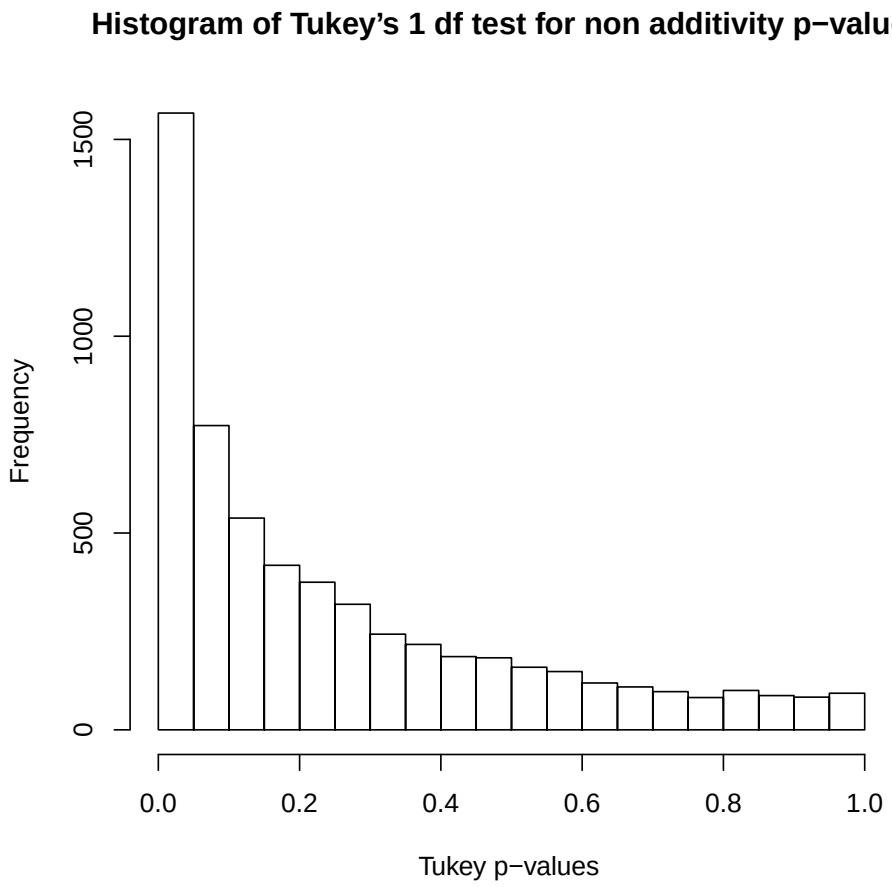

Figure 2.6: Histogram of Tukey p-values for each of the 5899 probes for chromosome of interest. The non-uniformity of this plot suggests that non-additivity may be present in these data.

show non-additivity according to the Tukey test but not ACMIF. About seven percent of the data sets show non-additivity according to ACMIF but not Tukey. Notice that the smallest percentage (4.64) of sets show non-additivity exists according to both Tukey and ACMIF methods. The non-zero off-diagonal counts suggest that there are probes which show non-additivity of a form that can only be detected by one of these two methods. The presence of these different classes of non-additivity emphasize the need for multiple methods to assess non-additivity in unreplicated factorial experiments.

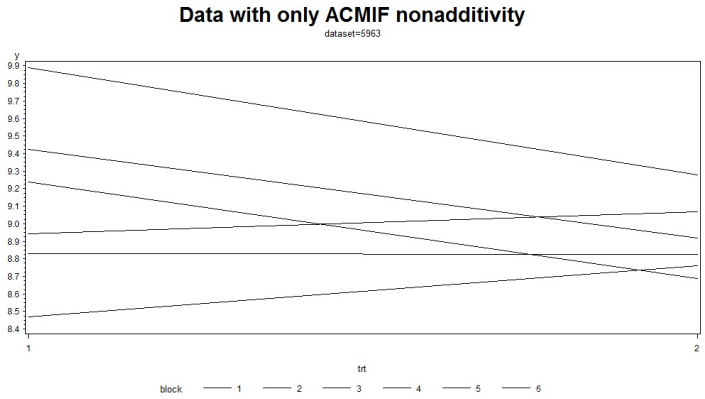

Figure 2.7: Data from probe 5963 shows non-additivity according to the ACMIF test but not the Tukey test. Notice the similarity between this pattern and the cancellatory pattern in Section 2.1.

Figure *A*.1 is a scatterplot of adjusted ACMIF p-values and Tukey p-values. This plot does not reveal any particulary interesting patterns.

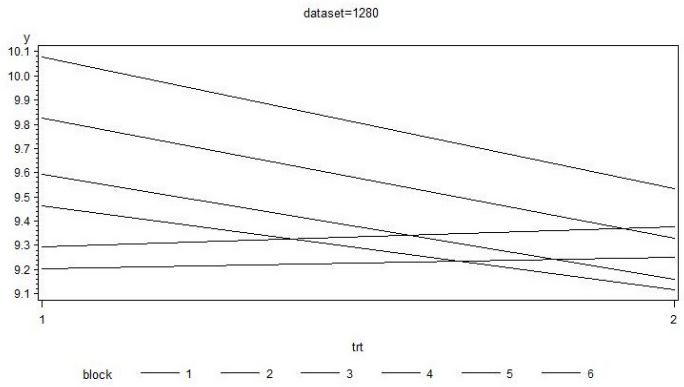

Figure 2.8: Data from probe 1280 shows non-additivity according to the ACMIF test but not the Tukey test. Notice the similarity between this pattern and the inert versus active pattern in Section 2.1.

2.5 Maximum interaction F test for large data sets.

One difficulty with the ACMIF approach is that the number of possible configurations c can be large for even moderate sized experiments. While our simulation study suggests that the Bonferroni adjustment performs well for moderate sized experiments, there are still remaining issues of algorithm coding and computational expense associated with a large number of analyses which can accompany the ACMIF method when many blocks are present.

One tangible benefit of the Bonferroni adjustment based on $c = 2^{b-1} - 1$ configurations is that any subset of interesting configurations may be screened in advance of testing the entire complement of configurations. For instance, if a dataset exhibits potential non-additivity in the interaction plot then the researcher can advance screen a subset of all possible configurations, and if these initial tests beat the $\frac{\alpha}{c}$ cutoff then the researcher may save the considerable computational expense of investigating all other configurations.

Suppose that upon looking at an interaction plot the researcher feels one possible configuration of blocks into groups is a good candidate for exposing treatment by group interaction of the form presented in 2.1. Let $F_\diamond$ be the treatment by group interaction F ratio for this configuration. Denote $p_\diamond$ as the corresponding p-value. Notice $F_{(C)} \geq F_\diamond$ which implies $p_{(c)} \leq p_\diamond$. If $p_\diamond < \frac{\alpha}{c}$ then $p_{(1)} < \frac{\alpha}{c}$ and we can reject $H_0 : g = 1$ and conclude that the additive model is not sufficient. On the other hand, if $p_\diamond \geq \frac{\alpha}{c}$ then we may not wish to render a final decision in the ACMIF framework without considering the balance of remaining configurations. It should be noted that the cutoff $\frac{\alpha}{c}$ is based on all possible configurations. As such, any number of p-values may be chosen as a subset of all possible p-values to screen. If any of the interaction p-values are lower than the cutoff we may reject H_0 with α level significance.

Notice that we do not change the Bonferroni cutoff to some smaller number of comparisons even though we are only screening a subset of the possible p-values. If we are only interested in screening one particularly suspicious configuration's p-value why not compare it to α instead of $\frac{\alpha}{c}$? We have found that for additive data sets there are virtually always some p-values that fall below reasonable α levels even in the absence of any grouping structure. Said another way, if we decide to cherry pick the most suspicious configuration and compare the associated p-value for the treatment$\times$group interaction to α, we will commit a type I error with much greater probability than α. Hence we strongly advise adhering to the Bonferroni adjustment based on all configurations for this screening procedure. We will implement this screening procedure on the third illustrative data set in Chapter 3.

2.6 Simulation study

A simulation study was conducted on the Bonferroni adjusted ACMIF in order to verify the level α properties and assess the power at various alternatives. In all simulations $N = 1,000$ MC data sets were generated. Data were generated from each combination of $a = 2, 3$ and $b = 4, 7, 11$ cases. The variance was investigated at levels $\sigma^2 = 1, 5, 10$ and the block variance was set to $\sigma_B^2 = 1$ for all cases. The type I error rate and power of ACMIF was compared with Tukey's one degree of freedom test for non-additivity in the $a = 2, 3$ case. Block membership in groups was split as evenly as possible. When $b = 4$ each group had two blocks. When $b = 7$ one group contained three blocks and the other group had four. When $b = 11$ the groups contained five and six blocks respectively. Data were generated in accordance with model (2.1) for a given configuration, and for various levels of treatment effect as specified in the tables below.

For the purpose of this simulation study the block term was modeled as a fixed effect. In practice mixed modeling procedures occasionally yield singular covariance matrix estimates when random effects are included in the model. For example in the mixed model considered here, whenever $MS(Block) < MS(Error)$ an algorithm will produce singular estimated covariance matrices for the estimates where MS stands for 'Mean Square'. This event happens with probability greater than zero even when the parameter values themselves provide a nonsingular matrix. In simulation this can become problematic if individual MC datasets encounter numerical problems such as this. To avoid these problems and maintain our focus on the latent group based interaction point we simplify the simulation by modeling blocks as a fixed effect for both ACMIF and Tukey's test. For the purpose of this study we used $\alpha = 0.05$ for all decisions. We will show that

Table 2.6: Non-additivity labels and treatment vectors for each group. The labels describe the above treatment means for the rest of Chapter 3.

a	Label	group 1 treatment mean vector	group 2 treatment mean vector
2	Additive	(0,20)	-
2	Inert vs. active	$(0, 0)$	$(0, 20)$
2	Average vs. active	$(10, 10)$	$(0, 20)$
2	Cancellatory	$(20, 0)$	$(0, 20)$
3	Additive	(0,10,20)	-
3	Inert vs. active	$(0, 0, 0)$	$(0, 10, 20)$
3	Average vs. active	$(10, 10, 10)$	$(0, 10, 20)$
3	Cancellatory	$(20, 10, 0)$	$(0, 10, 20)$
3	Accordion	$(20, 0, 20)$	$(0, 20, 0)$

the Bonferroni adjusted ACMIF gives a level α test under the null hypothesis, and for many types of latent group-based interaction is able to outperform Tukey's one degree of freedom test in terms of power. The success of ACMIF illustrates that the Bonferroni adjustment is useful despite its simplicity.

Table 2.6 shows the treatment means for the various forms of additivity and non-additivity under investigation. All tables in Chapter 2 use these labels for the sake of brevity. To deduce the parameter values which correspond to these mean vectors set the appropriate vector element equal to the systematic part of model (2.1) and solve for the parameters.

As an example consider the second row of Table 2.6. Set

$$\begin{aligned} \mu + \alpha_1 + \xi_1 + (\alpha\xi)_{11} &= 0 \\ \mu + \alpha_1 + \xi_2 + (\alpha\xi)_{12} &= 0 \\ \mu + \alpha_2 + \xi_1 + (\alpha\xi)_{21} &= 0 \\ \mu + \alpha_2 + \xi_2 + (\alpha\xi)_{22} &= 20 \end{aligned} \quad . \tag{2.9}$$

By also taking advantage of model sum to zero constraints we can solve for the following parameter values: $\mu = \alpha_2 = \xi_2 = (\alpha\xi)_{11} = (\alpha\xi)_{22} = 5$ and $\alpha_1 = \xi_1 = (\alpha\xi)_{12} = (\alpha\xi)_{21} = -5$. This general strategy can be used in any of the above cases to determine the model parameters for each simulation setting. We will not present all of these values since the group mean vectors more concisely identify the types of latent group interaction we have described.

Table 2.7 shows the empirical rejection rates of ACMIF and Tukey's single degree of freedom test for $N = 1000$ additive data sets. Since the null hypothesis is true in this case a level α test should have a rejection rate of α. One entry in the table had an empirical rejection rate significantly above 0.05 according to a one-sided binomial distribution test. However with 36 tests conducted, each at $\alpha = 0.05$, we would have expected 1.8 type I errors to occur. The rest of the tests are not significantly different from 0.05 which agrees with our earlier statements that both Tukey's test and ACMIF are level α tests.

Table 2.8 includes the MSE for each of the point estimators for the additive data. The formula for MSE is $MSE = \frac{1}{N}\sum(estimate - parameter\ value)^2$. Unsurprisingly the estimator based on the additive model $\tilde{\gamma}$ performs best across the board. This estimator is based on the model from which the simulation data was generated and hence it enjoys an advantage in this case. The second best point estimator seems to be γ^*. This makes

Table 2.7: Simulation results for additive data. First three columns describe experimental setup. Empirical rejection rates are given for both ACMIF and Tukey's test for non-additivity at $\alpha = 0.05$. Results are based on 1,000 MC data sets. The * denotes a run significantly different from actual error rate $\alpha = 0.05$ (unadjusted p-value=0.00207). Standard error for all proportions presented less than 0.0159.

Additive Data			Empirical rejection rates	
σ^2	b	a	ACMIF	Tukey
1	4	2	0.046	0.060
5	4	2	0.051	0.041
10	4	2	0.065*	0.054
1	7	2	0.037	0.045
5	7	2	0.032	0.056
10	7	2	0.033	0.042
1	11	2	0.037	0.043
5	11	2	0.051	0.034
10	11	2	0.047	0.054
1	4	3	0.052	0.049
5	4	3	0.051	0.048
10	4	3	0.049	0.060
1	7	3	0.030	0.052
5	7	3	0.036	0.051
10	7	3	0.037	0.045
1	11	3	0.041	0.040
5	11	3	0.044	0.057
10	11	3	0.047	0.043

Table 2.8: MSE results for four types of estimators for additive data. First three columns describe experimental setup. Other columns report mean squared error for the various estimators. Results are based on 1,000 MC data sets.

Additive data			MSE			
σ^2	b	a	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	4	2	0.50	3.33	0.64	2.96
5	4	2	2.50	15.68	3.39	13.95
10	4	2	4.90	34.30	6.71	30.48
1	7	2	0.28	3.81	0.45	3.71
5	7	2	1.48	18.34	2.25	17.85
10	7	2	3.03	39.15	4.61	38.14
1	11	2	0.18	3.05	0.28	3.04
5	11	2	0.86	16.30	1.53	16.25
10	11	2	1.78	32.90	3.50	32.81
1	4	3	0.52	3.08	0.71	2.71
5	4	3	2.46	14.45	3.48	12.76
10	4	3	4.77	28.66	6.06	25.18
1	7	3	0.27	2.93	0.43	2.84
5	7	3	1.36	15.22	2.09	14.72
10	7	3	3.00	30.12	4.28	29.12
1	11	3	0.19	2.31	0.25	2.30
5	11	3	0.93	11.64	1.58	11.59
10	11	3	1.73	23.72	2.83	23.62

Table 2.9: Squared bias results for four types of estimators for additive data. First three columns describe experimental setup. Other columns report squared bias for the various estimators. Results are based on 1,000 MC data sets.

Additive data			Bias^2			
σ^2	b	a	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	4	2	0.00	0.00	0.00	0.00
5	4	2	0.00	0.00	0.00	0.00
10	4	2	0.00	0.00	0.00	0.00
1	7	2	0.00	0.00	0.00	0.00
5	7	2	0.01	0.02	0.01	0.02
10	7	2	0.00	0.05	0.00	0.05
1	11	2	0.00	0.01	0.00	0.01
5	11	2	0.00	0.00	0.00	0.00
10	11	2	0.00	0.01	0.00	0.01
1	4	3	0.00	0.02	0.00	0.02
5	4	3	0.00	0.00	0.00	0.00
10	4	3	0.00	0.00	0.00	0.00
1	7	3	0.00	0.00	0.00	0.00
5	7	3	0.01	0.03	0.01	0.03
10	7	3	0.00	0.00	0.01	0.00
1	11	3	0.00	0.00	0.00	0.00
5	11	3	0.00	0.00	0.00	0.00
10	11	3	0.00	0.00	0.00	0.00

Table 2.10: Variance results for four types of estimators for additive data. First three columns describe experimental setup. Other columns report variance for the various estimators. Note $variance + Bias^2 = MSE$. Results are based on 1,000 MC data sets.

Additive data			Variance			
σ^2	b	a	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	4	2	0.50	3.33	0.64	2.96
5	4	2	2.50	15.68	3.39	13.95
10	4	2	4.90	34.30	6.71	30.48
1	7	2	0.28	3.81	0.45	3.71
5	7	2	1.47	18.32	2.24	17.83
10	7	2	3.03	39.09	4.60	38.09
1	11	2	0.18	3.04	0.28	3.03
5	11	2	0.86	16.30	1.53	16.25
10	11	2	1.77	32.89	3.50	32.80
1	4	3	0.52	3.06	0.71	2.69
5	4	3	2.45	14.45	3.48	12.76
10	4	3	4.77	28.66	6.06	25.18
1	7	3	0.27	2.93	0.43	2.84
5	7	3	1.35	15.19	2.08	14.69
10	7	3	3.00	30.12	4.27	29.12
1	11	3	0.19	2.31	0.25	2.30
5	11	3	0.93	11.64	1.58	11.58
10	11	3	1.73	23.72	2.83	23.62

sense since we expect γ^* to take the same value as the additive model 95% of the time. The weighted average $\dot{\gamma}$ seems to outperform $\breve{\gamma}$. This is reasonable since $\dot{\gamma}$ incorporates some information from the additive model while $\breve{\gamma}$ is based entirely on the group model. Table 2.9 shows the squared bias for the estimators in the additive data scenario, and Table 2.10 presents the variance for the additive data scenario. Squared bias is presented instead of bias since $MSE = Variance + Bias^2$. Table entries with bias of zero are due to rounding error.

Table 2.11 shows the empirical coverage rate of the four types of intervals for $1,000$ additive MC data sets. For the most part the interval based on $\tilde{\gamma}$ has the best coverage which is not surprising since the data is additive. However in some of the $a = 3$ cases presented the $\gamma^* \pm max(MOE)$ has slightly better empirical coverage. Notice the poor performance of $\dot{\gamma}$ interval in the additive case.

Tables $A.1$, $A.6$, $A.11$, $A.16$, 2.12, $A.21$ all report empirical power of ACMIF and Tukey's test for various treatment, block sizes, interaction types, and variances. These tables all have several things in common which make interpretation simple. As variance σ^2 increases, power decreases across the board. For the inert versus active interaction type Tukey seems to outperform ACMIF for higher variances. Increasing the number of blocks seems to increase the power in all cases as well. At lower variances and higher block sizes both methods have similar empirical power close to 1 for the configurations presented here. For the other two types of interaction, average versus active and cancellatory ACMIF outperform Tukey's test, sometimes by a wide margin. These tables illustrate the ability of ACMIF to perform well in all latent-group interaction cases as b increases. While Tukey's method enjoys superior performance in the inert versus active case this

Table 2.11: Coverage proportions for four types of intervals for additive data. First three columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Results are based on 1,000 MC data sets. Standard error for all proportions presented less than 0.0159.

Additive data			Empirical coverage proportion of confidence interval for each method.				
σ^2	b	a	$\tilde{\gamma}$ interval	$\breve{\gamma}$ interval	γ^* interval	$\gamma^* \pm max(MOE)$	$\dot{\gamma}$ interval
1	4	2	0.949	0.703	0.906	0.946	0.708
5	4	2	0.932	0.689	0.889	0.931	0.723
10	4	2	0.945	0.686	0.893	0.934	0.734
1	7	2	0.954	0.333	0.919	0.935	0.337
5	7	2	0.945	0.323	0.919	0.929	0.342
10	7	2	0.943	0.305	0.913	0.937	0.35
1	11	2	0.943	0.11	0.909	0.925	0.111
5	11	2	0.951	0.087	0.902	0.927	0.087
10	11	2	0.955	0.106	0.908	0.921	0.106
1	4	3	0.931	0.739	0.901	0.944	0.74
5	4	3	0.956	0.739	0.919	0.949	0.783
10	4	3	0.951	0.753	0.92	0.956	0.827
1	7	3	0.948	0.582	0.927	0.966	0.587
5	7	3	0.963	0.551	0.937	0.959	0.57
10	7	3	0.944	0.56	0.92	0.952	0.596
1	11	3	0.95	0.44	0.925	0.957	0.44
5	11	3	0.95	0.426	0.923	0.952	0.426
10	11	3	0.954	0.42	0.923	0.948	0.421

method does not perform well in the other latent group interaction cases in general.

Tables $A.2$, $A.7$, $A.12$, $A.17$, 2.13, $A.22$ all present MSE for the four point estimators $\tilde{\gamma}$, $\breve{\gamma}$, γ^*, $\dot{\gamma}$ for various treatment, block sizes, interaction types, and variances. Tables $A.3$, $A.8$, $A.13$, $A.18$, 2.14, $A.23$ present the associated squared bias and Tables $A.4$, $A.9$, $A.14$, $A.19$, 2.15, $A.24$ show the variance for these estimators in each of the simulated scenarios. In each of these tables the additive estimator $\tilde{\gamma}$ has poor performance compared to any of the other estimators. This comes as no surprise since the data from these tables all exhibit interaction. However this phenomenon does illustrate the danger of using the additive model when latent group-based interaction is present. Also $\breve{\gamma}$ has the best performance for all of these cases which is sensible since this estimator is based on the latent group model and the simulated data arise from this model. The real comparison then is between $\gamma*$ and $\dot{\gamma}$. For these simulations $\dot{\gamma}$ outperforms or ties γ^* in every case. Note that as b increases the number of configurations will increase and hence value of $p_{(1)}$ will tend to get smaller. Hence we expect $\dot{\gamma}$ to perform similarly to $\breve{\gamma}$, particularly for the higher block sizes. Since $\dot{\gamma}$ more closely follows the group model than the additive model with $p_{(1)}$ typically being very small it is not surprising that $\dot{\gamma}$ outperforms γ^* in the non-additive simulations but γ^* is superior to $\dot{\gamma}$ when the data is truly additive. Note that for the fully powered scenarios in this simulation $\gamma^* = \breve{\gamma}$ by definition.

