Order Restricted Testing of Random Effects in Generalized Linear Mixed Models

by

Voleak Choeurng

A Thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Master of Science

Ottawa-Carleton Institute for Mathematics and Statistics
Carleton University
Ottawa, Ontario, Canada
August 2013

Abstract

Generalized linear mixed models (GLMM) have been used in many areas of research to analyze longitudinal and clustered data with non-normal responses. In addition to the fixed effects parameters found in the generalized linear model (GLM), variance components associated with unobservable random effects are estimated in the GLMM. Moreover, it is well understood that order restricted inference methods that properly incorporate additional information by way of a restricted parameter space are more efficient than procedures that ignore this information. In this thesis, a distance statistic based on the Wald statistic is suggested for order restricted tests on the random components in the mixed model. The null distributions of the distance and the likelihood ratio test statistics are shown to be asymptotically equivalent and that of a chi-bar-square. An analysis conducted on data extracted from the 2011 National Youth Tobacco Survey will serve as an illustration of the proposed testing procedure.

Acknowledgments

I will always be grateful for the guidance, patience and support extended to me from my supervisor, Dr. Chul Gyu Park. If there is one thing that he has impressed upon me the most it is the idea that you must thoroughly understand the details, small as they might be, before you can hope to have a sense of the bigger picture. I sincerely thank him for all of the time he has committed to my development as a statistician.

I would also like to thank the members of my committee, Dr. Craig Leth-Steensen, Dr. Raluca Balan and Dr. Sanjoy Sinha for their valuable comments and recommendations in improving this thesis.

I wish to thank the Centers for Disease Control and Prevention for making their National Youth Tobacco Survey data readily accessible to the public.

And finally, I wish to thank my family and friends for always being there for me in so many different ways.

Table of Contents

List of Tables

Chapter 1

Introduction

It is often the case when conducting statistical tests in scientific disciplines that populations follow a natural ordering on the parameter space. For example, in a clinical trial studying the efficacy of a new drug we might expect that the mean response time to this particular drug would be ordered according to dosage levels. This type of partial information has been incorporated in inferential procedures by many researchers including Bartholomew (1959a, 1959b), Kudo (1963), Shapiro (1988), and so on. If the underlying information (given as constraints) is correct, the constrained inference achieves higher precision than its unconstrained counterparts. Numerous theories and methodologies on this subject are well summarized in Robertson et. al. (1988) and Silvapulle and Sen (2005).

At the same time, over the last few decades we have seen the development of both generalized linear models (GLMs) and generalized linear mixed models (GLMMs) as necessary extensions to both linear models (LMs) and linear mixed models (LMMs). The GLM was proposed to analyse data for which the response variables are not normally distributed as required by the linear models. This class of regression models assumes that all observations are independent of each other. However, there are many situations in which this assumption cannot be made. We come across these situations whenever data are seen to be clustered in some way such as the case with longitudinal

data where there exists dependency in the response, or survey data where subjects are observed nested within larger units. Therefore, in the GLMM we must model not only the conditional mean response but also random effects in order to reflect the dependence structure in responses. Although these random effects are unobservable, their variance components allow us to partition the overall variance in an explainable way. Often our analysis starts with prior knowledge that the variance component of one random effect should be at least the same as the variance component of another. Then, it is reasonable to compare those variance components under the given ordering information.

In the context of hypothesis testing in GLMMs, researchers have been typically concerned with the problem of investigating the presence of a random effect. If it can be concluded based on a testing procedure that the variance components associated with the random effects are zero, then the GLMM reduces to a simpler GLM. Lin (1997) proposed a global score test for the null hypothesis that all variance components are zero. It is a robust test in the sense that it does not require specifying the joint distribution of the random effects. This test is based on estimation of parameters through the penalised quasilikelihood of Breslow and Clayton (1993) that was originally proposed by Green (1987). Lin (1997) showed that the test performs well when the number of levels of each random effect was at least moderately large, but was unsatisfactory for binary response data.

Self and Liang (1987) extended earlier work by Chernoff (1954) in developing a likelihood ratio test (LRT) that could also be applied toward testing for the presence of random effects. They were able to show that the LRT in this case often had a mixed chi-squared distribution. Hence, statistical inferences involving random effects has received a fair amount of attention by various researchers. However, comparison of random components under their ordering information has not been studied in as much detail. This thesis is aimed to present a viable testing procedure for comparing

random effects when they are ordered in a certain way.

1.1 Motivation

As mentioned above, situations may occur where data are best modeled using a generalized linear mixed model with multiple random effects whose variance components are ordered. One example where this could potentially be seen is in cluster correlated data where stronger dependency might be assumed among observations within clusters. The idea can be explained in the following way. Suppose we were conducting a case study on patient care satisfaction in hospitals across Canada. Such a study would involve sampling, say, 20 hospitals from the population of all medical institutions in Canada. The random variation from one institution to another could then be modeled through a normally distributed random effect to be referred to as the hospital effect and would comprise of 20 levels. If the responses from patients in the same hospital are independent then the mean responses from all hospitals are likely to be similar given that the hospital effects are drawn from the same normal distribution. This would result in a relatively smaller variance associated with this random effect. If, however, the responses regarding patient satisfaction within the same hospitals were dependent instead, then responses from individual hospitals would be similar but different from those of other hospitals. Consequently, the levels of this random effect could vary more significantly than the prior case, producing a larger variance component for the hospital effect.

Another way to explain the relationship between dependent data and the size of a variance component is through the idea of the intracluster correlation (ICC), which was originally introduced by Dr. R.A. Fisher. Take the example of the linear mixed model in which the data are grouped into clusters $j = 1, ..., q$ given by, $Y_{ij} = X_{ij}^T\boldsymbol{\beta} + b_j + e_{ij}$, where b_j represents a random effect and e_{ij} a random error term. Set

$Var(e_{ij}) = \sigma_e^2$ and $Var(b_j) = \sigma_b^2$. Then the correlation between a pair