Tables $A.5$, $A.10$, $A.15$, $A.20$, 2.16, $A.25$ show the empirical coverage rates of the four different types of intervals for various treatment, block sizes, interaction types, and variances. Notice that the accordion type non-additivity is represented when $a = 3$ but not when $a = 2$ (in the $a = 2$ case cancellatory is similar to the accordion pattern.). The simple effect we are attempting to capture in this coverage experiment is the difference in

means between the first and second treatments in the first group when $a = 2$, and the difference in the first and third treatment means when $a = 3$. One interesting feature of this simulation is that we may actually expect coverage rates to be better when b is smaller. The intervals will shrink as b increases, and due to the modeling uncertainty involved in the estimators (remember we are estimating group membership of the blocks using the most significant configuration), there is a chance these intervals could be centered at the wrong value. Hence increasing b will not only shorten the interval but possibly increase the model uncertainty as well. This is because larger b means larger c there may be a greater chance that the most significant configuration is not the true configuration.

In Table $A.5$ all of the coverage proportions look reasonable. This is likely due to very wide intervals which capture the parameter value since b is small. In Table $A.10$ many of the intervals show worsening coverage as σ^2 increases. The interval based on the additive estimator $\tilde{\gamma}$ performs poorly overall regardless of σ^2, which makes sense since these data are not additive.

In the remainder of the coverage tables the coverages seem to lose performance as the variability in the data increases. The fact that the interval coverage suffers as σ^2 increases shows that we have achieved the nominal confidence level $100(1-\alpha)\%$. However for the nonadditive data $\dot{\gamma}$ does seem to perform the best. However it is constructed using an upper bound on the standard error and too-conservative critical value, and both of these features probably loan the $\dot{\gamma}$ interval an unfair advantage.

Table 2.12: Simulation results for various latent group non-additive scenarios. Three treatments and seven blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF different from Tukey's test for all cases when estimates not equal (p-values < 0.0001). Standard error for all proportions presented less than 0.0159.

$a = 3$; $b = 7$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2, b_3)$, $g_2 = (b_4, b_5, b_6, b_7)$			
		Proportion of MC rejections at $\alpha = 0.05$	
σ^2	Data type	ACMIF	Tukey
1	Inert vs. active	1.000	1.000
5		0.961	1.000
10		0.638	0.996
1	Average vs. active	1.000	0.218
5		0.964	0.176
10		0.681	0.147
1	Cancellatory	1.000	0.167
5		1.000	0.140
10		1.000	0.116
1	Accordion	1.000	1.000
5		1.000	0.972
10		1.000	0.852

Table 2.13: MSE results for four types of estimators under various latent group non-additive scenarios. Three treatments and seven blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators.

$a = 3$; $b = 7$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3)$, $g_2 = (b_4, b_5, b_6, b_7)$					
		MSE			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	130.71	0.61	0.61	0.61
5		133.41	3.15	8.47	3.15
10		132.14	20.99	53.53	20.90
1	Average vs. active	130.76	0.69	0.69	0.69
5		132.59	3.47	8.05	3.47
10		131.92	10.93	46.39	10.89
1	Cancellatory	522.40	0.63	0.63	0.63
5		522.33	3.37	3.37	3.37
10		525.66	7.26	7.26	7.26
1	Accordion	522.30	0.67	0.67	0.67
5		521.84	3.29	3.29	3.29
10		524.79	6.25	6.25	6.25

Table 2.14: Squared bias results for four types of estimators under various latent group non-additive scenarios. Three treatments and seven blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators.

$a = 3$; $b = 7$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3)$, $g_2 = (b_4, b_5, b_6, b_7)$					
		Bias^2			
1	Inert vs. active	130.46	0.00	0.00	0.00
5		131.98	0.01	0.24	0.01
10		129.12	0.02	12.51	0.03
1	Average vs. active	130.47	0.00	0.00	0.00
5		131.14	0.00	0.14	0.00
10		129.17	0.08	9.31	0.07
1	Cancellatory	522.14	0.00	0.00	0.00
5		520.93	0.00	0.00	0.00
10		522.64	0.00	0.00	0.00
1	Accordion	522.01	0.00	0.00	0.00
5		520.45	0.00	0.00	0.00
10		522.06	0.00	0.00	0.00

Table 2.15: Variance results for four types of estimators under various latent group non-additive scenarios. Three treatments and seven blocks were used. First two columns describe experimental setup. Other columns report variance of the various estimators.

$a = 3$; $b = 7$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3)$, $g_2 = (b_4, b_5, b_6, b_7)$					
		Variance			
1	Inert vs. active	0.25	0.61	0.61	0.61
5		1.43	3.14	8.23	3.14
10		3.02	20.97	41.03	20.87
1	Average vs. active	0.29	0.69	0.69	0.69
5		1.45	3.47	7.91	3.47
10		2.76	10.85	37.08	10.81
1	Cancellatory	0.26	0.63	0.63	0.63
5		1.40	3.37	3.37	3.37
10		3.02	7.26	7.26	7.26
1	Accordion	0.29	0.67	0.67	0.67
5		1.39	3.29	3.29	3.29
10		2.73	6.25	6.25	6.25

Figures *A*.2, 2.9, *A*.3, *A*.4, *A*.5, *A*.6 all show empirical power for various interaction types, treatment numbers, and block numbers for both methods. In general these plots reflect the previous tables by showing that Tukey is superior to ACMIF in the inert versus active data for low block sizes and high variances. As variance goes down ACMIF catches up to Tukey in the inert versus active case, and ACMIF outperforms Tukey in all other cases. One observation from these tables is that the Bonferroni corrected ACMIF does not lose power as the number of configurations increases from $c = 7$ up to at least $c = 1023$. The Bonferroni correction becomes increasingly severe as the number of tests increases, so the fact that the ACMIF criterion can withstand this increasing severity for the size experiments investigated in this simulation study is encouraging. We may speculate that for some $b > 11$ the increasing severity of the Bonferroni correction will eventually overwhelm the evidence provided by ACMIF. We present alternative method-

Table 2.16: Coverage proportions for four types of intervals under various latent group non-additive scenarios. Three treatments and seven blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159.

$a = 3$; $b = 7$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2, b_3)$, $g_2 = (b_4, b_5, b_6, b_7)$						
		Empirical coverage proportion of confidence interval.				
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\gamma^* \pm max(MOE)$	$\dot{\gamma}$
1	Inert vs. active	0.000	0.962	0.962	1.000	0.962
5		0.000	0.946	0.908	0.961	0.946
10		0.022	0.931	0.609	0.649	0.931
1	Average vs. active	0.000	0.941	0.941	1.000	0.941
5		0.000	0.943	0.908	0.962	0.943
10		0.021	0.949	0.647	0.682	0.949
1	Cancellatory	0.000	0.953	0.953	1.000	0.953
5		0.000	0.950	0.950	1.000	0.950
10		0.000	0.947	0.947	1.000	0.947
1	Accordion	0.000	0.951	0.951	1.000	0.951
5		0.000	0.951	0.951	1.000	0.951
10		0.000	0.947	0.947	1.000	0.947

ology in Chapter 3 which uses mixture model theory and is designed to be used on larger experiments.

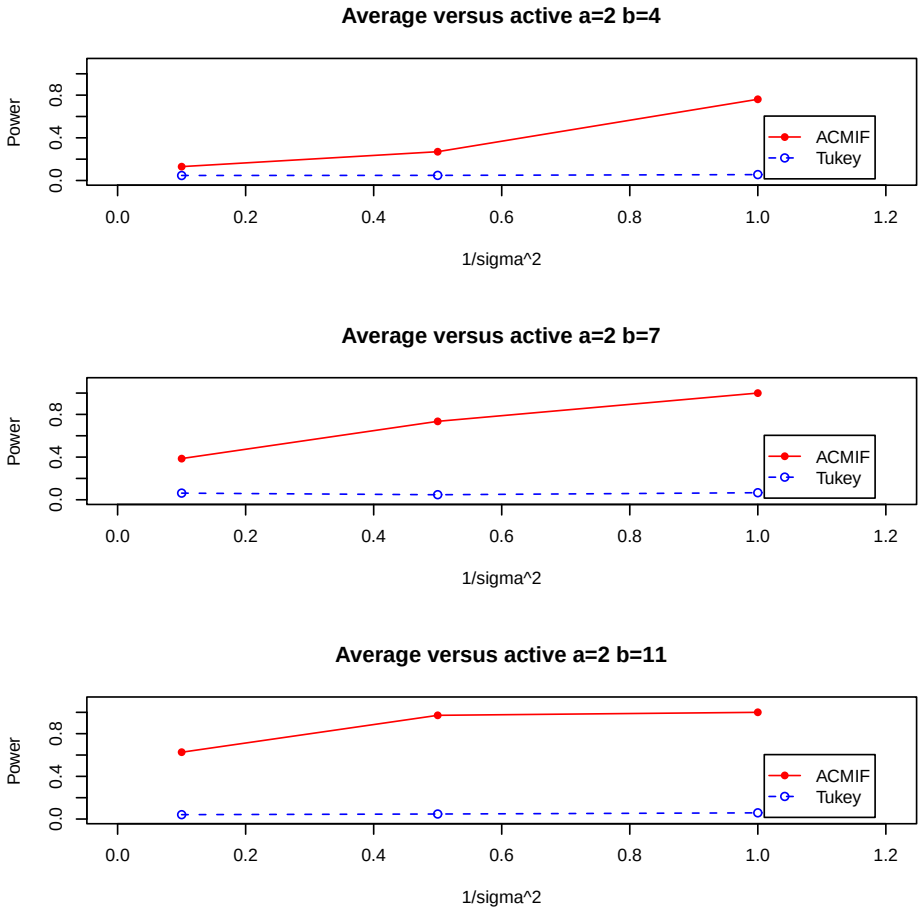

Figure 2.9: Power plots for average versus active data for both methods. $a = 2$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis.

Chapter 3

Bayes Factor approach

In addition to the all configurations method described in Chapter 2 we have also explored a Bayes Factor (BF) approximation for the assessment of latent group stucture. Since BF does not rely on asymptotic likelihood-based regularity conditions we have avoided some known problems in the frequentist paradigm. The strategy is to investigate latent grouping structure using multivariate normal finite mixture model theory. If the data set in question exhibits no latent grouping structure then the data vector from each block $\boldsymbol{y}_j$ for $j = 1, ..., b$ can be regarded as a random vector from the multivariate normal distribution with a common mean vector and covariance matrix. If the data has a two latent group structure, the $\boldsymbol{y}_j$ can be regarded as a draw from a two component multivariate normal mixture model where each of the two components corresponds to one of the two latent groups. We assume the components have different mean vectors but share a common covariance matrix. Generalization to the different covariance matrix case should be straightforward. Our goal is to fit each model, integrate over the respective parameter spaces to get the marginal distribution of the data for each model respectively, and use these marginal distributions to compute BF. Then we can use BF as a measure of evi-

dence in favor of one model over the other.

Let $\boldsymbol{y}$ denote the $a \times b$ matrix of data values where the jth column is denoted $\boldsymbol{y}_j$. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represent the mean vector and covariance matrix of the multivariate normal distribution. Let $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}, \omega)$ represent the two mean vectors, common covariance matrix, and mixing weight of a two component multivariate normal mixture distribution. Since we typically model blocks as random and we are considering the equal covariance matrix case for each component, $\boldsymbol{\Sigma}$ will exhibit compound symmetry of the following form:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 + \sigma_B^2 & \sigma_B^2 & \dots & \sigma_B^2 \\ \sigma_B^2 & \sigma^2 + \sigma_B^2 & \dots & \sigma_B^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_B^2 & \sigma_B^2 & \dots & \sigma^2 + \sigma_B^2 \end{bmatrix} .$$

Let $f_0(\boldsymbol{y}|\boldsymbol{\theta}_0)$ and $f_1(\boldsymbol{y}|\boldsymbol{\theta}_1)$ represent the conditional distribution of the data $\boldsymbol{y}$ given the parameters $\boldsymbol{\theta}_i$ for $i = \{0, 1\}$. Since we model data from different blocks as independent

$$f_i(\boldsymbol{y}|\boldsymbol{\theta}_i) = \prod_{j=1}^{b} f_i(\boldsymbol{y}_j|\boldsymbol{\theta}_i); i = \{0, 1\}; j = 1, ..., b \ . \tag{3.1}$$

Let the prior distribution of the parameters be denoted by $\pi(\boldsymbol{\theta}_i)$. Bayes factor is defined as:

$$BF = \frac{m_1(\boldsymbol{y})}{m_0(\boldsymbol{y})} \ . \tag{3.2}$$

where

$$m_i(\boldsymbol{y}) = \int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i, i \in \{0, 1\} \ . \tag{3.3}$$

is the marginal distribution of the data for model i. In this formulation, values of BF

greater than one give evidence that a two component mixture is more plausible for data $\boldsymbol{y}$ than a one component multivariate normal mixture. Values of BF lower than 1 suggest a one component multivariate normal distribution is more plausible than the two component mixture. Carrying out the necessary integration for the marginal distribution is analytically intractable in this case. Hence an approximation to the marginal distribution is needed. We describe an MC method to approximate the marginal distributions in the next section.

3.1 Approximation of the marginal distribution of the data.

Our marginal distribution approximation uses the following MC strategy:

$$\hat{m_i}(\boldsymbol{y}) = \frac{1}{M}\sum_{m=1}^{M} f_i(\boldsymbol{y}|\boldsymbol{\theta}_i^{(m)}) \xrightarrow{p} \int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta_i})d\boldsymbol{\theta}_i = m_i(\boldsymbol{y}) \ . \tag{3.4}$$

The above result 3.4 holds by the weak law of large numbers (WLLN) if $E_{\boldsymbol{\theta}_i}(f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)) < \infty$. We refer to the above method as a 'direct' MC approximation to distinguish it from an importance sampling based approximation to the marginal which was also investigated.

To implement the above approximation, draw $m = 1, ..., M$ random samples from the prior distribution $\pi(\boldsymbol{\theta}_i)$. Plug these M values into $f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)$ denoting the mth such value $f_i(\boldsymbol{y}|\boldsymbol{\theta}_i^{(\boldsymbol{m})})$ and compute the arithmetic mean. Note that for the remainder of this discussion the (m) superscript on $\boldsymbol{\theta}_i^{(m)}$ represents values that have been generated from a prior distribution in an MC fashion. As M increases this mean converges in probability

to the marginal distribution of interest. Several different prior specifications for $\pi(\boldsymbol{\theta_i})$ were investigated. The first prior under investigation was a vague proper prior for all of the parameters. The following parametric families were investigated for these priors

$$\begin{aligned}
\boldsymbol{\mu} \sim &\ MVN(\boldsymbol{\zeta}, \boldsymbol{\Lambda}) \\
\boldsymbol{\mu}_1 \sim &\ MVN(\boldsymbol{\zeta}_1, \boldsymbol{\Lambda}_1) \\
\boldsymbol{\mu}_2 \sim &\ MVN(\boldsymbol{\zeta}_2, \boldsymbol{\Lambda}_2) \\
\sigma^2_{(i)} \sim &\ IG(\nu_{(i)}, \rho_{(i)})\ i = \{0, 1\} \\
\sigma^2_{B(i)} \sim &\ IG(\nu_{B(i)}, \rho_{B(i)})\ i = \{0, 1\} \\
\omega \sim &\ Uniform(0, 1).
\end{aligned} \tag{3.5}$$

where $\pi(\boldsymbol{\theta}_1) = P(\boldsymbol{\mu}) \times P(\sigma^2_{(0)}) \times P(\sigma^2_{B(0)})$ and $\pi(\boldsymbol{\theta_2}) = P(\boldsymbol{\mu}_1) \times P(\boldsymbol{\mu}_2) \times P(\sigma^2_{(1)}) \times P(\sigma^2_{B(1)}) \times P(\omega)$.

One problem that occurred with vague specification is numerical underflow. The density $f_i(\boldsymbol{y}|\boldsymbol{\theta}_i^{(m)})$ can be extremely small when evaluated at parameter values drawn from a vague prior. If these values get too small then it becomes impossible for a computer to distinguish them from zero. In our investigations this happened with some regularity, which caused approximation to the marginal distribution to be too small for one or both models. While some vague prior specifications seemed to work for certain simulated data we were not able to find a general vague parameterization for proper priors that avoided numerical underflow for all data specifications.

Vague priors showed the tendency to generate values which would underflow when substituted into $f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)$, so our next strategy was to sample from some other prior which does not produce underflow values upon evaluation of $f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)$ for $i = \{0, 1\}$. The Empirical

Bayes (EB) method uses priors whose hyperparameters are obtained from the data. To emphasize that EB priors depend on the data and to distinguish EB priors from other classes of priors we will denote EB priors with $g(\boldsymbol{\theta}_i|\boldsymbol{y})$.

We now elaborate on the construction of $g(\boldsymbol{\theta}_i|\boldsymbol{y})$. In our model the parameters are independent so in the two component case $g(\boldsymbol{\theta}_1^{(m)}|\boldsymbol{y}) = p(\boldsymbol{\mu_1}) * p(\boldsymbol{\mu_2}) * p(\omega) * p(\sigma^2) * p(\sigma_B^2)$. In the one component case $g(\boldsymbol{\theta}_0^{(m)}|\boldsymbol{y}) = p(\boldsymbol{\mu}) * p(\sigma^2) * p(\sigma_B^2)$. Since numerical underflow can become problematic if $g(.)$ is too vague, we have chosen to characterize $g(\boldsymbol{\theta}_i^{(m)}|\boldsymbol{y})$ using the MLEs of the parameters from the data. The general strategy is to compute MLE estimates as well as the estimated covariance matrix of these estimates based on the data for all parameters in both models. Then we use this information to solve for the hyperparameters for the prior distributions. In the one component case the "*optim*" function was used in the R package to generate MLEs as well as the second derivative of the log likelihood, which yields the hessian matrix for the one component model. The provided hessian was then inverted to obtain the covariance matrix of the parameters. In the two component case the expectation maximization (EM) algorithm was used to generate MLEs and a bootstrap routine was implemented to estimate the covariance matrix of the parameters. We implemented code for the EM algorithm which maintains the compound symmetry of $\boldsymbol{\Sigma}$. Full details of the algorithm are included in section 4.1, and the code may be found in the Appendix B.

The following table summarizes the parametric specification for the joint density of the parameters, $g(\boldsymbol{\theta}_i|\boldsymbol{y})$ where $i = \{0, 1\}$. The hat notation is used to denote the MLE of

the associated parameter:

$$\begin{aligned}
\boldsymbol{\mu} \sim &\quad MVN(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Lambda}} = \widehat{Cov}(\hat{\boldsymbol{\mu}})) \\
\boldsymbol{\mu}_i \sim &\quad MVN(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Lambda}}_i = \widehat{Cov}(\hat{\boldsymbol{\mu}}_i)) \\
\sigma^2_{(i)} \sim &\quad IG(\hat{\alpha}_{(i)}, \hat{\beta}_{(i)}) \\
\sigma^2_{B(i)} \sim &\quad IG(\hat{\alpha}_{B(i)}, \hat{\beta}_{B(i)}) \\
\omega \sim &\quad Beta(\hat{a}, \hat{b}) \ .
\end{aligned} \tag{3.6}$$

The MLEs and estimated covariance matrices of the parameter estimates were used to construct the priors for $\sigma^2_{(i)}$, $\sigma^2_{B(i)}$, and ω. To illustrate how this was done, consider the expression for the mean and variance from the inverse gamma distribution.

$$\begin{aligned}
\text{mean} = &\quad \tfrac{\beta}{\alpha-1} \text{ for } \alpha > 1 \\
\text{variance} = &\quad \tfrac{\beta^2}{(\alpha-1)^2(\alpha-2)} \text{ for } \alpha > 2.
\end{aligned} \tag{3.7}$$

Using $\sigma^2_{(1)}$ as an example, replace the "mean" expression with the MLE $\hat{\sigma}^2_{(1)}$ and the "variance" with $\widehat{Var}(\hat{\sigma}^2_{(1)})$. Then solve for $\alpha_{(i)}$ and $\beta_{(i)}$ to get $\hat{\alpha}_{(i)}$ and $\hat{\beta}_{(i)}$, which are used in the generating function $g(.)$. The same method is used for all variance components in both models as well as the mixing weight ω in the two component model to yield the following estimates:

$$\begin{aligned}
\hat{\alpha}_{(i)} = &\quad \tfrac{(\hat{\sigma}^2_{(i)})^2}{\widehat{Var}(\hat{\sigma}^2_{(1)})} + 2 \\
\hat{\beta}_{(i)} = &\quad \tfrac{(\hat{\sigma}^2_{(i)})^3}{\widehat{Var}(\hat{\sigma}^2_{(1)})} + \hat{\sigma}^2_{(1)} \\
\hat{\alpha}_{B(i)} = &\quad \tfrac{(\hat{\sigma}^2_{B(i)})^2}{\widehat{Var}(\hat{\sigma}^2_{B(1)})} + 2 \\
\hat{\beta}_{B(i)} = &\quad \tfrac{(\hat{\sigma}^2_{B(i)})^3}{\widehat{Var}(\hat{\sigma}^2_{B(1)})} + \hat{\sigma}^2_{B(1)} \\
\hat{a} = &\quad \tfrac{\hat{\omega}^2(1-\hat{\omega})}{\widehat{Var}(\hat{\omega})} - \hat{\omega} \\
\hat{b} = &\quad \tfrac{\hat{\omega}^2(1-\hat{\omega})}{\widehat{Var}(\hat{\omega})} + \hat{\omega} - 1 \ .
\end{aligned} \tag{3.8}$$

In addition to the direct MC approximation we also explored an importance sampling approximation to the marginal distribution:

$$m_i^{IS}(\boldsymbol{y}) \approx \frac{1}{M}\sum_{m=1}^{M}\frac{f_i(\boldsymbol{y}|\boldsymbol{\theta}_i^{(m)})\pi(\boldsymbol{\theta}_i^{(m)})}{g(\boldsymbol{\theta}_i^{(m)}|\boldsymbol{y})} \ . \tag{3.9}$$

The prior distribution on the parameters is denoted, $\pi(\boldsymbol{\theta}_i^{(m)})$, the indices $m = 1, ..., M$ refer to the mth MC draw from $g(\boldsymbol{\theta}_i^{(m)}|\boldsymbol{y})$, which is the joint density of the parameters. Now $g(.)$ is a generating function and the ratio of $\pi(.)$ to $g(.)$ is called the importance weight.

To motivate the importance sampling, we want

$$\begin{aligned} m_i(\boldsymbol{y}) = & \int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \\ = & E_\pi\left[f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\right] \\ = & E_\pi\left[f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\frac{g(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}\right] \text{ for some density } g(.), \\ = & \int f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)\frac{g(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}d\boldsymbol{\theta}_i \\ = & E_g\left[f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\frac{\pi(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}\right] \ . \end{aligned} \tag{3.10}$$

If the above expectation is finite then by the strong law of large numbers

$$\frac{1}{M}\sum_{m=1}^{M}\frac{f_i(\boldsymbol{y}|\boldsymbol{\theta}_i^{(m)})\pi(\boldsymbol{\theta}_i^{(m)})}{g(\boldsymbol{\theta}_i^{(m)}|\boldsymbol{y})} \xrightarrow{p} E_g\left[f_i(\boldsymbol{y}|\boldsymbol{\theta}_i)\frac{\pi(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}\right] = m_i(\boldsymbol{y}) \ . \tag{3.11}$$

So by choosing some density $g(.)$, randomly drawing $l = 1, .., M$ observations from $g(.)$, and computing $\frac{1}{M}\sum_{m=1}^{M}\frac{f_i(\boldsymbol{y}|\boldsymbol{\theta}_i^{(m)})\pi(\boldsymbol{\theta}_i^{(m)})}{g(\boldsymbol{\theta}_i^{(m)}|\boldsymbol{y})}$ we are able to obtain another MC approximation to the marginal distribution of the data under the ith model where $i = \{0, 1\}$.

Within the importance sampling framework we chose the following parametric families of $\pi(\boldsymbol{\theta}_i)$ and $g(\boldsymbol{\theta}_i|\boldsymbol{y})$ to investigate. The prior distribution on location parameters $\boldsymbol{\mu}, \boldsymbol{\mu_1}, \boldsymbol{\mu_2}$ is taken to be multivariate normal. The mixing weight ω in the two component model is distributed according to a beta distribution. Variance components σ^2 and σ_B^2 come from inverse-gamma distributions. We are assuming compound symmetry in both one and two component cases, and in the two component case we are assuming both groups have equivalent covariance matrices. However σ^2 and σ_B^2 will clearly have different estimates in the one versus two component model for a given data set, so we denote the variance components from the former model $\sigma^2_{(0)}$ and $\sigma^2_{B(0)}$ and for the latter model we denote variance components $\sigma^2_{(1)}$ and $\sigma^2_{B(1)}$. Preliminary work was done on four specifications of $\pi(\boldsymbol{\theta}_i)$ and among these choices a vague prior performed the best and will be employed henceforth. Various hyperparameterizations were explored among them $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \mathbf{0}$, $\hat{\boldsymbol{\Lambda}} = 800\boldsymbol{I} + 200\mathbf{1}$, $\hat{\alpha}_{(i)} = \hat{\beta}_{B(i)} = \hat{\alpha}_{B(i)} = \hat{\beta}_{(i)} = 1000$, and $\hat{a} = \hat{b} = 1$.

Importance sampling is generally used for numerical integration in the situation when the numerator density ($\pi(.)$ in our case) is difficult to sample from directly. However in this case we are using importance sampling in an attempt to strike a balance between vague proper priors and a data informed generator function g(.). In our investigations the direct method yields estimates which are more stable than the importance sampling method. BF approximations from the importance sampling method are typically extremely small or large and seldom near one. Also the directional correctness (i.e. number of times BF was on the correct side of one for simulated data) was better under the direct MC approximation. The direct MC method was used for the illustrative example and the simulation study.

The ratio of the approximated marginal distributions $\frac{\hat{m_2}(\boldsymbol{y})}{\hat{m_1}(\boldsymbol{y})}$ serves as the approximation

to Bayes Factor.

3.2 Implementation of EM Algorithm for Two Component Model

The expectation-maximization (EM) algorithm was implemented to calculate the MLEs and estimate MLE covariance matrices from the two component model. The EM algorithm was developed by Dempster et al. (1977) as a general method of obtaining MLEs in the context of incomplete data problems. EM type problems arise when the data at hand can be viewed as incomplete, meaning there is some desirable data that is not available. In the finite mixture model scenario the incomplete data is the component memberships of the data points, or blocks in our case. Conceptually, we would view the combination of the available data $\boldsymbol{y}$ along with the knowlege of which component each block arose from as the complete data. While we do know $\boldsymbol{y}$ we do not know which blocks belong in which group for known g. We now describe how the EM algorithm is used to obtain MLEs and their corresponding estimated covariance matrices for the $g = 2$ two component model. We loosely follow the notation established in McLachlan and Peel (2000). Let $\boldsymbol{\Psi}$ be the vector of all mixture model parameters.

$$\boldsymbol{\Psi} = \{\omega_1, ..., \omega_{g-1}, \boldsymbol{\theta}_1^T, ..., \boldsymbol{\theta}_g^T\} \ . \tag{3.12}$$

where $\boldsymbol{\theta_k}$ is the vector of parameters associated with the kth component for $k = 1, ..., g$, and $\omega_1, ..., \omega_{g-1}$ represents the $g - 1$ mixing weights where the gth mixing weight is determined by $1 - \sum_{k=1}^{g-1} = \omega_g$. Consider the data $\boldsymbol{y} = \{\boldsymbol{y}_1^T, ..., \boldsymbol{y}_b^T\}^T$ where $\boldsymbol{y}_j$ is the $a \times 1$

vector corresponding to block j for $j = 1, ..., b$. The ith element of $\boldsymbol{y}_j$ represents the data point for treatment level i in block j for $i = 1, ..., a$. In the two component model, each $\boldsymbol{y}_j$ belongs to one of the two groups. Let $\boldsymbol{z} = \{\boldsymbol{z}_1^T, ..., \boldsymbol{z}_b^T\}$ be the component labels for the blocks and let $\boldsymbol{z}_j$ represent the $g \times 1$ component label for the jth block where in this case $g = 2$.

$$\boldsymbol{z}_j = \begin{cases} (1,0) & \text{if } \boldsymbol{y}_j \in \text{group 1} \\ (0,1) & \text{if } \boldsymbol{y}_j \in \text{group 2} \end{cases}$$

In practice the component labels are a latent variable: we do not know which block belongs in which group. Hence the $\boldsymbol{z}$ is the missing data, and we conceptually think of the combination of $\boldsymbol{z}$ and $\boldsymbol{y}$ as the complete data. In other words the complete data are both the response values from the jth block and also the identity of the component from which the jth block arose. So the complete data density for block j is $f_c(\boldsymbol{y}_j, \boldsymbol{\Psi}, \boldsymbol{z}_j) = \prod_{k=1}^{g} \omega_k f_k(\boldsymbol{y}_j; \boldsymbol{\theta}_k)^{\boldsymbol{z}_{kj}}$. Note that this previous expression is always the product of the mixing weight and component density corresponding to the component from which $\boldsymbol{y}_j$ originates. We model blocks as independent of one another so the log-likelihood of the complete data is the product of the block densities. The complete log-likelihood will be referred to as $\log L_c(\boldsymbol{\Psi}_1 | \boldsymbol{y}, \boldsymbol{z})$ and can be written as follows.

$$\begin{aligned} \log L_c(\boldsymbol{\Psi} | \boldsymbol{y}, \boldsymbol{z}) = & \log \prod_{j=1}^{b} \prod_{k=1}^{g} \omega_k f_k(\boldsymbol{y}_j; \boldsymbol{\theta}_k)^{\boldsymbol{z}_{kj}} \\ = & \sum_{j=1}^{b} \sum_{k=1}^{g} \boldsymbol{z}_{kj} \{\log(\omega_k) + \log(f_k(\boldsymbol{y}_j; \boldsymbol{\theta}_k))\} \end{aligned} . \tag{3.13}$$

The EM algorithm works by iteratively taking the expectation of the complete log-likelihood with respect to $\boldsymbol{z}_j$ given the observed data $\boldsymbol{y}$, and then maximizing the parameters over the entire parameter space to obtain updated parameter estimates. This iteration between "E" and "M" steps continues until the change in likelihood between

two steps becomes arbitrarily small. Notice that since z_{kj} is an indicator variable then for the uth step of the algorithm $E_{\boldsymbol{\Psi}(u)}\left[z_{kj}|\boldsymbol{y}\right] = P_{\boldsymbol{\Psi}(u)}(z_{kj}=1|\boldsymbol{y})$. Denote this probability $P_{\boldsymbol{\Psi}(u)}(z_{kj}=1|\boldsymbol{y}) = \tau_k(\boldsymbol{y}_j, \boldsymbol{\Psi^{(u)}})$ and let:

$$\begin{aligned} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(u)}) = & \quad E_{\boldsymbol{\Psi}(u)}\left[\log L_C(\boldsymbol{\Psi})|\boldsymbol{y}\right] \\ = & \quad \textstyle\sum_{j=1}^{b}\sum_{k=1}^{g} E_{\boldsymbol{\Psi}(u)}\left[\boldsymbol{z}_{kj}|\boldsymbol{y}\right]\left\{\log(\omega_k) + \log(f_k(\boldsymbol{y}_j;\boldsymbol{\theta}_k))\right\} \end{aligned} \quad . \tag{3.14}$$

With the expected log likelihood computed for the uth step, we get the $(u+1)th$ estimators by maximizing $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(u)})$ over $\boldsymbol{\Psi}$. The complete data MLE for ω_k for $k = 1, ..., g$ is

$$\hat{\omega}_k = \frac{1}{b}\sum_{j=1}^{b} z_{kj} \ . \tag{3.15}$$

For the $(u+1)th$ step we replace z_{kj} above with its expectation which is the probability of block j belonging to group k.

$$\omega_i^{u+1} = \frac{1}{b}\sum_{j=1}^{b} \tau_k(\boldsymbol{y}_j, \boldsymbol{\Psi^{(u+1)}}) \ . \tag{3.16}$$

To ease our notational burden we let $\tau_{kj} = \tau_k(\boldsymbol{y}_j, \boldsymbol{\Psi})$ and $\tau_{kj}^{(u)}$ be the estimated probability associated with the uth iteration of the EM algorithm. The rest of the $(u+1)th$ MLEs can be obtained by finding the appropriate roots of the following expression

$$\sum_{j=1}^{b}\sum_{k=1}^{g} \tau_k(\boldsymbol{y}_j, \boldsymbol{\Psi^{(u+1)}})\{\log(\omega_k) + \log(f_k(\boldsymbol{y}_j;\boldsymbol{\theta}_k))\} = 0 \ . \tag{3.17}$$

Evaluating these roots leads to the following estimators

$$\boldsymbol{\mu}^{(u+1)} = \frac{\sum_{j=1}^{b}\tau_{kj}^{(u)}\boldsymbol{y}_j}{\sum_{j=1}^{b}\tau_{kj}^{(u)}} \ . \tag{3.18}$$

$$\boldsymbol{\Sigma}_i^{(u+1)} = \frac{\sum_{j=1}^{b} \tau_{kj}^{(u)} (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(u+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(u+1)})^T}{\sum_{j=1}^{b} \tau_{kj}^{(u)}} \ . \tag{3.19}$$

The above expression (3.19) applies when we have different covariance matrices for each component. To apply the assumption of equal variances we use the following estimator

$$\boldsymbol{\Sigma}^{(u+1)} = \frac{1}{b} \sum_{k=1}^{g} \sum_{j=1}^{b} \tau_{kj}^{(u)} \boldsymbol{\Sigma}_i^{(u+1)} \ . \tag{3.20}$$

In this way we continue iterating between expectation steps and maximization steps until the increase in likelihood is trivially small from one step to the next. Denote the likelihood at step u as $L^{(u)}(\boldsymbol{\Psi})$. In Dempster et al. (1977) it was shown that the likelihood does not decrease as u increases, i.e. $L^{(u+1)}(\boldsymbol{\Psi}) \geq L^{(u)}(\boldsymbol{\Psi})$. So then the EM iterations continue until $L^{(u+1)}(\boldsymbol{\Psi}) - L^{(u)}(\boldsymbol{\Psi}) \leq \epsilon$ for $\epsilon > 0$.

The MLEs generated by the EM algorithm were then used to produce a bootstrap estimate of the covariance matrix of the estimators. In order to produce the bootstrap covariance matrix the following steps can be taken.

Step 1: Generate bootstrap data sets. To do this, denote $\hat{F} = L(\hat{\boldsymbol{\Psi}})$ i.e. the likelihood of the data substituting the MLE estimates $\hat{\boldsymbol{\Psi}}$ from the EM algorithm in for the true parameters $\boldsymbol{\Psi}$. Then generate data sets $\boldsymbol{y}_q^*$ from $\hat{F}$ for $q = 1, ..., B$ such that

$$\boldsymbol{y}_1^*, ..., \boldsymbol{y}_B^* \overset{iid}{\sim} \hat{F} \ . \tag{3.21}$$

Step 2: Apply EM algorithm to bootstrap data sets to get bootstrap MLEs. Denote the estimates from the qth bootstrap sample $\boldsymbol{y}_q^*$ as $\hat{\boldsymbol{\Psi}}_q^*$.

Step 3: Use the bootstrap MLEs to estimate the covariance matrix of the parameters.

$$Cov^*(\hat{\boldsymbol{\Psi}}^*) \approx \frac{1}{B-1}\sum_{q=1}^{B}\left[\hat{\boldsymbol{\Psi}}_q^* - \bar{\boldsymbol{\Psi}}^*\right]\left[\hat{\boldsymbol{\Psi}}_q^* - \bar{\boldsymbol{\Psi}}^*\right]^T . \tag{3.22}$$

where

$$\bar{\boldsymbol{\Psi}}^* = \frac{1}{b}\sum_{q=1}^{B}\hat{\boldsymbol{\Psi}}_q^* . \tag{3.23}$$

The appropriate subsections of the estimated covariance matrix can then be used to characterize $\hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Lambda}}_i$, and to characterize $\hat{\alpha}_{(i)}, \hat{\beta}_{(i)}, \hat{\alpha}_{B(i)}, \hat{\beta}_{B(i)}, \hat{a}$ and $\hat{b}$ as described above.

The EM algorithm requires starting values for the parameters in order to perform the first M step. Without any advance knowledge of which blocks belong in which group we do not have natural estimators to initialize the algorithm. In the one component case the sample means can be used as starting values for $\boldsymbol{\mu}$. For other parameters we employ a graphical approach. When unreplicated experiments are conducted in practice, one of the first indications of possible non-additivity frequently comes from non-parallel lines in an interaction plot. We use the following method for generating starting values.

First, inspect an interaction plot of the data. Assume that there are g groups (we consider the $g = 2$ case). Visually guess which blocks may be in which of the g groups, and guess what the mean of each group is at each treatment level based on the plot. Finally use the visual evidence in the plot to initialize σ^2 and σ_B^2.

One criticism of generating starting values would be that the practitioner may be "stack-

ing the deck" in favor of a two component model: By choosing initial values which corroborate specific patterns for a given dataset perhaps we will get a two component fit when the data at hand is truly from a one component model. While this is a valid concern bear in mind we will rely on our BF approximation to render a formal decision as to whether to favor a one or two component model. At this stage in the method we are not choosing between $g = 1$ or $g = 2$ components, but rather fitting a two component model so we can approximate BF later. Care should be taken that reasonable starting values are used based on the data to help ensure the EM algorithm provides reasonable estimates and associated covariances for the parameters in the two component model. Another way to initialize the starting values for component mean vectors is to use k-means clustering to cluster the b data vectors. Then use the cluster centroids as the initial estimates for the mean vectors.

3.3 Illustrative example

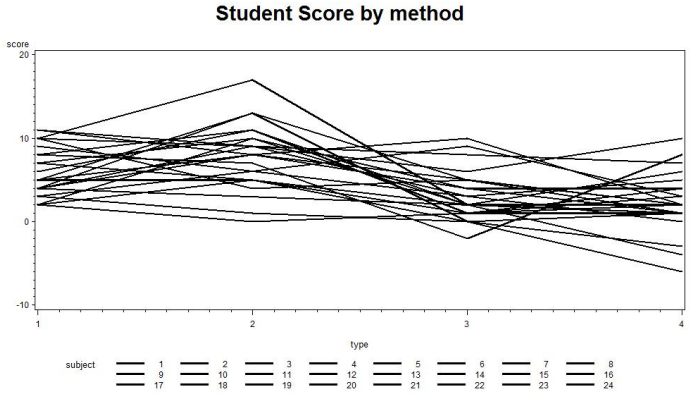

Figure 3.1: Interaction plot.

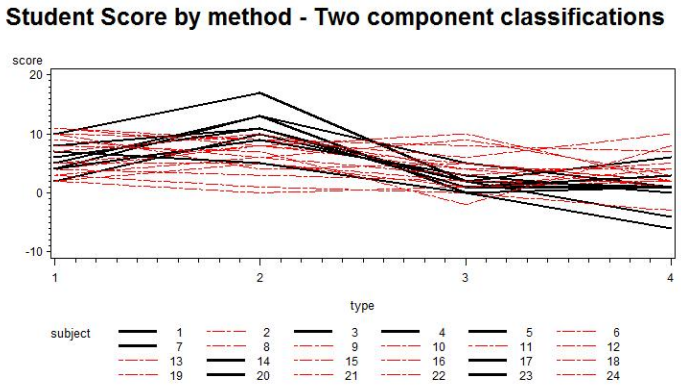

Figure 3.2: Interaction plot coded by most likely classification based on a two component model.

Our second illustrative example was taken from Daniel (2005) based on a paper by Brooks et al. (2002). The goal of the study is to compare various kitchen training programs efficacy at improving learning disabled students' culinary skill. The training programs are **1** traditional training, **2** virtual training, **3** workbook training, and **4** no training to serve as a control. Twenty four learning disabled catering students participated in the study. Each student received each training program one time and their change in food preparation score was measured as a response. The original paper states that half of the students were exposed to virtual training before the study began.

The interaction plot in 3.1 suggests that the relationship subject and treatment have with score may be non-additive. A two component model was fit to the data, and the second interaction plot in 3.2 shows the classification of each patient into one of two groups based on most probable assignment. The visual evidence of two components seems compelling based on this second plot.

For this data set the Tukey one degree of freedom test has a p=value of 0.7578. Based on this test we would conclude that any non-addivity in this data is not of the form in model (1.3). Our BF approximation for this data set is 0.0015. Based on this BF we would conclude that the one component model is more plausible than the two component model. We also screened the most probable configuration in 3.2 using a Bonferroni corrected cutoff value in the ACMIF framework. With 24 students serving as blocks, there are $c = 8,388,607$ configurations which put these students into $g = 2$ groups. The raw interaction p-value for the configuration of interest is $3.08 * 10^{-7}$. The Bonferroni adjusted cutoff based on the $\alpha = 0.05$ significance level is $\frac{\alpha}{8,388,607} = 5.96 * 10^{-9}$. Since our interaction p-value is not smaller than the Bonferroni cutoff we would fail to reject the null hypothesis at the $\alpha = 0.05$ significance level.

These analyses tell a cautionary tale. The visual evidence in the interaction plot, the nice fit of the two component model, and even some extra knowledge about some students having different exposure levels to some training methods all suggest a two component model is reasonable. However none of the methods under review favor a two component model over a one component model in an inferential framework.

3.4 Simulation Study

A simulation study was conducted to determine the efficacy of the BF approximation for deciding between $g = 1$ and $g = 2$ components. The $N = 300$ MC data sets were generated in each of the experimental configurations. BF was approximated for each set

as well as Tukey's one degree of freedom test for non-additivity.

BF summarizes evidence of one scientific hypothesis over another. A $BF = 10$ suggests that the two component model is 10 times as likely as the one component model in this case. But how does one decide how large (or small) BF needs to be in order to decide whether to favor one model or the other? Should the cutoff for favoring the two component model be 1? 10? 100? $1,000,000$? Once we decide on a cutoff, how can we compare the performance of this method to other methods which operate in the frequentist paradigm?

Even though $\alpha = 0.05$ is a widely accepted cutoff value for frequentist methods, this (or any other) cutoff may be regarded as subjective as well. One thing that unifies these different approaches is the decision theory underlying both methods. No matter which method is used, we want to make a decision as to whether or not non-additivity is present in the data at hand. With the above concerns in mind we use the following strategy to make decisions in each case.

First, we will generate $N = 300$ additive datasets and approximate BF for each. If $BF < \frac{1}{10}$ we will favor the hypothesis of $g = 1$ i.e. additivity. If $BF > 10$ we will favor the $g = 2$ hypothesis. We will also report the proportion of Tukey rejections at the $\alpha = 0.05$ level. We choose these cutoffs for the purpose of summarization even though this is not an apples to apples comparison between the methods.

For the purposes of this simulation study, σ_B^2 was set equal to zero and excluded from the modeling algorithm and BF approximation. The optimization of likelihood functions for

random effects such as we have described with σ_B^2 faces occasional numerical difficulties. One such difficulty occurs if the optimization algorithm estimates $\hat{\sigma}_B^2 < 0$. This happens with some regularity even when σ_B^2 is truly greater than zero in the data generation. While a negative variance component can still result in a positive definite $\hat{\Sigma}$ there is interpretational difficulty associated with allowing a variance component to be negative since variances in general are positive.

By default the MIXED procedure in SAS automatically excludes the random effect with variance component σ_B^2 if the corresponding estimate is negative. We experimented with a similar algorithm. We generated data with nonzero σ_B^2 and for each of the MC data sets we calculated estimates of all parameters described above. For the group of MC data sets that have $\hat{\sigma}_B^2 < 0$ we complete the BF approximation omitting σ_B^2 from all calculations. For example this means we would have $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is an $a \times a$ identity matrix. For MC sets with $\hat{\sigma}_B^2 > 0$ we complete the BF approximation with σ_B^2 in the way described above. We refer to this algorithm as the "split by fit" method.

The split by fit method has some drawbacks. It is not clear in general that the BF approximations arising from the different parameterizations are comparable. Virtually every step of the approximation to BF is different under the different parameterizations, so it is perhaps unsurprising that the BF that resulted from the simple model seemed to differ substantially from the BF which arose from the larger model.

One way to resolve this would be to run a great number of MC datasets through the split by fit method and assess the performance of the BF approximations which arise from each model separately. This would increase computation time however and distract us

from the main purpose of placing blocks into groups on the basis of treatment by group interaction. With these concerns in mind we have chosen to simplify the simulation study by omitting σ_B^2 from all calculations.

The ability of BF to distinguish between one and two component models as σ^2 increases is of interest. In the case of finite mixture models, increasing the variance can actually make a two component model resemble a one component model more closely. Figure 3.3 shows simulated univariate two component mixtures. In each of the three cases the components are centered at zero and sixty respectively. When σ^2 is low the two components are quite distinct. However as σ^2 increases the components become less distinct and eventually the mixture distribution becomes unimodal. In a case such as this it would seem possible that a one component model may effectively model the data. Based on the pattern in Figure 3.3 we might expect to see BF decrease as σ^2 increases in our simulation study. In the BF approximation method each block is treated as an observation from either a multivariate normal distribution or a two component mixture of multivariate normal distributions. Thus increasing the number of blocks in an experiment is equivalent to increasing the sample size in other types of statistical applications. We expect that as block number b increases the performance of the BF approximation will improve.

Table 3.1 provides the labels for each of the data types and the corresponding means for each group.

Figures 3.4, 3.5, $A.7$, $A.8$, and $A.9$ each show histograms of $N = 300$ log base 10 BF approximations from four levels of σ^2. In these figures $\log_{10}(BF) < 0$ represents evidence in favor of $g = 1$ while $\log_{10}(BF) > 0$ represents evidence in favor of $g = 2$. In the addi-

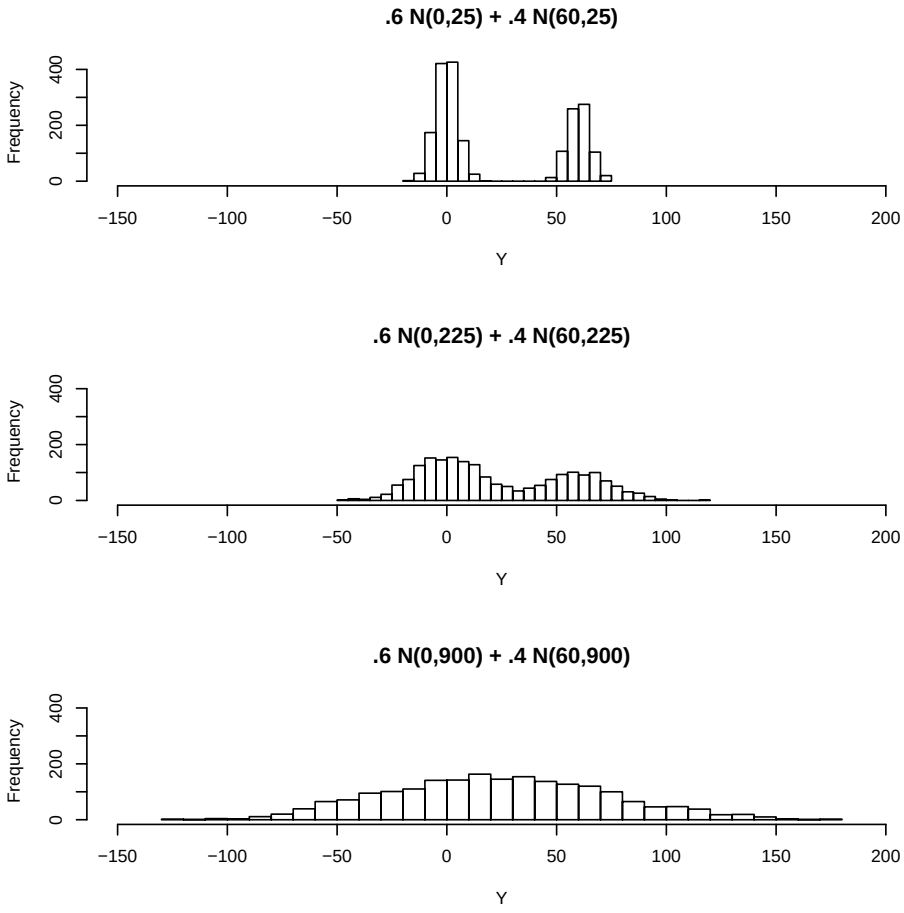

Figure 3.3: Histograms of univariate data in two component mixtures. First component is centered at zero and second component is centered at 60.

Table 3.1: Data labels and corresponding treatment means for BF simulation study.

Data type	Group 1 treatment means	Group 2 treatment means
Additive	(5,10,15,20)	-
Accordion	(0,4,0,4)	(4,0,4,0)
Cancellatory	(6,4,2,0)	(0,2,4,6)
Inert vs. active	(0,0,0,0)	(1.25,2.5,3.75,5)
Average vs. active	(6.25,6.25,6.25,6.25)	(2.5,5,7.5,10)

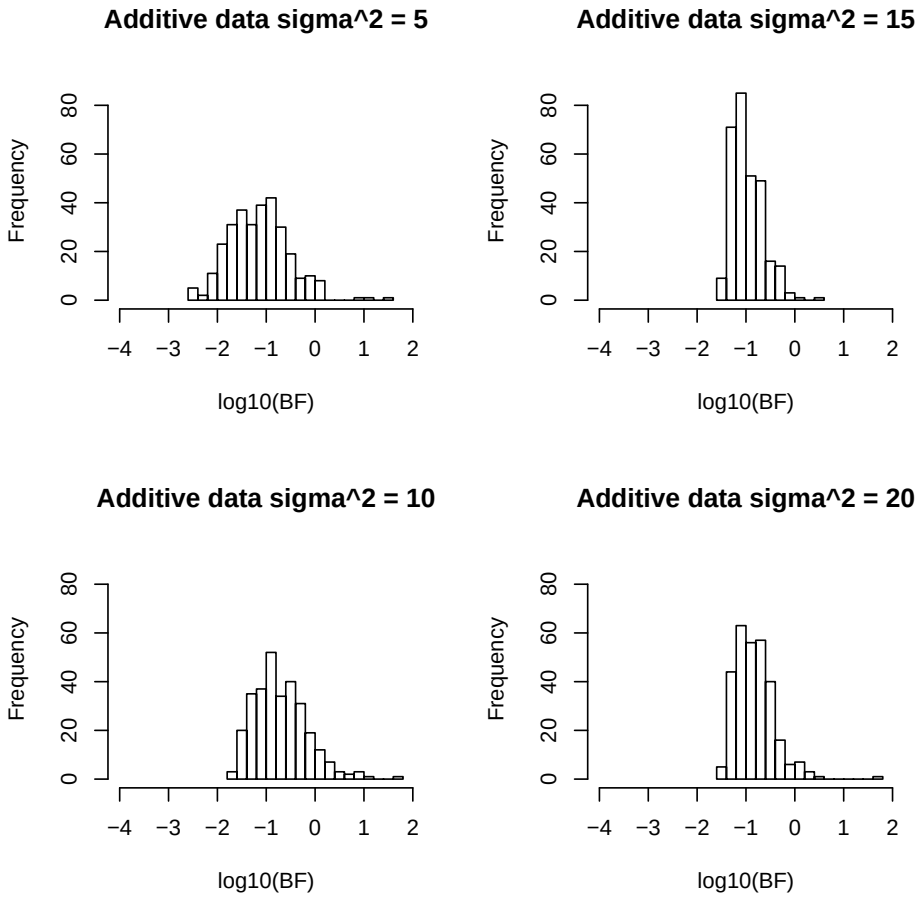

Figure 3.4: Histograms of $\log_{10}(BF)$ Bayes factors in additive data for $N = 300$ MC data sets and various levels of σ^2.

Table 3.2: Simulation distributional results: relative frequencies of BF approximation appear in columns 2-5. Median BF appears in column 6. Tukey's test rejection rates ($\alpha = 0.05$) in additive case appear in the last column.

$a = 4; b = 50$; $\omega = 0.6$; Data type: Additive ; MC reps: 300						
σ^2	$0 < BF < \frac{1}{10}$	$\frac{1}{10} < BF < 1$	$1 < BF < 10$	$BF > 10$	$median(BF)$	Tukey RR
5	0.5967	0.3667	0.0300	0.0067	0.0712	0.0533
10	0.3167	0.5867	0.0900	0.0067	0.1646	0.0600
15	0.5500	0.4433	0.0067	0.0000	0.0889	0.0567
20	0.3746	0.5853	0.0368	0.0033	0.1349	0.0500

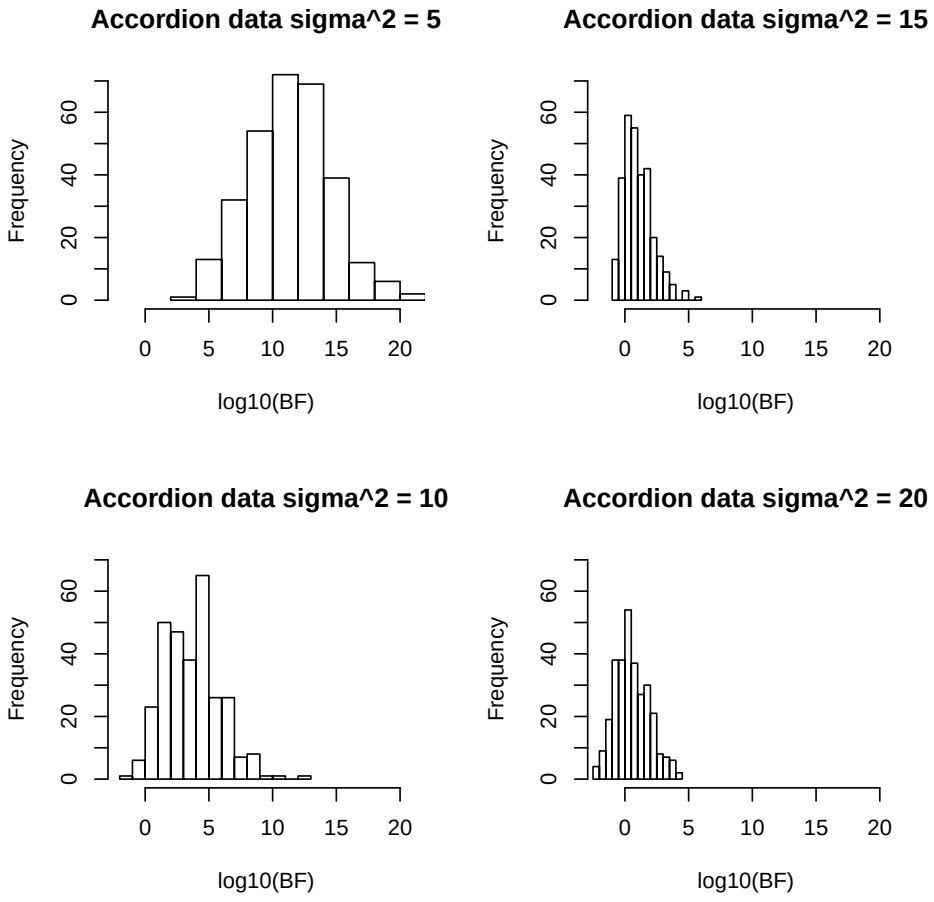

Figure 3.5: Histograms of $\log_{10}(BF)$ BF in accordion data for $N = 300$ MC data sets and various levels of σ^2.

Table 3.3: Simulation distributional results: relative frequencies of BF approximation appear in columns 2-5. Median BF appears in column 6. Tukey's test rejection rates ($\alpha = 0.05$) in accordion case appear in the last column.

$a = 4; b = 50$; $\omega = 0.6$; Data type: Accordion ; MC reps: 300 ;						
σ^2	$0 < BF < \frac{1}{10}$	$\frac{1}{10} < BF < 1$	$1 < BF < 10$	$BF > 10$	$median(BF)$	Tukey RR
5	0	0	0	1.000	2.85×10^{11}	0.0900
10	0.0033	0.0200	0.0767	0.900	4851.00	0.1033
15	0	0.1733	0.3800	0.4467	6.47	0.0667
20	0.1067	0.2533	0.3033	0.3367	2.71	0.0800

tive case we see that most of the $\log_{10}(BF)$ values are below zero corresponding to the correct decision. In all other cases we see that for low variance $\log_{10}(BF) > 0$ in most cases, but as σ^2 increases $\log_{10}(BF)$ decreases. Figure 3.3 illustrates why this may be so. If the data truly comes from a mixture model then as σ^2 increases it is harder for BF to make the correct conclusion. For additive data there is no apparent association between BF and σ^2. One encouraging feature of these histograms is that for all of the non-additive cases we see very similar patterns in BF as σ^2 increases.

Tables 3.2, 3.3, $A.26$, $A.27$, and $A.28$ show some simulation results. Each table has the proportion of BF in various regions of the real line along with the proportion of rejections from Tukey's one degree of freedom test for non-additivity at $\alpha = 0.05$. This is not an apples to apples comparison. Tukey's test is a frequentist method aimed at controlling the type I error rate under the assumption that the parameters are fixed but unknown constants which must be estimated. The BF approximation is simply a measure of plausibility of one model over another. We do not endorse $\frac{1}{10}$ and 10 or any other cutoff as a formal rule for decision making. Rather these tables are aimed at summarizing results for two separate and not directly comparable methods. The BF approximation performs very well in the additive case. When group-based interaction is present the performance of BF is best at lower variances and diminishes as σ^2 increases. We again see Tukey's method has good power in the inert versus active scenario but has lower power for all other non-additive configurations.

For this study $N = 300$ MC data sets were generated for each experimental setting in the tables of results. To approximate the marginal distributions m_1 and m_2, $M = 1000$ draws were made from $g(\theta_1|\boldsymbol{y})$ and $g(\theta_2|\boldsymbol{y})$ for the importance sampling for each of the

N data sets. For the two component model, bootstrap standard errors were computed on the basis of $B = 52$ bootstrap samples. It has been shown that B between 50 and 100 is generally sufficient for bootstrap estimation of standard errors (Efron and Tibshirani, 1993).

Towards the beginning of Chapter 3 we describe the compound symmetry of $\mathbf{\Sigma}$. In every case we generate data according to this form. However even when data is generated from a model with this covariance structure, optimizing routines occasionally have trouble fitting models with this covariance structure. The off-diagonal elements of a covariance matrix can be negative while the matrix is still positive definite in the case where variables are negatively correlated. However in our case the off-diagonal elements of $\mathbf{\Sigma}$ are all equal to σ_B^2 which must be positive.

By default some standard software packages such as the Mixed procedure in SAS exclude random effects from the model when they have negative variance component estimates. In this case if $\hat{\sigma}_B^2 \leq 0$ then σ_B^2 would be set to zero and $\mathbf{\Sigma} = \sigma^2 \boldsymbol{I}$ where I represents the $a \times a$ identity matrix. The model would not include any estimate for σ_B^2 and inference using the remaining parameter estimates in the model.

Our method of dealing with problematic estimates for σ_B^2 is similar in spirit. For our study the first step is to fit both one and two component models to each of the N data sets. Then we screen the estimates from each model. All of the data sets with $\hat{\sigma}_{B(1)}^2$ or $\hat{\sigma}_{B(0)}^2$ equal to or less than zero are placed in a separate array from the data sets which accommodate a positive estimate of σ_B^2. The problematic sets then go through the entire BF approximation algorithm using models that do not include σ_B^2 while the remainder

of sets have BF approximated using the more complicated covariance structure.

For this study the following prior distributions were used to characterize $\pi(\boldsymbol{\theta}_0)$ and $\pi(\boldsymbol{\theta}_1)$.

$$\begin{aligned} \boldsymbol{\mu} \stackrel{d}{=} \boldsymbol{\mu}_1 \stackrel{d}{=} \boldsymbol{\mu}_2 \sim & \quad Multivariate\ Normal(\mathbf{0}, \boldsymbol{\Lambda}_v) \\ \sigma^2 \stackrel{d}{=} \sigma_B^2 \sim & \quad Inverse\ Gamma(2.001, 10) \\ \omega \sim & \quad Uniform(0,1) \end{aligned} \quad . \tag{3.24}$$

where $\boldsymbol{I}$ represents the $a \times a$ identity matrix, $\mathbf{1}$ represents an $a \times a$ matrix of ones, $\mathbf{0}$ represents an a dimensional vector of zeroes, and $\boldsymbol{\Lambda}_v = 800\boldsymbol{I} + 200\mathbf{1}$.

Using the above priors and the method described earlier in this chapter, BF was approximated for each data set in a variety of experimental settings. The one component case results are summarized in Table 3.2.

Chapter 4

Concluding Remarks

Unreplicated factorial experiments can be useful when physical and/or economic constraints prohibit the researcher from replicating treatments on multiple experimental units. We have seen that one limitation of unreplicated factorial experiments is that the full set of standard interaction parameters cannot be tested using standard linear models theory due to lack of an appropriate error term.

Many of the current methods to detect non-additivity focus on restricted parameterizations for interaction. The two methods we have developed and presented are slightly different than these. Instead of focusing on a restricted form of interaction we consider the levels on one of the factors to belong to some smaller number of latent groups or environments.

We have developed an all configurations based test to detect this structure in the frequentist paradigm. The all configurations method ACMIF is practical for data with fewer blocks since c, the number of groupings, increases exponentially as a function of b, the

number of blocks. We have also presented a MC method approximation to Bayes Factor to detect this structure. Our BF approximation relies on mixture model theory and as such is well suited for experiments with a larger number of blocks.

We have conducted simulation studies to compare both of our methods to Tukey's one degree of freedom test for non-additivity. Our overall finding is that while Tukey's test seems to be able to detect some forms of latent group non-additivity (such as the 'inert versus active' case), our methods provide more powerful tests of non-additivity for latent group-based interaction data.

Estimation of a simple effect in the possible presence of the latent group structure is an interesting problem. If one concludes that $g = 2$ on the basis of the BF approximation then it would make sense to conduct statistical inference on the basis of the two component model which was fit to approximate BF. In the all configurations framework the issue of statistical inference is not fully resolved. Our simulation study indicates that the types of intervals based on ACMIF have less than $1 - \alpha$ coverage as σ^2 increases. The reason for this is the missing group labels. Even when we correctly conclude $g = 2$ we still do not know which blocks are truly grouped together. We have explored using the "most plausible" group defined as the grouping which corresponds to the maximum F interaction statistic across all possible configurations. However there is absolutely no guarantee that this is truly the correct configuration and as such there is unresolved model uncertainty associated with this decision.

For unreplicated factorial experiments there is no gold-standard method for detecting non-additivity. Virtually every method available including the two presented here assess

non-additivity under additional modeling assumptions. The degree to which any of these methods succeed at detecting non-additivity is related directly to the form of the non-additivity present in a given data set. As such the two methods we present should be regarded as two additional tools in the non-additivity detection utility belt. Rather than supplant existing methods we have striven to provide new methods to assess latent group-based non-additivity.

4.1 Future research.

While the work presented in this dissertation is an exciting first look at latent group-based interaction there are numerous possible directions for the research to go from here. Simulation results could be conducted to compare these two methods to competitors besides Tukey's 1 degree of freedom test. However not all methods work on experiments of any size. For instance, the Mandel model (1.4) is not available unless there are at least $a, b > 2$ levels in the experiment. Further work is also needed to achieve confidence intervals with $1 - \alpha$ coverage regardless of σ^2.

With any Bayesian analysis, investigation of the choice and impact of the prior distribution is of interest. In our case we attempted to use vague proper priors and Empirical Bayes MLE driven priors. However many other priors can be investigated in this problem.

Another way to view the mixture problem is to test for the presence of a mixture at each treatment level individually rather than as one multivariate mixture model. For instance in the culinary example shown in Figure 3.1 there may be mixtures at some treatment levels but not others. Perhaps there is no mixture for standard training or workbook

training (coded 1 and 3 in the figure), but there are mixtures for computer training and no training (coded 2 and 4). That is, maybe some subset of students perform differently than other for computer training but all the students perform similarly on average for workbook training. The current method assumes that if there are indeed $g = 2$ groups the mixture manifests itself across all treatment levels. To explore this notion one could use a univariate EM algorithm to fit mixture models and obtain an approximation to the marginal strategy at each treatment level. These algorithms may avoid certain numerical difficulties since typically univariate optimization and numerical integration is simpler than the multivariate counterparts.

Another area of interest may be to find the class of experiment sizes for which both ACMIF and BF are effective and compare the performance of the two methods directly. The original intent was for ACMIF to be a method to handle smaller experiments and BF to handle larger experiments in terms of b. ACMIF is less effective in larger experiments due to computation time, and BF would seem to be less effective in smaller experiments due to the nature of the underlying mixture model theory. Perhaps for some b we can compare the two methods directly .

REFERENCES

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control 19*, 716–723.

Alin, A. and S. Kurt (2006). Testing non-additivity (interaction) in two-way anova tables with no replication. *Statistical Methods in Medical Research 15*, 63–85.

Barker, C., L. Stefanski, and J. Osborne (2009). The orthogonal interactions model for unreplicated factorial experiments, unpublished manuscript.

Brooks, B., F. Rose, E. Attree, and A. Elliot-Square (2002). An evaluation of the efficacy of training people with learning disabilities in a virtual environment. *Disability and Rehabilitation 24*, 622–626.

Casella, G. and R. Berger (2002). *Statistical Inference, Second Edition.*

Celeux, G., F. Forbes, C. Robert, and D. Titterington (2006). Deviance information criterion for missing data models. *Bayesian Analysis 1*, 651–674.

Daniel, W. (2005). *Biostatistics: A Foundation for Analysis in the Health Sciences Eighth edition.*

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society 39*, 1–38.

Efron, B. and R. Tibshirani (1993). *An introduction to the bootstrap.*

Garel, B. (1996). Recent asymtotic results in testing for mixtures. *Computational Statistics and Data Analysis 51*, 5295–5304.

Ghosh, J. and P. Sen (1984). On the asymptotic performance of the log likelohood ratio statistic for the mixture model and related results. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, vol. II Wadsworth, Montery, 789806.*.

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

Hirotsu, C. (1982). An approach to defining the pattern of interaction effects in a two-way layout. *Annals of the Institute of Statistical Mathematics 35*, 77–90.

Johnson, D. and F. Graybill (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 F. Graybill (1972b). Estimation of σ^2 in a two-way classification model with interaction. *Journal of the American Statistical Association 67*, 338–394.

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.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*.

Pardo, L. and M. Pardo (2005). Nonadditivity in loglinear models using ϕ-divergences and mles. *Journal of Statistical Planning and Inference 127*, 237–252.

Rao, P. (1998). *Statistical Research Methods in the Life Sciences*.

Richardson, S. and P. Green (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B (Methodological) 59*, 731–792.

Schwartz, G. (1978). Estimating the dimmension of a model. *Annals of Statistics 6*, 461–464.

Shinawi, M. and S. Wai Cheung (2008). The array cgh and its clinical applications. *Drug Discovery Today 13*, 760–770.

Spiegelhalter, D., N. Best, B. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B 64*, 583–640.

Titterington, D., A. Smith, and U. Markov (1985). *Statistical analysis of finite mixture distributions*. New York:Wiley.

Truong, V. D., Y. T. Pascua, R. Reynolds, and R. L. Thompson (2010). Evaluation of post-harvest handling practices and biochemical changes affecting french fry quality of long-term stored sweetpotatoes. *Golden LEAF funded project, Progress Report, NC State University*.

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

Tukey, J. (1955). Answers to query 113. *Biometrics 11*, 111.

Zafar-Yab (1993). On testing for non-additivity in factorial experiments. *Biometrical Journal 35*, 925–932.

APPENDICES

Appendix A

Tables and Figures

A.1 Tables

Table A.1: Simulation results for various latent group non-additive scenarios. Two treatments and four blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF is different from Tukey's test for all cases (p-values < 0.0001). Standard error for all proportions presented less than 0.0159.

$a = 2$; $b = 4$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2)$, $g_2 = (b_3, b_4)$			
		Proportion of MC rejections at $\alpha = 0.05$	
σ^2	Data type	ACMIF	Tukey
1	Inert vs. active	0.770	0.915
5		0.228	0.574
10		0.145	0.349
1	Average vs. active	0.761	0.055
5		0.270	0.048
10		0.130	0.047
1	Cancellatory	0.997	0.065
5		0.669	0.052
10		0.440	0.060

Table A.2: MSE results for four types of estimators under various latent group non-additive scenarios. Two treatments and four blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators.

$a = 2$; $b = 4$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2)$, $g_2 = (b_3, b_4)$					
		MSE			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	100.89	1.02	23.73	1.02
5		102.42	6.53	79.93	6.48
10		105.12	37.06	90.86	33.90
1	Average vs. active	100.33	1.00	24.42	1.01
5		101.44	8.45	75.95	8.16
10		104.95	30.11	93.28	27.99
1	Cancellatory	400.39	1.07	2.32	1.08
5		403.45	5.48	137.09	5.53
10		404.97	10.37	230.67	10.54

Table A.3: Squared bias results for four types of estimators under various latent group non-additive scenarios. Two treatments and four blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators.

$a = 2$; $b = 4$; $\sigma^2_B = 1$; MC reps:1000 $g_1 = (b_1, b_2)$, $g_2 = (b_3, b_4)$					
		Bias^2			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	100.40	0.00	5.17	0.00
5		100.03	0.00	58.04	0.08
10		99.95	0.64	71.08	1.27
1	Average vs. active	99.85	0.00	5.60	0.00
5		98.97	0.00	50.61	0.06
10		100.18	0.18	74.81	0.60
1	Cancellatory	399.87	0.00	0.00	0.00
5		400.74	0.00	42.05	0.00
10		399.77	0.00	123.66	0.08

Table A.4: Variance results for four types of estimators under various latent group non-additive scenarios. Two treatments and four blocks were used. First two columns describe experimental setup. Other columns report variance of the various estimators.

$a = 2$; $b = 4$; $\sigma^2_B = 1$; MC reps:1000 $g_1 = (b_1, b_2)$, $g_2 = (b_3, b_4)$					
		Variance			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	0.49	1.02	18.55	1.02
5		2.39	6.52	21.89	6.39
10		5.16	36.42	19.78	32.63
1	Average vs. active	0.48	1.00	18.82	1.01
5		2.48	8.45	25.34	8.11
10		4.77	29.94	18.48	27.39
1	Cancellatory	0.52	1.07	2.32	1.07
5		2.71	5.47	95.04	5.52
10		5.19	10.37	107.01	10.46

Table A.5: Coverage proportions for four types of intervals under various latent group non-additive scenarios. Two treatments and four blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159.

$a = 2$; $b = 4$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2)$, $g_2 = (b_3, b_4)$						
		Empirical coverage proportion of confidence interval.				
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\gamma^* \pm max(MOE)$	$\dot{\gamma}$
1	Inert vs. active	1.000	0.955	0.955	1.000	0.956
5		0.998	0.956	0.961	0.999	0.965
10		0.977	0.926	0.939	0.980	0.955
1	Average vs. active	1.000	0.950	0.950	1.000	0.952
5		0.998	0.940	0.944	0.999	0.948
10		0.969	0.933	0.937	0.979	0.958
1	Cancellatory	1.000	0.937	0.937	1.000	0.938
5		1.000	0.949	0.949	1.000	0.960
10		1.000	0.941	0.941	1.000	0.955

Table A.6: Simulation results for various latent group non-additive scenarios. Two treatments and seven blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF different from Tukey's test for all cases when estimates not equal (p-values < 0.0001). Standard error for all proportions presented less than 0.0159.

$a = 2$; $b = 7$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2, b_3)$, $g_2 = (b_4, b_5, b_6, b_7)$			
		Proportion of MC rejections at $\alpha = 0.05$	
σ^2	Data type	ACMIF	Tukey
1	Inert vs. active	1.000	1.000
5		0.729	0.991
10		0.345	0.876
1	Average vs. active	1.000	0.066
5		0.735	0.047
10		0.386	0.062
1	Cancellatory	1.000	0.043
5		1.000	0.063
10		0.961	0.053

Table A.7: MSE results for four types of estimators under various latent group non-additive scenarios. Two treatments and seven blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators.

$a = 2$; $b = 7$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3)$, $g_2 = (b_4, b_5, b_6, b_7)$					
		MSE			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	130.75	0.63	0.63	0.63
5		131.56	3.36	37.82	3.36
10		134.68	15.38	89.78	15.27
1	Average vs. active	130.47	0.69	0.69	0.69
5		133.06	3.42	37.78	3.43
10		134.17	19.06	85.74	18.86
1	Cancellatory	522.80	0.69	0.69	0.69
5		526.76	3.35	3.35	3.35
10		527.74	6.59	26.57	6.60

Table A.8: Squared bias results for four types of estimators under various latent group non-additive scenarios. Two treatments and seven blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators.

$a = 2$; $b = 7$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3)$, $g_2 = (b_4, b_5, b_6, b_7)$					
		Bias^2			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	130.46	0.00	0.00	0.00
5		130.14	0.00	8.45	0.00
10		131.85	0.04	53.37	0.06
1	Average vs. active	130.19	0.00	0.00	0.00
5		131.54	0.00	8.05	0.00
10		131.08	0.06	44.82	0.07
1	Cancellatory	522.52	0.00	0.00	0.00
5		525.33	0.00	0.00	0.00
10		524.92	0.00	0.70	0.00

Table A.9: Variance results for four types of estimators under various latent group non-additive scenarios. Two treatments and seven blocks were used. First two columns describe experimental setup. Other columns report variance of the various estimators.

$a = 2$; $b = 7$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3)$, $g_2 = (b_4, b_5, b_6, b_7)$					
		Variance			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	0.29	0.63	0.63	0.63
5		1.42	3.36	29.37	3.36
10		2.83	15.34	36.41	15.21
1	Average vs. active	0.28	0.69	0.69	0.69
5		1.51	3.42	29.73	3.43
10		3.09	19.00	40.91	18.79
1	Cancellatory	0.28	0.69	0.69	0.69
5		1.43	3.35	3.35	3.35
10		2.82	6.59	25.87	6.60

Table A.10: Coverage proportions for four types of intervals under various latent group non-additive scenarios. Two treatments and seven blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159.

$a = 2$; $b = 7$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2, b_3)$, $g_2 = (b_4, b_5, b_6, b_7)$						
		Empirical coverage proportion of confidence interval.				
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\gamma^* \pm max(MOE)$	$\dot{\gamma}$
1	Inert vs. active	0.034	0.959	0.959	1.000	0.959
5		0.242	0.937	0.725	0.787	0.937
10		0.344	0.940	0.533	0.571	0.941
1	Average vs. active	0.028	0.953	0.953	1.000	0.953
5		0.244	0.950	0.730	0.776	0.951
10		0.350	0.929	0.542	0.588	0.931
1	Cancellatory	0.000	0.948	0.948	1.000	0.948
5		0.039	0.957	0.957	1.000	0.957
10		0.127	0.959	0.923	0.964	0.960

Table A.11: Simulation results for various latent group non-additive scenarios. Two treatments and eleven blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF different from Tukey's test for all cases when estimates not equal (p-values < 0.0001). Standard error for all proportions presented less than 0.0159.

$a = 2$; $b = 11$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2, b_3, b_4, b_5)$, $g_2 = (b_6, b_7, b_8, b_9, b_{10}, b_{11})$			
		Proportion of MC rejections at $\alpha = 0.05$	
σ^2	Data type	ACMIF	Tukey
1	Inert vs. active	1.000	1.000
5		0.965	1.000
10		0.609	0.995
1	Average vs. active	1.000	0.058
5		0.972	0.047
10		0.626	0.041
1	Cancellatory	1.000	0.043
5		1.000	0.042
10		1.000	0.052

Table A.12: MSE results for four types of estimators under various latent group non-additive scenarios. Two treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators.

$a = 2$; $b = 11$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3, b_4, b_5)$, $g_2 = (b_6, b_7, b_8, b_9, b_{10}, b_{11})$					
		MSE			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	119.10	0.38	0.38	0.38
5		119.76	5.92	8.63	5.92
10		121.97	49.46	66.32	49.45
1	Average vs. active	119.01	0.38	0.38	0.38
5		120.19	3.08	6.22	3.08
10		119.91	46.92	61.71	46.91
1	Cancellatory	475.90	0.42	0.42	0.42
5		478.79	2.10	2.10	2.10
10		479.95	4.28	4.28	4.28

Table A.13: Squared bias results for four types of estimators under various latent group non-additive scenarios. Two treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators.

$a = 2$; $b = 11$; $\sigma^2_B = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3, b_4, b_5)$, $g_2 = (b_6, b_7, b_8, b_9, b_{10}, b_{11})$					
		Bias^2			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	118.91	0.00	0.00	0.00
5		118.84	0.02	0.18	0.02
10		120.10	3.75	22.85	3.75
1	Average vs. active	118.82	0.00	0.00	0.00
5		119.23	0.00	0.08	0.00
10		118.16	2.69	18.92	2.69
1	Cancellatory	475.72	0.00	0.00	0.00
5		477.88	0.00	0.00	0.00
10		478.06	0.00	0.00	0.00

Table A.14: Variance results for four types of estimators under various latent group non-additive scenarios. Two treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report variance of the various estimators.

$a = 2$; $b = 11$; $\sigma^2_B = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3, b_4, b_5)$, $g_2 = (b_6, b_7, b_8, b_9, b_{10}, b_{11})$					
		Variance			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	0.19	0.38	0.38	0.38
5		0.92	5.90	8.45	5.90
10		1.87	45.70	43.47	45.69
1	Average vs. active	0.19	0.38	0.38	0.38
5		0.96	3.08	6.14	3.08
10		1.75	44.24	42.79	44.22
1	Cancellatory	0.19	0.42	0.42	0.42
5		0.91	2.10	2.10	2.10
10		1.90	4.28	4.28	4.28

Table A.15: Coverage proportions for four types of intervals under various latent group non-additive scenarios. Two treatments and eleven blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159.

$a = 2$; $b = 11$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2, b_3, b_4, b_5)$, $g_2 = (b_6, b_7, b_8, b_9, b_{10}, b_{11})$						
		Empirical coverage proportion of confidence interval.				
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\gamma^* \pm max(MOE)$	$\dot{\gamma}$
1	Inert vs. active	0.000	0.959	0.959	1.000	0.959
5		0.002	0.951	0.919	0.96	0.951
10		0.015	0.843	0.527	0.576	0.843
1	Average vs. active	0.000	0.947	0.947	1.000	0.947
5		0.000	0.938	0.910	0.970	0.938
10		0.015	0.852	0.555	0.601	0.852
1	Cancellatory	0.000	0.940	0.940	1.000	0.940
5		0.000	0.939	0.939	1.000	0.939
10		0.000	0.947	0.947	1.000	0.947

Table A.16: Simulation results for various latent group non-additive scenarios. Three treatments and four blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF different from Tukey's test for all cases (p-values < 0.0001). Standard error for all proportions presented less than 0.0159.

$a = 3$; $b = 4$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2)$, $g_2 = (b_3, b_4)$			
		Proportion of MC rejections at $\alpha = 0.05$	
σ^2	Data type	ACMIF	Tukey
1	Inert vs. active	0.999	1.000
5		0.538	0.966
10		0.286	0.794
1	Average vs. active	0.997	0.220
5		0.541	0.205
10		0.288	0.180
1	Cancellatory	1.000	0.076
5		0.990	0.072
10		0.858	0.079
1	Accordion	1.000	0.467
5		0.997	0.378
10		0.940	0.311

A.2 Figures

Table A.17: MSE results for four types of estimators under various latent group non-additive scenarios. Three treatments and four blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators.

$a = 3$; $b = 4$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2)$, $g_2 = (b_3, b_4)$					
		MSE			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	100.81	1.02	1.12	1.02
5		103.94	7.61	50.74	7.56
10		107.23	37.02	79.98	35.27
1	Average vs. active	99.88	0.93	1.24	0.93
5		103.98	5.84	50.14	5.95
10		106.41	37.29	77.07	35.20
1	Cancellatory	400.88	1.02	1.02	1.02
5		404.50	5.05	9.15	5.06
10		404.96	9.92	65.36	9.99
1	Accordion	399.20	1.08	1.08	1.08
5		406.09	5.00	6.16	5.01
10		409.76	10.95	36.37	11.00

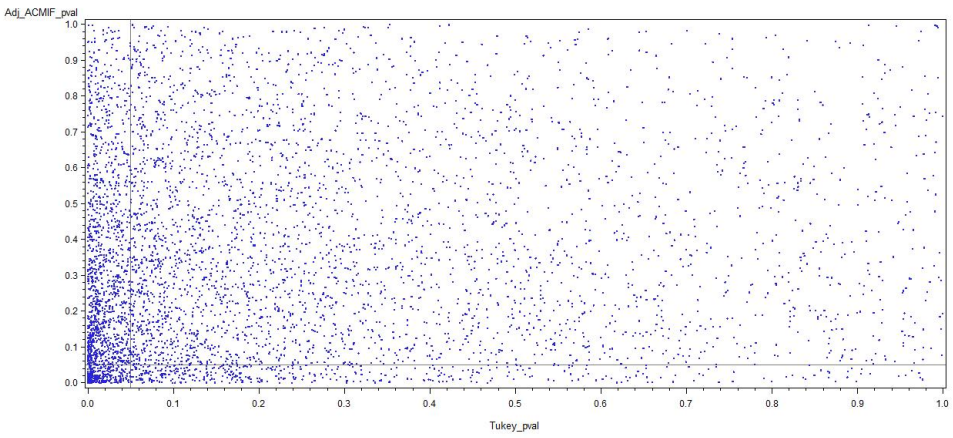

Figure A.1: Plot of Tukey p-values vs. adjusted ACMIF p-values for lymphoma example

Table A.18: Squared bias results for four types of estimators under various latent group non-additive scenarios. Three treatments and four blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators.

$a = 3$; $b = 4$; $\sigma^2_B = 1$; MC reps:1000 $g_1 = (b_1, b_2)$, $g_2 = (b_3, b_4)$					
		Bias^2			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	100.34	0.00	0.00	0.00
5		101.43	0.01	19.89	0.05
10		102.43	0.54	48.92	0.95
1	Average vs. active	99.36	0.00	0.00	0.00
5		101.73	0.03	19.79	0.07
10		101.59	0.69	47.41	1.11
1	Cancellatory	400.36	0.00	0.00	0.00
5		402.02	0.00	0.04	0.00
10		400.20	0.00	6.94	0.00
1	Accordion	398.64	0.00	0.00	0.00
5		403.47	0.02	0.03	0.02
10		404.55	0.00	1.39	0.01

Table A.19: Variance results for four types of estimators under various latent group non-additive scenarios. Three treatments and four blocks were used. First two columns describe experimental setup. Other columns report variance the various estimators.

$a = 3$; $b = 4$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2)$, $g_2 = (b_3, b_4)$					
		Variance			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	0.47	1.02	1.12	1.02
5		2.51	7.60	30.85	7.52
10		4.80	36.48	31.06	34.32
1	Average vs. active	0.53	0.93	1.24	0.93
5		2.25	5.81	30.35	5.88
10		4.82	36.60	29.66	34.09
1	Cancellatory	0.52	1.02	1.02	1.02
5		2.48	5.05	9.11	5.06
10		4.76	9.92	58.42	9.99
1	Accordion	0.55	1.08	1.08	1.08
5		2.61	4.99	6.12	4.99
10		5.20	10.94	34.98	10.99

Table A.20: Coverage proportions for four types of intervals under various latent group non-additive scenarios. Three treatments and four blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159.

$a = 3$; $b = 4$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2)$, $g_2 = (b_3, b_4)$						
		Empirical coverage proportion of confidence interval.				
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\gamma^* \pm max(MOE)$	$\dot{\gamma}$
1	Inert vs. active	0.529	0.955	0.954	0.999	0.955
5		0.599	0.947	0.752	0.800	0.953
10		0.646	0.909	0.694	0.765	0.935
1	Average vs. active	0.564	0.945	0.942	0.997	0.945
5		0.593	0.953	0.743	0.790	0.955
10		0.636	0.906	0.701	0.780	0.935
1	Cancellatory	0.504	0.956	0.956	1.000	0.956
5		0.547	0.944	0.937	0.993	0.946
10		0.581	0.950	0.882	0.932	0.957
1	Accordion	0.998	0.933	0.933	1.000	0.933
5		0.927	0.948	0.947	0.999	0.951
10		0.855	0.950	0.930	0.980	0.957

Table A.21: Simulation results for various latent group non-additive scenarios. Three treatments and eleven blocks were used. First two columns describe experimental setup. Third and fourth column provide empirical power for ACMIF and Tukey's test at $\alpha = 0.05$. McNemar's test indicates ACMIF different from Tukey's test for all cases when estimates not equal (p-values < 0.0001). Standard error for all proportions presented less than 0.0159.

$a = 3$; $b = 11$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2, b_3, b_4, b_5)$, $g_2 = (b_6, b_7, b_8, b_9, b_{10}, b_{11})$			
		Proportion of MC rejections at $\alpha = 0.05$	
σ^2	Data type	ACMIF	Tukey
1	Inert vs. active	1.000	1.000
5		1.000	1.000
10		0.911	0.999
1	Average vs. active	1.000	0.171
5		1.000	0.177
10		0.899	0.150
1	Cancellatory	1.000	0.170
5		1.000	0.121
10		1.000	0.096
1	Accordion	1.000	1.000
5		1.000	0.958
10		1.000	0.868

Table A.22: MSE results for four types of estimators under various latent group non-additive scenarios. Three treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report mean squared error for the various estimators.

$a = 3$; $b = 11$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3, b_4, b_5)$, $g_2 = (b_6, b_7, b_8, b_9, b_{10}, b_{11})$					
		MSE			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	119.06	0.39	0.39	0.39
5		120.67	6.56	6.56	6.56
10		118.40	43.18	47.62	43.18
1	Average vs. active	119.22	0.39	0.39	0.39
5		120.31	5.92	5.92	5.92
10		119.42	36.33	41.15	36.33
1	Cancellatory	475.63	0.38	0.38	0.38
5		477.45	1.80	1.80	1.80
10		480.10	5.35	5.35	5.35
1	Accordion	476.15	0.39	0.39	0.39
5		477.81	1.93	1.93	1.93
10		479.87	4.25	4.25	4.25

Table A.23: Squared bias results for four types of estimators under various latent group non-additive scenarios. Three treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report squared bias for the various estimators.

$a = 3$; $b = 11$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3, b_4, b_5)$, $g_2 = (b_6, b_7, b_8, b_9, b_{10}, b_{11})$					
		Bias^2			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	118.88	0.00	0.00	0.00
5		119.79	0.06	0.06	0.06
10		116.64	1.98	4.07	1.98
1	Average vs. active	119.03	0.00	0.00	0.00
5		119.40	0.04	0.04	0.04
10		117.60	1.38	3.41	1.38
1	Cancellatory	475.46	0.00	0.00	0.00
5		476.60	0.00	0.00	0.00
10		478.26	0.00	0.00	0.00
1	Accordion	475.96	0.00	0.00	0.00
5		476.92	0.00	0.00	0.00
10		477.97	0.00	0.00	0.00

Table A.24: Variance results for four types of estimators under various latent group non-additive scenarios. Three treatments and eleven blocks were used. First two columns describe experimental setup. Other columns report variance of the various estimators.

$a = 3$; $b = 11$; $\sigma_B^2 = 1$; MC reps:1000 $g_1 = (b_1, b_2, b_3, b_4, b_5)$, $g_2 = (b_6, b_7, b_8, b_9, b_{10}, b_{11})$					
		Variance			
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\dot{\gamma}$
1	Inert vs. active	0.18	0.39	0.39	0.39
5		0.89	6.49	6.49	6.49
10		1.76	41.20	43.55	41.20
1	Average vs. active	0.18	0.39	0.39	0.39
5		0.91	5.88	5.88	5.88
10		1.82	34.96	37.74	34.95
1	Cancellatory	0.17	0.38	0.38	0.38
5		0.85	1.80	1.80	1.80
10		1.83	5.35	5.35	5.35
1	Accordion	0.18	0.39	0.39	0.39
5		0.90	1.93	1.93	1.93
10		1.90	4.25	4.25	4.25

Table A.25: Coverage proportions for four types of intervals under various latent group non-additive scenarios. Three treatments and eleven blocks were used. First two columns describe experimental setup. Remaining columns report proportion of MC datasets for which associated confidence interval covers true parameter value. Standard error for all proportions presented less than 0.0159.

$a = 3$; $b = 11$, $\sigma_B^2 = 1$ MC reps:1000 $g_1 = (b_1, b_2, b_3, b_4, b_5)$, $g_2 = (b_6, b_7, b_8, b_9, b_{10}, b_{11})$						
		Empirical coverage proportion of confidence interval.				
σ^2	Data type	$\tilde{\gamma}$	$\breve{\gamma}$	γ^*	$\gamma^* \pm max(MOE)$	$\dot{\gamma}$
1	Inert vs. active	0.000	0.942	0.942	1.000	0.942
5		0.000	0.942	0.942	0.986	0.942
10		0.000	0.865	0.792	0.836	0.865
1	Average vs. active	0.000	0.950	0.950	1.000	0.950
5		0.000	0.934	0.934	0.99	0.934
10		0.000	0.886	0.804	0.839	0.886
1	Cancellatory	0.000	0.959	0.959	1.000	0.959
5		0.000	0.962	0.962	1.000	0.962
10		0.000	0.952	0.952	0.999	0.952
1	Accordion	0.000	0.953	0.953	1.000	0.953
5		0.000	0.957	0.957	1.000	0.957
10		0.000	0.940	0.940	1.000	0.94

Table A.26: Simulation distributional results for BF approximation and Tukey's test rejection rates in cancellatory case.

$a = 4; b = 50$; $\omega = 0.6$; Data type: Cancellatory ; MC reps: 300 ;						
σ^2	$0 < BF < \frac{1}{10}$	$\frac{1}{10} < BF < 1$	$1 < BF < 10$	$BF > 10$	$median(BF)$	Tukey RR
5	0.0000	0.0000	0.0000	1.0000	1.09×10^{15}	0.1100
10	0.0000	0.0067	0.0100	0.9833	3.9×10^{5}	0.0933
15	0.0233	0.0533	0.1233	0.8000	429.800	0.1133
20	0.0233	0.0767	0.2133	0.6867	68.386	0.0533

Table A.27: Simulation distributional results for BF approximation and Tukey's test rejection rates in inert versus active case.

$a = 4; b = 50$; $\omega = 0.6$; Data type:Inert vs. active ; MC reps: 300 ;						
σ^2	$0 < BF < \frac{1}{10}$	$\frac{1}{10} < BF < 1$	$1 < BF < 10$	$BF > 10$	$median(BF)$	Tukey RR
5	0.0000	0.0000	0.0000	1.0000	6.59×10^7	0.8200
10	0.0833	0.1333	0.2000	0.5833	28.2308	0.4200
15	0.1133	0.2700	0.2867	0.3300	3.5589	0.2000
20	0.0633	0.3833	0.3533	0.2000	1.2499	0.1866

Table A.28: Simulation distributional results for BF approximation and Tukey's test rejection rates in average versus active case.

$a = 4; b = 50$; $\omega = 0.6$; Treatment means: ; MC reps: 300 ;						
σ^2	$0 < BF < \frac{1}{10}$	$\frac{1}{10} < BF < 1$	$1 < BF < 10$	$BF > 10$	$median(BF)$	Tukey RR
5	0.0200	0.0367	0.0733	0.8700	3396.38	0.1200
10	0.0600	0.3967	0.3433	0.2000	1.38	0.0733
15	0.3200	0.4300	0.1667	0.0833	0.24	0.0733
20	0.2267	0.5767	0.1433	0.0533	0.23	0.0566

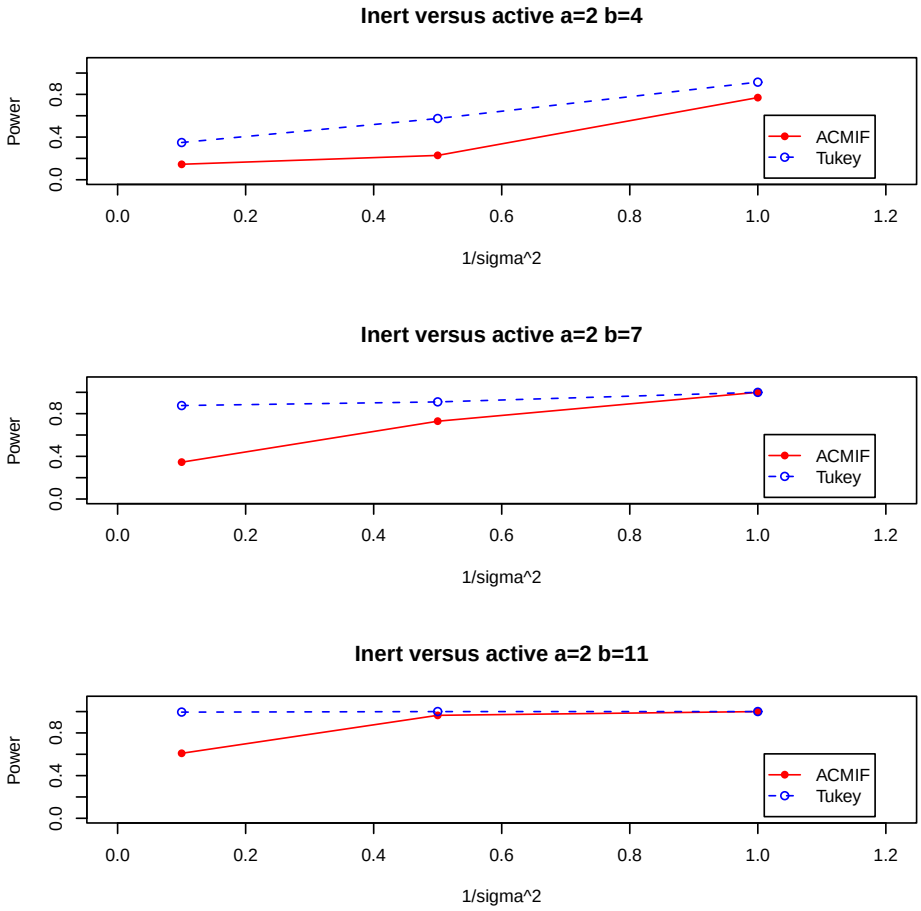

Figure A.2: Power plots for inert versus active data for both methods. $a = 2$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis.

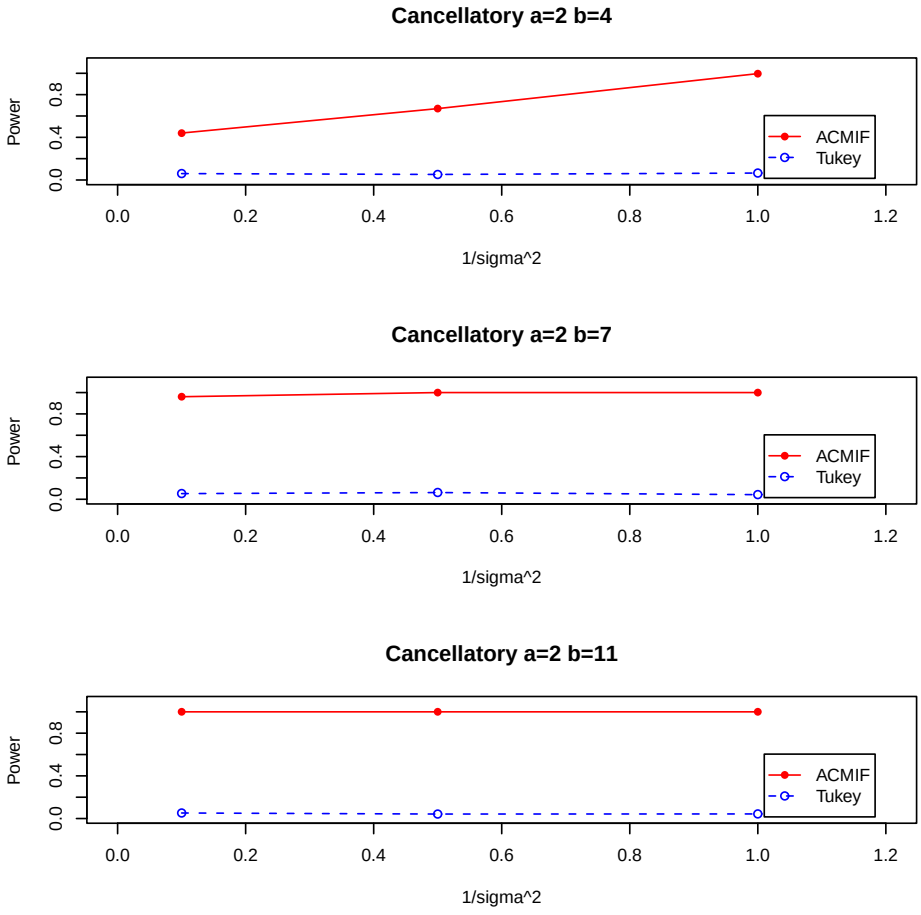

Figure A.3: Power plots for cancellatory data for both methods. $a = 2$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis.

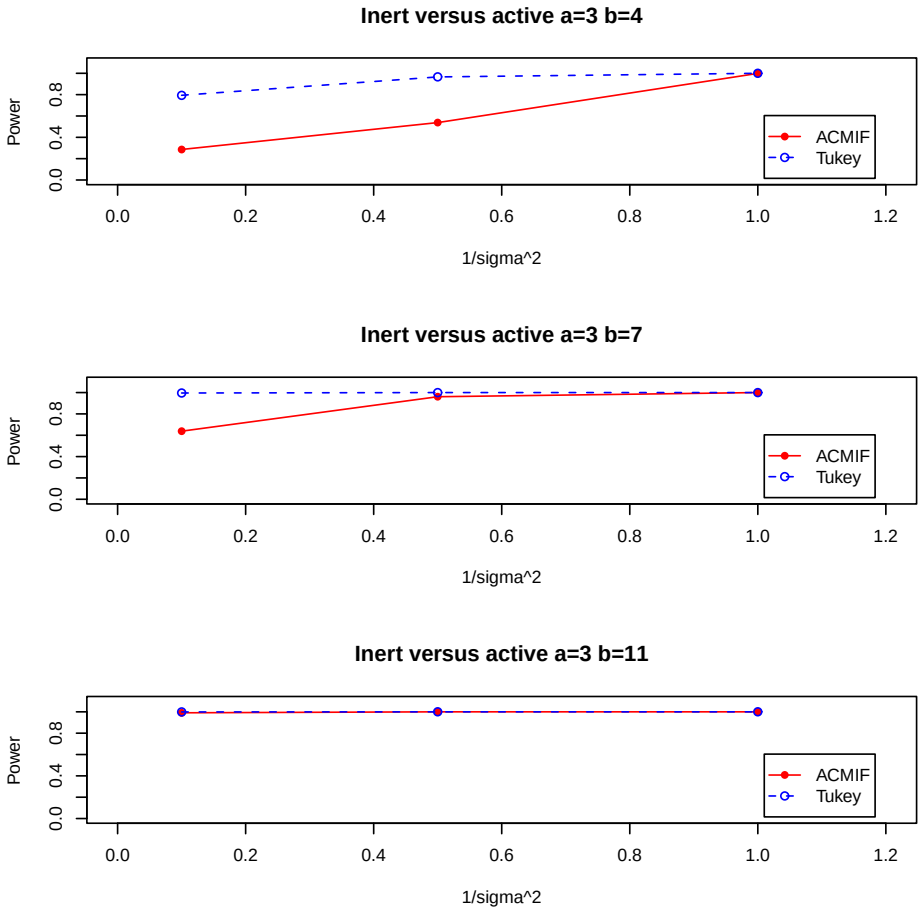

Figure A.4: Power plots for inert versus active data for both methods. $a = 3$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis.

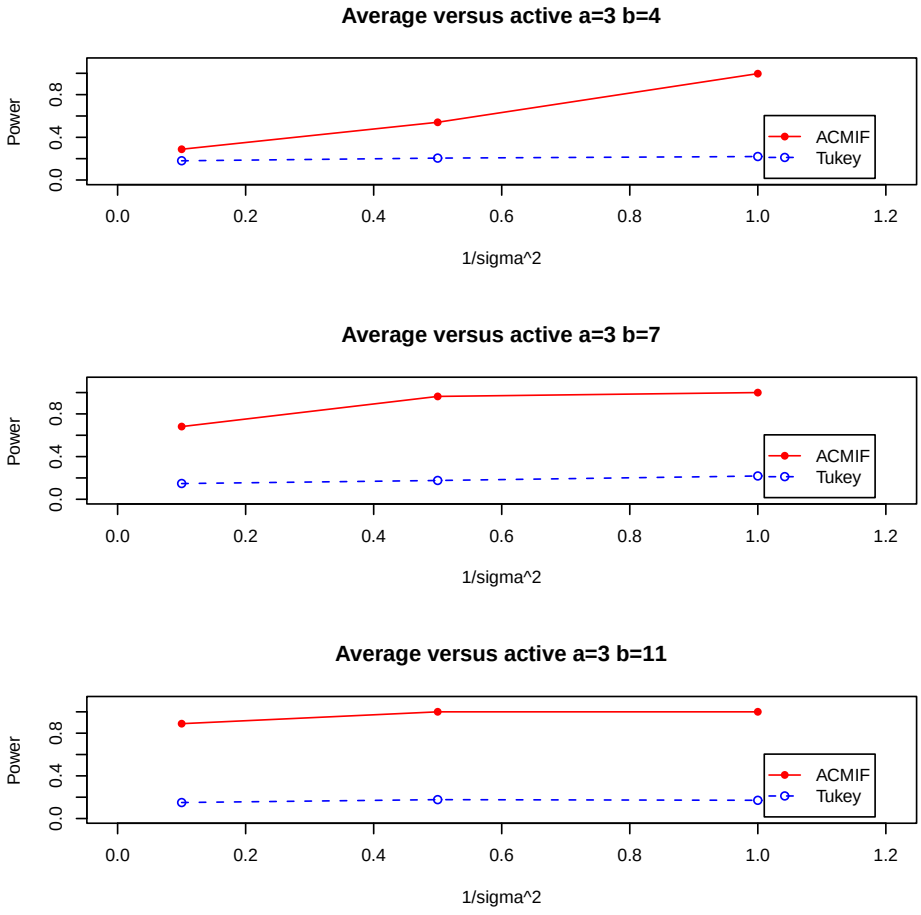

Figure A.5: Power plots for average versus active data for both methods. $a = 3$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis.

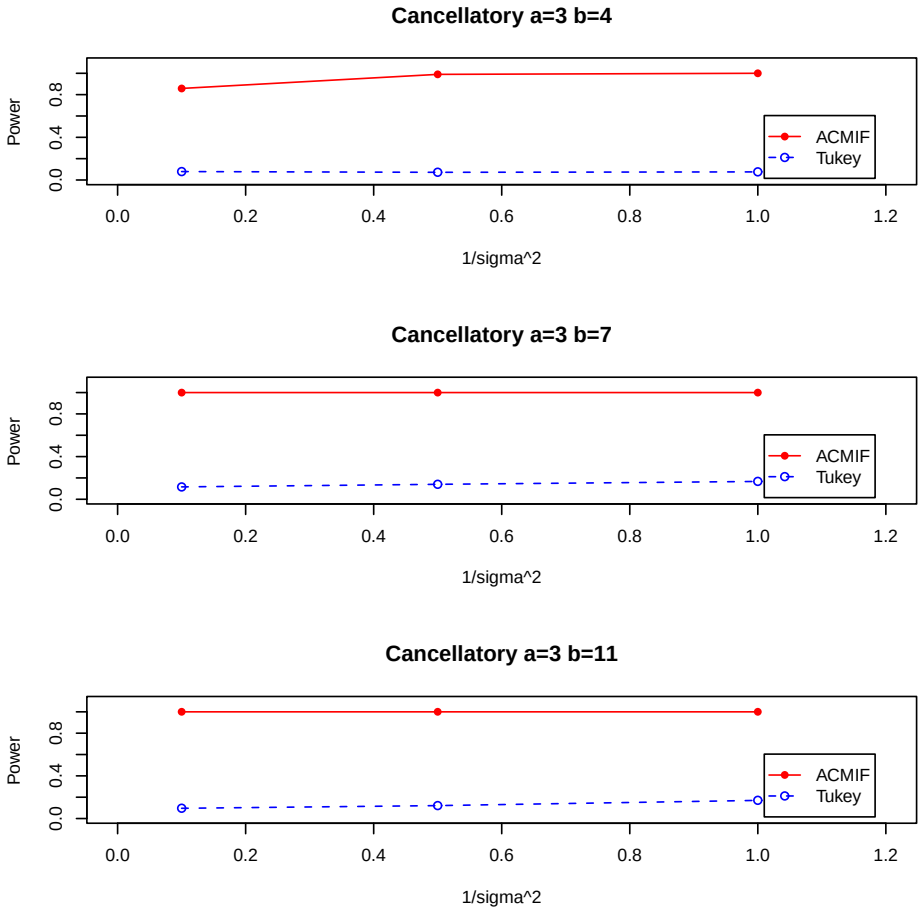

Figure A.6: Power plots for cancellatory data for both methods. $a = 3$, $b = \{4, 7, 11\}$. $\frac{1}{\sigma^2}$ plotted on x axis and empirical power at $\alpha = 0.05$ plotted on y axis.

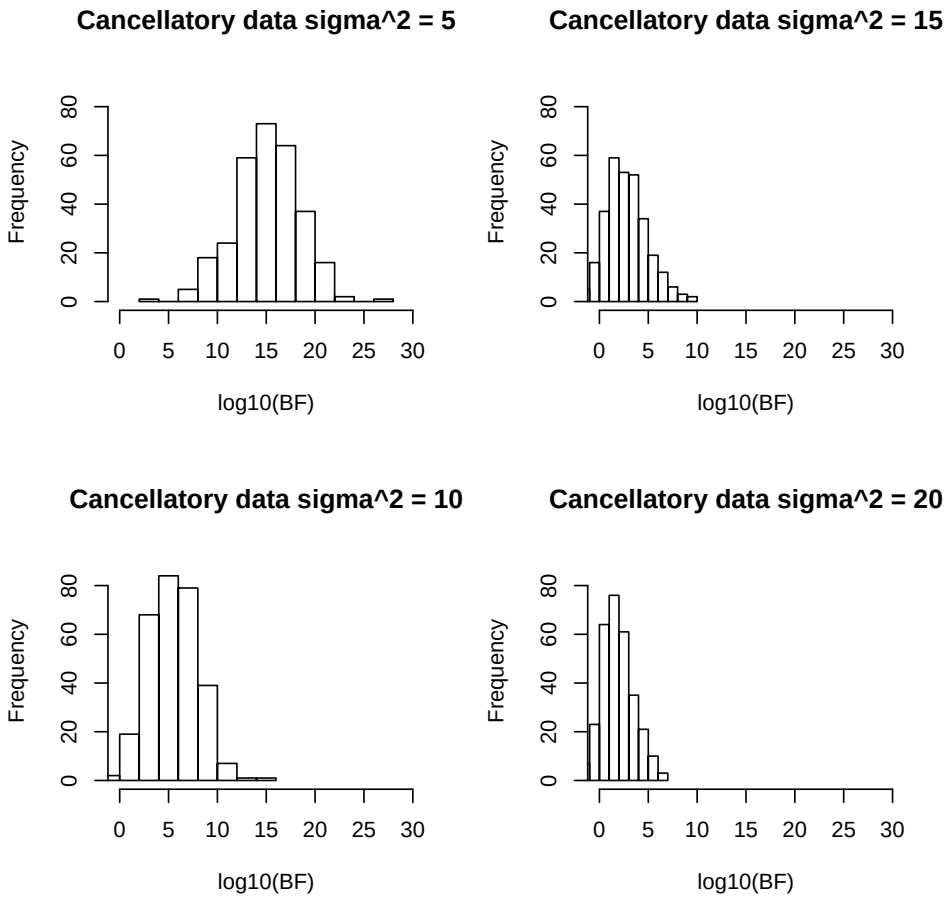

Figure A.7: Histograms of $log_{10}(BF)$ in cancellatory data for $N = 300$ MC data sets and various levels of σ^2.

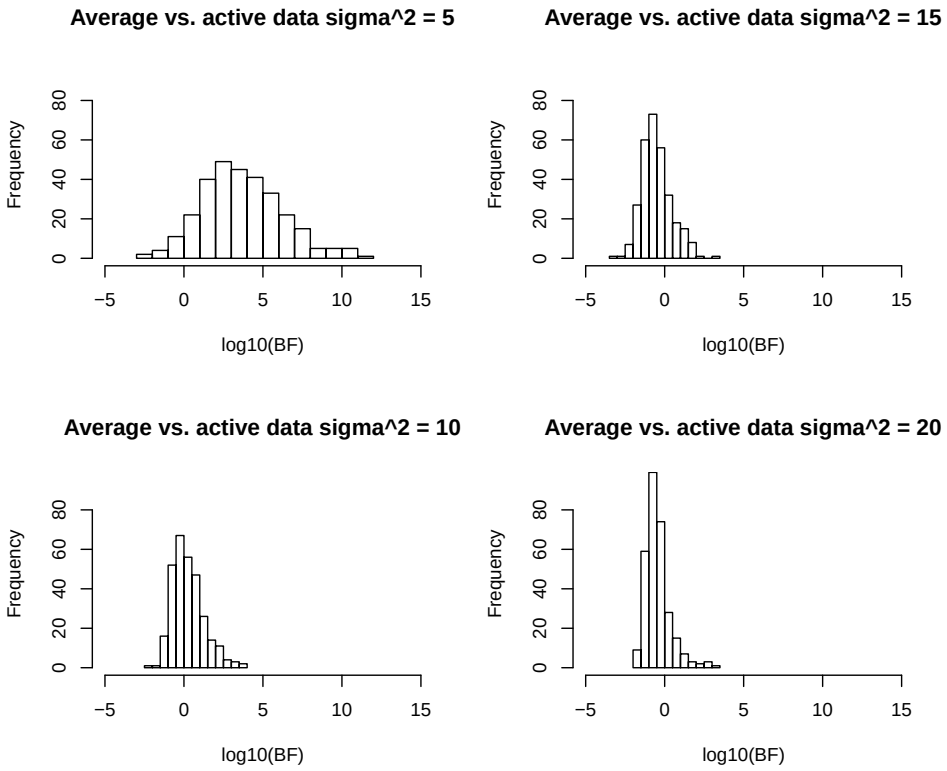

Figure A.8: Histograms of $log_{10}(BF)$ in average versus active data for $N = 300$ MC data sets and various levels of σ^2.

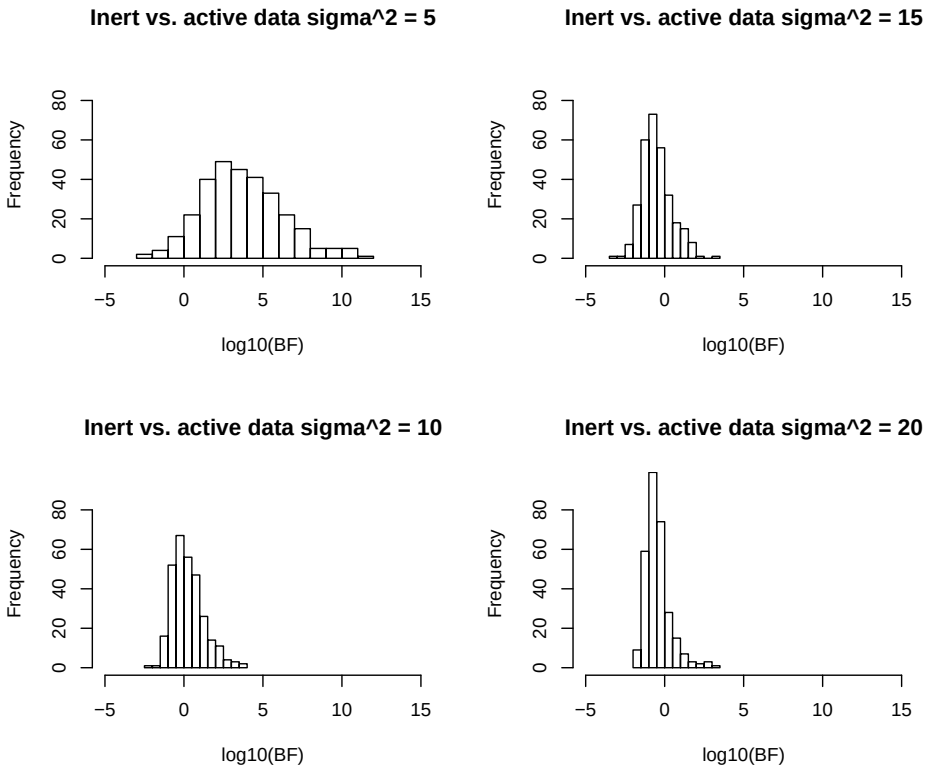

Figure A.9: Histograms of $log_{10}(BF)$ inert versus active data for $N = 300$ MC data sets and various levels of σ^2.

Appendix B

Computer Codes

B.1 SAS code

The following SAS macros were used for the simulation and illustrative example in Chapter 3.

```
%macro datagen2a4b(sigma,t11,t12,t21,t22,sigmaB);
data full;
array evec{7} e1-e7;
   do dataset=1 to &nsims;
   do block=1 to 4;
   do i=1 to 7;evec{i}=0;end;
   evec{block}=1;
   if block > 1 then evec{block+3}=1;
   else do; evec{5}=1;evec{6}=1;evec{7}=1;end;
   B=&sigmaB*rannor(&seed);
   do trt=1 to 2;
   if (trt=1) and (block in (1,2)) then teff=&t11;
   if (trt=2) and (block in (1,2)) then teff=&t12;
   if (trt=1) and (block in (3,4)) then teff=&t21;
   if (trt=2) and (block in (3,4)) then teff=&t22;
```

```
   epsi=&sigma*rannor(&seed+trt);
   y = B + 5 + teff + epsi;
   output;
   end;
   end;
   end;
   run;

data simsetup;
nsims=&nsims;
nblocks=4;
sigma=&sigma;
t11=&t11;
t12=&t12;
t21=&t21;
t22=&t22;
sigmaB=&sigmaB;
run;

ods listing;
proc print data=simsetup;
run;
ods listing close;
%mend;

%macro datagen2a7b(sigma,t11,t12,t21,t22,sigmaB);

data subfull;
   do dataset=1 to &nsims;
   do block=1 to 7;
   B=&sigmaB*rannor(&seed);
   do trt=1 to 2;
   if (trt=1) and (block in (1,2,3)) then teff=&t11;
   if (trt=2) and (block in (1,2,3)) then teff=&t12;
   if (trt=1) and (block in (4,5,6,7)) then teff=&t21;
   if (trt=2) and (block in (4,5,6,7)) then teff=&t22;
   epsi=&sigma*rannor(&seed+trt);
   y = B + 5 + teff + epsi;
   output;
   end;
   end;
```

```
   end;
run;

data full;
   set subfull;
   array evec{63} e1-e63;
   do e=1 to 63; evec{e}=0;end;
      do grp=1 to 3; *scrolls through group;
         if grp=1 then do; *distinguishes 1 environment;
            do e=1 to 7;
                  evec{e}=(block=e);
               end;
            end;
         else if grp=2 then do; *distibguishes two environments;
            e=8;
            do i=1 to 6;
            do j=(i+1) to 7;
               evec{e}=(block=i) or (block=j);  e=e+1;
            end;
            end;
          end;
         else if grp=3 then do;
 e=29;
  do i=1 to 5;
          do j=(i+1) to 6;
  do k=(j+1) to 7;
   evec{e}=(block=1)or(block=j)or(block=k); e=e+1;
   end;
   end;
   end;
   end;
      end;
run;
data simsetup;
nsims=&nsims;
nblocks=7;
sigma=&sigma;
t11=&t11;
t12=&t12;
t21=&t21;
t22=&t22;
```

```
sigmaB=&sigmaB;
run;

ods listing;
proc print data=simsetup;
run;
ods listing close;
%mend;

%macro datagen2a11b(sigma,t11,t12,t21,t22,sigmaB);

data subfull;
   do dataset=1 to &nsims;
   do block=1 to 11;
   B=&sigmaB*rannor(&seed);
   do trt=1 to 2;
   if (trt=1) and (block in (1,2,3,4,5)) then teff=&t11;
   if (trt=2) and (block in (1,2,3,4,5)) then teff=&t12;
   if (trt=1) and (block in (6,7,8,9,10,11)) then teff=&t21;
   if (trt=2) and (block in (6,7,8,9,10,11)) then teff=&t22;

   epsi=&sigma*rannor(&seed+trt);
   y = B + 5 + teff + epsi;
   output;
   end;
   end;
   end;
run;

data full;
   set subfull;
   array evec{1023} e1-e1023;
   do e=1 to 1023; evec{e}=0;end;
      do grp=1 to 5; *scrolls through group;

         if grp=1 then do; *distinguishes 1 environment;
            do a=1 to 11;
                  evec{a}=(block=a);
               end;
            end;
         else if grp=2 then do; *distibguishes two environments;
```

```
            a=12;
            do b=1 to 10;
            do c=(b+1) to 11;
               evec{a}=(block=b) or (block=c);  a=a+1;
            end;            end;          end;
           else if grp=3 then do; *distinguished three environments;
 a=67;
  do b=1 to 9;
          do c=(b+1) to 10;
  do d=(c+1) to 11;
   evec{a}=(block=b)or(block=c)or(block=d); a=a+1;
   end;    end;    end;    end;

else if grp=4 then do; *distinguish four environments;
        a=232;
 do b=1 to 8;
 do c=(b+1) to 9;
 do d=(c+1) to 10;
 do e=(d+1) to 11;
  evec{a}=(block=b)or(block=c)or(block=d)or(block=e); a=a+1;
   end;    end;    end;    end;    end;

 else if grp=5 then do; *distinguish five environments;
 a=562;
  do b=1 to 7;
  do c=(b+1) to 8;
  do d=(c+1) to 9;
  do e =(d+1) to 10;
  do f = (e+1) to 11;
evec{a}=(block=b)or(block=c)or(block=d)or(block=e)or(block=f); a=a+1;
end;end;end;end;end;end;

      end;
run;
data simsetup;
nsims=&nsims;
nblocks=11;
sigma=&sigma;
t11=&t11;
t12=&t12;
t21=&t21;
```

```
t22=&t22;
sigmaB=&sigmaB;
run;

ods listing;
proc print data=simsetup;
run;
ods listing close;
%mend;


%macro datagen3a4b(sigma,t11,t12,t21,t22,t13,t23,sigmaB);
data full;
array evec{7} e1-e7;
   do dataset=1 to &nsims;
   do block=1 to 4;
   do i=1 to 7;evec{i}=0;end;
   evec{block}=1;
   if block > 1 then evec{block+3}=1;
   else do; evec{5}=1;evec{6}=1;evec{7}=1;end;
   B=&sigmaB*rannor(&seed);
   do trt=1 to 3;
   if (trt=1) and (block in (1,2)) then teff=&t11;
   if (trt=2) and (block in (1,2)) then teff=&t12;
   if (trt=3) and (block in (1,2)) then teff=&t13;
   if (trt=1) and (block in (3,4)) then teff=&t21;
   if (trt=2) and (block in (3,4)) then teff=&t22;
   if (trt=3) and (block in (3,4)) then teff=&t23;
   epsi=&sigma*rannor(&seed+trt);
   y = B + 0 + teff + epsi;
   output;
   end;
   end;
   end;
   run;

data simsetup;
nsims=&nsims;
nblocks=4;
sigma=&sigma;
t11=&t11;
```

```
t12=&t12;
t13=&t13;
t21=&t21;
t22=&t22;
t23=&t23;
sigmaB=&sigmaB;
run;

ods listing;
proc print data=simsetup;
run;
ods listing close;
%mend;

%macro datagen3a7b(sigma,t11,t12,t13,t21,t22,t23,sigmaB);

data subfull;
   do dataset=1 to &nsims;
   do block=1 to 7;
   B=&sigmaB*rannor(&seed);
   do trt=1 to 3;
   if (trt=1) and (block in (1,2,3)) then teff=&t11;
   if (trt=2) and (block in (1,2,3)) then teff=&t12;
   if (trt=3) and (block in (1,2,3)) then teff=&t13;
   if (trt=1) and (block in (4,5,6,7)) then teff=&t21;
   if (trt=2) and (block in (4,5,6,7)) then teff=&t22;
   if (trt=3) and (block in (4,5,6,7)) then teff=&t23;
   epsi=&sigma*rannor(&seed+trt);
   y = B + 5 + teff + epsi;
   output;
   end;
   end;
   end;
run;

data full;
   set subfull;
   array evec{63} e1-e63;
   do e=1 to 63; evec{e}=0;end;
      do grp=1 to 3; *scrolls through group;
         if grp=1 then do; *distinguishes 1 environment;
```

```
            do e=1 to 7;
                  evec{e}=(block=e);
               end;
            end;
         else if grp=2 then do; *distibguishes two environments;
            e=8;
            do i=1 to 6;
            do j=(i+1) to 7;
               evec{e}=(block=i) or (block=j);  e=e+1;
            end;
            end;
          end;
         else if grp=3 then do;
 e=29;
  do i=1 to 5;
          do j=(i+1) to 6;
  do k=(j+1) to 7;
   evec{e}=(block=1)or(block=j)or(block=k); e=e+1;
   end;
   end;
   end;
   end;
      end;
run;
data simsetup;
nsims=&nsims;
nblocks=7;
sigma=&sigma;
t11=&t11;
t12=&t12;
t13=&t13;
t21=&t21;
t22=&t22;
t23=&t23;
sigmaB=&sigmaB;
run;

ods listing;
proc print data=simsetup;
run;
ods listing close;
```

```
%mend;


%macro datagen3a11b(sigma,t11,t12,t13,t21,t22,t23,sigmaB);

data subfull;
   do dataset=1 to &nsims;
   do block=1 to 11;
   B=&sigmaB*rannor(&seed);
   do trt=1 to 3;
   if (trt=1) and (block in (1,2,3,4,5)) then teff=&t11;
   if (trt=2) and (block in (1,2,3,4,5)) then teff=&t12;
   if (trt=3) and (block in (1,2,3,4,5)) then teff=&t13;
   if (trt=1) and (block in (6,7,8,9,10,11)) then teff=&t21;
   if (trt=2) and (block in (6,7,8,9,10,11)) then teff=&t22;
   if (trt=3) and (block in (6,7,8,9,10,11)) then teff=&t23;
   epsi=&sigma*rannor(&seed+trt);
   y = B + 5 + teff + epsi;
   output;
   end;
   end;
   end;
run;

data full;
   set subfull;
   array evec{1023} e1-e1023;
   do e=1 to 1023; evec{e}=0;end;
      do grp=1 to 5; *scrolls through group;

         if grp=1 then do; *distinguishes 1 environment;
            do a=1 to 11;
                  evec{a}=(block=a);
               end;
            end;
         else if grp=2 then do; *distibguishes two environments;
            a=12;
            do b=1 to 10;
            do c=(b+1) to 11;
               evec{a}=(block=b) or (block=c);  a=a+1;
            end;            end;          end;
```

```
          else if grp=3 then do; *distinguished three environments;
 a=67;
  do b=1 to 9;
         do c=(b+1) to 10;
  do d=(c+1) to 11;
   evec{a}=(block=b)or(block=c)or(block=d); a=a+1;
   end;    end;    end;    end;

else if grp=4 then do; *distinguish four environments;
        a=232;
 do b=1 to 8;
 do c=(b+1) to 9;
 do d=(c+1) to 10;
 do e=(d+1) to 11;
  evec{a}=(block=b)or(block=c)or(block=d)or(block=e); a=a+1;
   end;    end;    end;    end;    end;

 else if grp=5 then do; *distinguish five environments;
 a=562;
  do b=1 to 7;
  do c=(b+1) to 8;
  do d=(c+1) to 9;
  do e =(d+1) to 10;
  do f = (e+1) to 11;
evec{a}=(block=b)or(block=c)or(block=d)or(block=e)or(block=f); a=a+1;
end;end;end;end;end;end;

      end;
run;
data simsetup;
nsims=&nsims;
nblocks=11;
sigma=&sigma;
t11=&t11;
t12=&t12;
t13=&t13;
t21=&t21;
t22=&t22;
t23=&t23;
sigmaB=&sigmaB;
run;
```

```
ods listing;
proc print data=simsetup;
run;
ods listing close;
%mend;

%macro intcont_mix;
ods listing close;
proc means data=full max;
var block;
ods output summary=blocknum;
run;

data blocknum;
set blocknum;
c = 2**(block_max-1)-1;
call symput('c',c);
run;

proc glm data=full;
   by dataset;
   class block trt;
   model y=block trt/clparm;
   estimate "linear_add" trt -1 1 0;
   ods output estimates=addests overallanova=adddf;
run;

data addests;
set addests;
addlower=lowercl;
addupper=uppercl;
run;

data adddfe;
set adddf;
if source ne 'Error' then delete;
adddfe=df;
keep dataset adddfe;
run;
```

```
%do i=1 %to &c;

proc glm data=full;
by dataset;
   class e&i trt block;
   model y=e&i|trt block(e&i)/clparm;
   estimate "t3vt1_env1" trt -1 1 0 e&i*trt 0 0 0 -1 1 0; *Double check;
     ods output overallanova=configMSE&i modelanova=F&i estimates=grpests&i.;
run;quit;

data configMSE&i;
set  configMSE&i;
if source ne 'Error' then delete;
config=&i;
dfgrp=df;
keep dataset config ms dfgrp;
run;

data F&i;
set  F&i;
if source ne "e&i*trt" then delete;
if hypothesistype ne 3 then delete;
config=&i;
run;


data grpests&i;
set grpests&i;
config=&i;
grpupper=uppercl;
grplower=lowercl;
run;

proc append base=configmseraw data = configmse&i; run;
proc append base=Fraw data=F&i;run;
proc append base=grpestsraw data=grpests&i;run;



*clean out old data sets;
proc datasets;
```

```
   delete  configmse&i F&i grpest&i;
run;quit;
%end;

data configMSE;
set configMSEraw;
run;

data F;
set Fraw;
run;

data grpests;
set grpestsraw;
run;

proc datasets;
delete  configmseraw Fraw grpestsraw;
run;quit;

proc glm data=full;
by dataset;
   class trt block;
   model y=trt block;
   ods output overallanova=RCBDMSE;
run;quit;

data rcbdmse;
set  rcbdmse;
if source ne "Error" then delete;
RCBDMSE = MS;
keep RCBDMSE dataset;
run;

proc sort data=configmse;
by dataset config;
run;
proc sort data=F;
by dataset config;
run;
```

```
proc means data=configmse var;
class dataset;
var ms;
ods output summary=VarMSE;
run;

proc means data=F max;
class dataset;
var Fvalue;
ods output summary=Fmax;
run;


*Determine which configuration was used;
data findconfig;
merge F Fmax;
by dataset;
if Fvalue_max ne fvalue then delete;
sigconfig=config;
keep sigconfig dataset;
run;

*Tie the used configuration to the grpest set;
proc sort data=grpests; by dataset;
run;

data bestgrpest;
merge grpests findconfig;
by dataset;
if sigconfig ne config then delete;
run;

proc means data=F min;
class dataset;
var ProbF;
ods output summary=Pmin;
run;

data varspread;
merge RcbdMSE VarMSE Fmax Pmin;
```

```
by dataset;
CVsq=MS_var/(RCBDMSE**2);
CV=sqrt(CVsq);
Fmax=Fvalue_max;
Pmin=Probf_min;
drop Fvalue_max Probf_min;
if Pmin < (0.05/&c) then Bonreject=1; else Bonreject=0;
run;

ods listing;
proc means data=varspread mean stderr;
var Bonreject;
ods output summary=disp2;
run;
ods listing close;


%mend;

%macro tukeytest;
ods listing close;
proc glm data=full;
by dataset;
class trt block;
model y = trt block;
output out=predset p=p;
run;
quit;

proc glm data=predset;
by dataset;
class trt block;
model y= trt block p*p;
ods output overallanova=TukMSE modelanova=tests;
run;
quit;

data tests_clean;
set tests;
if (hypothesistype=3) and (source ='p*p');
if probf<0.05 then tukreject =1; else tukreject=0;
```

```
run;

data tukmse;
set tukmse;
if source ne 'Error' then delete;
keep dataset MS;
run;
/*
data tests_chisq;
set tests_clean;
chisq=ss/(&sigma**2);
tukeysig_known=(chisq>cinv(.95,1));
run;

ods listing;
proc means data=tests_chisq;
title "Tukeys rejection rate for known sigma^2";
var tukeysig_known;
run;

proc means data=tukmse;
title 'MSE from tukey model';
var MS;
run;
*/

ods listing;
proc means data=tests_clean mean stderr;
title 'Tukey Method' ;
var tukreject;
output out= Tukey_reject;
ods output summary=disp3;
run;
title;
%mend;
%macro mcnemartest;
data mcset;
merge varspread tests_clean;
by dataset;
run;
```

```
ods trace on;
proc freq data=mcset;
tables tukreject*bonreject/agree;
ods output McNemarsTest=Mcnemar_comp;
run;

data Mcnemar_comp;
set Mcnemar_comp;
if _N_ ne 3 then delete;
keep nvalue1;run;
%mend;
%macro homebrew;
data mcset_hb;
merge varspread tests_clean;
by dataset;
if (bonreject=0)&(tukreject=1) then b=1; else b=0;
if (bonreject=1)&(tukreject=0) then c=1; else c=0;
run;

proc means data=mcset_hb sum;
var b c;
ods output summary=mcset_stats;
run;

data mcset_stats;
set mcset_stats;
if ((b_sum + c_sum) ne 0) then S = (abs(b_sum-c_sum) )**2 /(b_sum + c_sum);
else S=0;
mcp=1-probchi(S,1,0);
run;
%mend;
%macro gather;
data display;
merge simsetup disp2 disp3  covers;
run;
proc append base=table data=display;
run;
%mend;

%macro addtarget(sigma,t11,t12,t21,t22,t13,t23,sigmaB);
data target;
```

```
truval=&t12-&t11;
run;

data bonstrip;
set varspread;
keep dataset bonreject;
run;

data bestgrpest;
set  bestgrpest;
stderrgrp=stderr;
grpestimate=estimate;
run;
data addests;
set  addests;
stderradd=stderr;
addestimate=estimate;
run;

data adddfe;
set adddfe;
if _n_ ne 1 then delete;
keep adddfe;
call symput('adddfe',adddfe);
run;

data grpdf;
set configmse;
if _n_ ne 1 then delete;
keep dfgrp;
call symput('grpdf',dfgrp);
run;

data addtarget;
merge bestgrpest addests bonstrip pmin;
by dataset;
groupdf=&grpdf;
adddf=&adddfe;
truval=&t12-&t11;
addlength=addupper-addlower;
grplength=grpupper-grplower;
```

```
MOEadd =tinv(.975,&adddfe)*stderradd;
MOEgrp =tinv(.975,&grpdf)*stderrgrp;
MOE = max(MOEadd,MOEgrp);
gammadots=addestimate*(probf_min) + grpestimate*(1-probf_min);
if bonreject=1 then gammastar=grpestimate; else gammastar=addestimate;
sedots=sqrt((probf_min*stderradd)**2 + ((1-probf_min)*stderrgrp)**2 +
2*probf_min*(1-probf_min)*(stderradd*stderrgrp)**2);
dotlower = gammadots - tinv(.975,&grpdf)*sedots;
dotupper = gammadots + tinv(.975,&grpdf)*sedots;
if bonreject=1 then uselower=grplower;
if bonreject=1 then useupper=grpupper;
if bonreject=0 then uselower=addlower;
if bonreject=0 then useupper=addupper;
if bonreject=1 then widelower=grpestimate - MOE;
if bonreject=1 then wideupper=grpestimate + MOE;
if bonreject=0 then widelower=addestimate - MOE;
if bonreject=0 then wideupper=addestimate + MOE;
if (truval>grplower)and(truval<grpupper) then grpcover=1; else
grpcover=0;
if (truval>uselower)and(truval<useupper) then cover=1; else cover=0;
if (truval>widelower)and(truval<wideupper) then widecover=1; else
widecover=0;
if (truval>addlower)and(truval<addupper) then addcover=1; else
addcover=0;
if (truval>dotlower)and(truval<dotupper) then dotcover=1; else
dotcover=0;
sqerror_add =(addestimate-truval)**2;
sqerror_grp =(grpestimate-truval)**2;
sqerror_star=(gammastar-truval)**2;
sqerror_dots=(gammadots-truval)**2;
run;

proc means data=addtarget;
var cover widecover addcover grpcover dotcover sqerror_add sqerror_grp
sqerror_star sqerror_dots;
ods output summary=covers;
run;
%mend;
```

B.2 R code

The following R code is an example of the code used for the simulations in Chapter 4. Inputs can be changed to get additional scenarios.

```
library(mvtnorm)

numsets=300

N=50
n<-N
mu11=0
mu12=4
mu13=0
mu14=4
mu21=4
mu22=0
mu23=4
mu24=0
w=.6
sigsq=10

NSAMP<-1000 #number of draws to approximate marginals - importance sampling;
B<-52 #number of bootstrap samples to get SEs

#Starting values

mu110=.5
mu120=6
mu130=-1
mu140=5
mu210=3.5
mu220=.25
mu230=4.4
mu240=.1
w0=1/2
sigsq0=15
```

```
m1NB<-rep(-99,numsets)
m2NB<-rep(-99,numsets)

m2NBNG<-rep(-99,numsets)
m1NBNG<-rep(-99,numsets)

############################################################
###########Define the data generation function###########
############################################################
Makedat<-function(seed,numsets,N,mu11,mu12,mu13,mu14,mu21,mu22,mu23
,mu24,w,sigsq){
#initialize data storage

Y<-array(0,c(N,4,numsets))
#Create appropriate covariance matrix
SIG<-diag(sigsq,4)

parmvect<-matrix(c(w,mu11,mu12,mu13,mu14,mu21,mu22,mu23,mu24,sigsq),
10,1)
parmmat<-matrix(rep(parmvect,numsets),10,numsets,byrow=FALSE)

#generate 'numset' datasets
for(i in 1:numsets){
comp1<-rbinom(1,N,w)
comp2<-N-comp1

if(comp1 == 0){
Y[,,i]<-rmvnorm(comp2,c(mu21,mu22,mu23,mu24),SIG)}

if(comp1 == N){
Y[,,i]<-rmvnorm(comp1,c(mu11,mu12,mu13,mu14),SIG)}

if((comp1 != N)&&(comp1 != 0)){
Y1<-rmvnorm(comp1,c(mu11,mu12,mu13,mu14),SIG)
Y2<-rmvnorm(comp2,c(mu21,mu22,mu23,mu24),SIG)
Y[,,i]<-rbind(Y1,Y2)
}}
return(Y)
}

############################################################
```

```
#Write function 'optim' will maximize:1 comp  NO  blocks#
#########################################################
OnecompNB<-function(parm,Yset){
#if Y is supplied in a matrix, convert it to an array to ensure
compatibility
if(is.matrix(Yset)==TRUE){
Yset<-array(Yset,c(dim(Yset)[1],dim(Yset)[2],1))}
#get loop limits from Y array
N<-dim(Yset)[1]
#initialize storage
logdens<-c(0,N)
for(i in 1:N){ #open data points loop
logdens[i]<-dmvnorm(Yset[i,,],parm[1:4],diag(parm[5],4),log=TRUE)
}#end data points loop
loglike<- -sum(logdens)
return(loglike)
}#close Onecomp function

#########################################################
##Define the multivariate EM function  NO  blocks########
#########################################################

EMNB<-function(seed,EMMC,Y,mu110,mu120,mu130,mu140,mu210,mu220,
mu230,mu240,w0,sigsq0){
#if Y is supplied in a matrix, convert it to an array to ensure
compatibility
if(is.matrix(Y)==TRUE){
Y<-array(Y,c(dim(Y)[1],dim(Y)[2],1))}
#get loop limits from Y array
N<-dim(Y)[1]
numsets<-dim(Y)[3]
#construct initial covariance matrix (NO compound symmetry)
SIG0<-diag(sigsq0,4)
####Initialize convergence storage##########
w1S<-rep(0,EMMC)
w2S<-rep(0,EMMC)
tau1S<-matrix(0,EMMC,N)
tau2S<-matrix(0,EMMC,N)
mu1S<-matrix(0,EMMC,4)
mu2S<-matrix(0,EMMC,4)
SIG1S<-array(0,c(4,4,EMMC))
```

```
SIG2S<-array(0,c(4,4,EMMC))
SIGS <-array(0,c(4,4,EMMC))
####Initialize numset estimate storage
w1hat<-rep(0,numsets)
w2hat<-rep(0,numsets)
mu1hat<-matrix(0,numsets,4)
mu2hat<-matrix(0,numsets,4)
sigsqhat<-rep(0,numsets)
for(g in 1:numsets){  #begin numsets loop
#Define initial values as current
w1cur<-w0
w2cur<- 1-w1cur
mu1cur<-c(mu110,mu120,mu130,mu140)
mu2cur<-c(mu210,mu220,mu230,mu240)
sigsqcur<-sigsq0
SIGcur<-diag(sigsqcur,4)
for(i in 1:EMMC){
##Define vector of probability estimates
#taui is the current guess for component membership probability
for the entire data set
tau1<- (w1cur*dmvnorm(Y[,,g],mu1cur,SIGcur)) / (w1cur*dmvnorm(Y[,,g]
,mu1cur,SIGcur) +w2cur*dmvnorm(Y[,,g],mu2cur,SIGcur))
tau2<- (w2cur*dmvnorm(Y[,,g],mu2cur,SIGcur)) / (w1cur*dmvnorm(Y[,,g]
,mu1cur,SIGcur) +w2cur*dmvnorm(Y[,,g],mu2cur,SIGcur))
w1next<-mean(tau1)
w2next<-mean(tau2)
#Prepare next Mus
tau1num<-matrix(0,N,4)
tau2num<-matrix(0,N,4)
for(j in 1:N){
tau1num[j,]<-tau1[j]*Y[j,,g]
tau2num[j,]<-tau2[j]*Y[j,,g]}
mu1next<-apply(tau1num,2,sum)/sum(tau1)
mu2next<-apply(tau2num,2,sum)/sum(tau2)
#put mus in matrix for subtraction from Y
mu1nextmat<-matrix(rep(mu1next,N),N,4,byrow=TRUE)
mu2nextmat<-matrix(rep(mu2next,N),N,4,byrow=TRUE)
centY1<-Y[,,g]-mu1nextmat
centY2<-Y[,,g]-mu2nextmat
cent1array<-array(0,c(4,4,N))
cent2array<-array(0,c(4,4,N))
```

```
for(j in 1:N){
cent1array[,,j]<-tau1[j]*matrix(centY1[j,],4,1)%*%matrix(centY1[j,]
,1,4)
cent2array[,,j]<-tau2[j]*matrix(centY2[j,],4,1)%*%matrix(centY2[j,]
,1,4)}
C1comp1list<-rep(0,N)
C1comp2list<-rep(0,N)
C2comp1list<-rep(0,N)
C2comp2list<-rep(0,N)
#Compute c1 constant for all stuff - related to variance - check with
 McLachlan
for(k in 1:N){
C1comp1list[k]<-t(centY1[k,])%*%centY1[k,]
C1comp2list[k]<-t(centY2[k,])%*%centY2[k,]
C2comp1list[k]<-t(centY1[k,])%*%matrix(1,4,4)%*%centY1[k,]
C2comp2list[k]<-t(centY2[k,])%*%matrix(1,4,4)%*%centY2[k,]}
T<-sum(c(tau1,tau2))
C1<-sum(c(tau1*C1comp1list,tau2*C1comp2list))
C2<-sum(c(tau1*C2comp1list,tau2*C2comp2list))
sigsqnext<-(1/12)*( (4*C1-C2)/T )
SIGnext<-diag(sigsqnext,4)
#SIG1next<-apply(cent1array,c(1,2),sum)/sum(tau1)
#SIG2next<-apply(cent2array,c(1,2),sum)/sum(tau2)
#SIGnext<-(sum(tau1)*SIG1next + sum(tau2)*SIG1next)/N
#Calculate 'current' and 'next' likelihoods
curloglike<- sum(log((w1cur*dmvnorm(Y[,,g],mu1cur,SIGcur)) + (w2cur*
dmvnorm(Y[,,g],mu2cur,SIGcur))))
nextloglike<-sum(log((w1next*dmvnorm(Y[,,g],mu1next,SIGnext)) + (w2next
*dmvnorm(Y[,,g],mu2next,SIGnext))))
numits<-i
w1cur     <-w1next
w2cur     <-w2next
mu1cur    <-mu1next
mu2cur    <-mu2next
SIGcur    <-SIGnext
sigsqcur  <-sigsqnext
w1S[i]<-w1cur
w2S[i]<-w2cur
tau1S[i,]<-tau1
tau2S[i,]<-tau2
mu1S[i,]<-mu1cur
```

```
mu2S[i,]<-mu2cur
#SIG1S[,,i]<-SIG1cur
#SIG2S[,,i]<-SIG2cur
#SIGS[,,i]<-SIGcur
if((nextloglike-curloglike)<1e-100){
break}
} #end EM loop
w1hat[g]    <-w1cur
w2hat[g]    <-1-w1cur
tau1hat  <-tau1
tau2hat  <-tau2
mu1hat[g,]   <-mu1cur
mu2hat[g,]   <-mu2cur
SIGhat   <-SIGcur
sigsqhat[g] <-sigsqcur

} #end numsets loop
return(list(w1hat,mu1hat,mu2hat,sigsqhat))
} #end EM function

###############Write a function for inverse gamma density
densigamma<-function(x,shape,scale){
exp( shape*log(scale) - lgamma(shape) + (-shape-1)*log(x) -
scale/x )}

###############Write a function for two component density#####
dens2norm<-function(Y,mean1,mean2,sigma,omega){
omega*dmvnorm(Y,mean1,sigma) + (1-omega)*(dmvnorm(Y,mean2,sigma))
}

##################################
####End function definitions#####
##################################
Y<-Makedat(seed,numsets,N,mu11,mu12,mu13,mu14,mu21,mu22,mu23,mu24
,w,sigsq)

#####################################################################
#####################################################################
#####DO the simple section first, NB stands for no block#######
#####################################################################
#####################################################################
```

```
#Get one component estimates and standard errors for each dataset
onecomphessianNB<-array(0,c(5,5,numsets))
onecompcovmatNB<-array(0,c(5,5,numsets))
onecompmeanNB<-matrix(0,numsets,5)
####Get one component estimates for data sets without blocking terms
for(u in 1:numsets){
onecompestNB<-optim(c(apply(Y[,,u],2,mean),sigsq0),OnecompNB,hessian=TRUE,
Yset=Y[,,u])

onecompmeanNB[u,]<-onecompestNB$par
onecomphessianNB[,,u]<-onecompestNB$hessian
onecompcovmatNB[,,u]<-solve(onecomphessianNB[,,u])
} #Ends dataset loop - Ends one component modeling

twocompestNB<-EMNB(seed=0,EMMC=1000,Y=Y,mu110,mu120,mu130,mu140,mu210,
mu220,mu230,mu240,w0,sigsq0)

#Compute bootstrap SEs for the no blocking variability case
#########################################################################
#Now get SEs from two component fit using bootstrap method NO blocks#
#########################################################################
estsetNB<-twocompestNB

what<-estsetNB[[1]]
mu1hat<-estsetNB[[2]]
mu2hat<-estsetNB[[3]]
sigsqhat<-estsetNB[[4]]

vectest<-matrix(0,10,numsets)
vectest[1,]<-estsetNB[[1]]
vectest[2:5,]<-t(estsetNB[[2]])
vectest[6:9,]<-t(estsetNB[[3]])
vectest[10,]<-estsetNB[[4]]

tallmeanmat<-matrix(0,10,numsets)
tallstore<-array(0,c(10,B,numsets))
covpart<-array(0,c(10,10,B,numsets))
SEmat<-matrix(0,10,numsets)
#Initialize storage for estimates across B runs and numsets data sets
w1store<-matrix(0,B,numsets)
```

```
mu1store<-array(0,c(B,4,numsets))
mu2store<-array(0,c(B,4,numsets))
sigsqstore<-matrix(0,B,numsets)

tallmat<-array(0,c(10,B,numsets))
matpart<-array(0,c(10,10,B,numsets))
for(g in 1:numsets){ #begin numsets loop
#Generate the bootstrap samples and call this array W
W<-array(0,c(N,4,B))

#Create appropriate covariance matrix based on estimates
SIGhat<-diag(sigsqhat[g],4)

for(i in 1:B){ #Open W generation loop
comp1<-rbinom(1,N,what[g])
comp2<-N-comp1

if(comp1 == 0){
W2<-rmvnorm(comp2,mu2hat[g,],SIGhat)
W[,,i]<-W2}

if(comp1 == N){
W[,,i]<-rmvnorm(comp1,mu1hat[g,],SIGhat)}

if((comp1 != N)&&(comp1 != 0)){
W1<-rmvnorm(comp1,mu1hat[g,],SIGhat)
W2<-rmvnorm(comp2,mu2hat[g,],SIGhat)
W[,,i]<-rbind(W1,W2)}
} #end W generation loop

bootlist<-EMNB(seed=0,EMMC=1000,Y=W,mu110,mu120,mu130,mu140,mu210,
mu220,mu230,mu240,w0,sigsq0)
w1store[,g]<-bootlist[[1]]
mu1store[,,g]<-bootlist[[2]]
mu2store[,,g]<-bootlist[[3]]
sigsqstore[,g]<-bootlist[[4]]


} #close dataset loop
#Create Psihatbar - the mean across B of Psihat estimators
w1bar<-apply(w1store,2,mean)
```

```
mu1bar<-apply(mu1store,c(2,3),mean)
mu2bar<-apply(mu2store,c(2,3),mean)
sigsqbar<-apply(sigsqstore,2,mean)

tallmeanmat[1,]<-w1bar
tallmeanmat[2:5,]<-mu1bar
tallmeanmat[6:9,]<-mu2bar
tallmeanmat[10,]<-sigsqbar

for(g in 1:numsets){
#combine the bootstrap estimates into vectors of parameters
tallstore[1,,g]<-w1store[,g]
tallstore[2:5,,g]<-t(mu1store[,,g])
tallstore[6:9,,g]<-t(mu2store[,,g])
tallstore[10,,g]<-sigsqstore[,g]

for(b in 1:B){
matpart[,,b,g]<-(tallstore[,b,g]-tallmeanmat[,g])%*%
t(tallstore[,b,g]-tallmeanmat[,g])}}
COVMATSNB<-apply(matpart,c(1,2,4),sum)/(B-1)
#extract standard errors from the covariance matrices
for(g in 1:numsets){
SEmat[,g]<-sqrt(diag(COVMATSNB[,,g]))
}

twocompSENB<-SEmat

##End bootstrap error section

#Provides storage for MLEs for both one and two components

sigsqhat1NB<-onecompmeanNB[,5]
VARsigsq1hatNB<-onecompcovmatNB[5,5,]

alphahat_sigsq1NB<-(sigsqhat1NB^2)/VARsigsq1hatNB + 2
betahat_sigsq1NB <-(sigsqhat1NB^3)/VARsigsq1hatNB + sigsqhat1NB

sigsqhat2NB <-twocompestNB[[4]]
VARsigsqhat2NB <-(twocompSENB[10,])^2

alphahat_sigsq2NB<-(sigsqhat2NB^2)/VARsigsqhat2NB + 2
```

```
betahat_sigsq2NB <-(sigsqhat2NB^3)/VARsigsqhat2NB + sigsqhat2NB

whatNB<-twocompestNB[[1]]
VARwhatNB<-(twocompSENB[1,])^2

ahatNB<- whatNB^2*(1-whatNB)/VARwhatNB - whatNB
bhatNB<- whatNB*(1-whatNB)^2/VARwhatNB - 1 + whatNB

ADAhatNB <-onecompcovmatNB[1:4,1:4,]
ADA1hatNB<-COVMATSNB[2:5,2:5,]
ADA2hatNB<-COVMATSNB[6:9,6:9,]

#ALTER ANY UNACCEPTABLE MLES
for(u in 1:numsets){
if( (alphahat_sigsq1NB[u]<2)|(betahat_sigsq1NB[u]<0)){
alphahat_sigsq1NB[u]<-2.0001
betahat_sigsq1NB[u] <-0.0001}

if( (alphahat_sigsq2NB[u]<2)|(  betahat_sigsq2NB[u]<0)){
alphahat_sigsq2NB[u]<-2.0001
betahat_sigsqNB[u] <-0.0001}

if( (ahatNB[u]<0)|(bhatNB[u]<0) ){
ahatNB[u]<-1
bhatNB[u]<-1}

if( det(ADAhatNB[,,u])<=0){
ADAhatNB[,,u]<-diag(1,4)}
if( det(ADA1hatNB[,,u])<=0){
ADA1hatNB[,,u]<-diag(1,4)}
if( det(ADA2hatNB[,,u])<=0){
ADA2hatNB[,,u]<-diag(1,4)}

}

like1storeNB<-matrix(0,NSAMP,numsets)
like2storeNB<-matrix(0,NSAMP,numsets)
prior1storeNB<-matrix(0,NSAMP,numsets)
prior2storeNB<-matrix(0,NSAMP,numsets)
gprior1storeNB<-matrix(0,NSAMP,numsets)
gprior2storeNB<-matrix(0,NSAMP,numsets)
```

```
#####################################################
#####Begin iterating over data sets###############
#####################################################
for(u in 1:numsets){
#MLEs two component model
mu11hatNB<-twocompestNB[[2]][u,1]
mu12hatNB<-twocompestNB[[2]][u,2]
mu13hatNB<-twocompestNB[[2]][u,3]
mu14hatNB<-twocompestNB[[2]][u,4]
mu21hatNB<-twocompestNB[[3]][u,1]
mu22hatNB<-twocompestNB[[3]][u,2]
mu23hatNB<-twocompestNB[[3]][u,3]
mu24hatNB<-twocompestNB[[3]][u,4]
sigsq2hat<-twocompestNB[[4]][u]
whatNB<-twocompestNB[[1]][u]

#MLEs for one compont model
#Pull MLEs from above info - once component model
mu1hatNB    <-onecompestNB$par[1]
mu2hatNB    <-onecompestNB$par[2]
mu3hatNB    <-onecompestNB$par[3]
mu4hatNB    <-onecompestNB$par[4]

########Draw from distributions characterized by these MLEs#####

DrawC21_LocNB<-rmvnorm(NSAMP,c(mu11hatNB,mu12hatNB,mu13hatNB,mu14hatNB),
ADA1hatNB[,,u])
DrawC22_LocNB<-rmvnorm(NSAMP,c(mu21hatNB,mu22hatNB,mu23hatNB,mu24hatNB),
ADA2hatNB[,,u])
DrawC1_LocNB <-rmvnorm(NSAMP,c(mu1hatNB ,mu2hatNB ,mu3hatNB ,mu4hatNB ),
ADAhatNB[,,u])

drawmu11NB<-DrawC21_LocNB[,1]
drawmu12NB<-DrawC21_LocNB[,2]
drawmu13NB<-DrawC21_LocNB[,3]
drawmu14NB<-DrawC21_LocNB[,4]
drawmu21NB<-DrawC22_LocNB[,1]
drawmu22NB<-DrawC22_LocNB[,2]
drawmu23NB<-DrawC22_LocNB[,3]
drawmu24NB<-DrawC22_LocNB[,4]
```

```
drawwNB<-rbeta(NSAMP,ahatNB[u],bhatNB[u])

drawmu1NB<-DrawC1_LocNB[,1]
drawmu2NB<-DrawC1_LocNB[,2]
drawmu3NB<-DrawC1_LocNB[,3]
drawmu4NB<-DrawC1_LocNB[,4]

drawSigsq2NB<-(rgamma(NSAMP,shape=alphahat_sigsq2NB[u],scale=1/
betahat_sigsq2NB[u])^-1)
drawSigsq1NB<-(rgamma(NSAMP,shape=alphahat_sigsq1NB[u],scale=1/
betahat_sigsq1NB[u])^-1)
}

###Iterate over importance sample draws###
for(u in 1:numsets){
parmdens1NB<-rep(0,NSAMP)
parmdens2NB<-rep(0,NSAMP)
#cycle through MC reps
for(j in 1:NSAMP){

parmdens1NB[j]<-exp(log(dmvnorm(c(drawmu1NB[j],drawmu2NB[j],
drawmu3NB[j],drawmu4NB[j]),
c(mu1hatNB,mu2hatNB,mu3hatNB,mu4hatNB),ADAhatNB[,,u]))+
log(densigamma(drawSigsq1NB[j],alphahat_sigsq1NB[u],
betahat_sigsq1NB[u])))

parmdens2NB[j]<-exp(
log(dmvnorm(c(drawmu11NB[j],drawmu12NB[j],drawmu13NB[j],
drawmu14NB[j]),c(mu11hatNB,mu12hatNB,
mu13hatNB,mu14hatNB),ADA1hatNB[,,u]))+
log(dmvnorm(c(drawmu21NB[j],drawmu22NB[j],drawmu23NB[j],
drawmu24NB[j]),c(mu21hatNB,mu22hatNB,
mu23hatNB,mu24hatNB),ADA2hatNB[,,u]))+
log(densigamma(drawSigsq2NB[j],alphahat_sigsq2NB[u],
betahat_sigsq2NB[u])) +
log(dbeta(drawwNB[j],ahatNB[u],bhatNB[u])) )

} #end importance sampling loop

#}#END dataset loop
```

```
####Use vague priors assessed at the draws
prior1NB<-rep(0,NSAMP)
prior2NB<-rep(0,NSAMP)
for(j in 1:NSAMP){ #begin new importance sampling

prior1NB[j]<-
dmvnorm(c(drawmu1NB[j],drawmu2NB[j],drawmu3NB[j],
drawmu4NB[j]),c(0,0,0,0),(diag(rep(800,4)))) *
densigamma(drawSigsq1NB[j],2.001,10)

prior2NB[j]<-
dmvnorm(c(drawmu11NB[j],drawmu12NB[j],drawmu13NB[j],
drawmu14NB[j]),c(0,0,0,0),(diag(rep(800,4))))*
dmvnorm(c(drawmu21NB[j],drawmu22NB[j],drawmu23NB[j],
drawmu24NB[j]),c(0,0,0,0),(diag(rep(800,4))))*
densigamma(drawSigsq2NB[j],2.001,10)*
dbeta(drawwNB[j],1,1)
}

#The funcion g in the notation IS the joint density of
#the parameters
gprior1NB<-parmdens1NB
gprior2NB<-parmdens2NB

like1NB<-rep(0,NSAMP)
dens1NB<-rep(0,n)

like2NB<-rep(0,NSAMP)
dens2NB<-rep(0,n)
#Marginal for two components

##THIS IS THE POINT THE ALGORITHM REALLY SLOWS DOWN
#cycle through MC reps
for(j in 1:NSAMP){

#attempt to use apply statement
dens1NB<-apply(matrix(Y[,,u],n,4,byrow=FALSE),1,dmvnorm
,mean=c(drawmu1NB[j],
drawmu2NB[j],drawmu3NB[j],drawmu4NB[j]),sigma=
matrix(c(
```

```
drawSigsq1NB[j] ,0,0,0,
0,drawSigsq1NB[j] ,0,0,
0,0,drawSigsq1NB[j] ,0,
0,0,0,drawSigsq1NB[j] ),4,4),log=T)

like1NB[j]<-exp(sum(dens1NB))

}

#cycle through MC reps
for(j in 1:NSAMP){

dens2NB<-apply(matrix(Y[,,u],n,4,byrow=FALSE),1,
dens2norm,
mean1=c(drawmu11NB[j],drawmu12NB[j],drawmu13NB[j],
drawmu14NB[j]),
mean2=c(drawmu21NB[j],drawmu22NB[j],drawmu23NB[j],
drawmu24NB[j]),
sigma=matrix(c(
drawSigsq2NB[j] ,0,0,0,
0,drawSigsq2NB[j] ,0,0,
0,0,drawSigsq2NB[j] ,0,
0,0,0,drawSigsq2NB[j] ),4,4),
omega=drawwNB[j])

like2NB[j]<-prod(exp(log(dens2NB)))
}

like1storeNB[,u]<-like1NB
like2storeNB[,u]<-like2NB
prior1storeNB[,u]<-prior1NB
prior2storeNB[,u]<-prior2NB
gprior1storeNB[,u]<-gprior1NB
gprior2storeNB[,u]<-gprior2NB

m2NB[u]<-mean(like2NB*prior2NB/gprior2NB)
m1NB[u]<-mean(like1NB*prior1NB/gprior1NB)

m2NBNG[u]<-mean(like2NB*gprior2NB/gprior2NB)
m1NBNG[u]<-mean(like1NB*gprior1NB/gprior1NB)
```

```
} #end data set loop

BFNB<-m2NB/m1NB
BFNBNG<-m2NBNG/m1NBNG

save.image(file='Space')
```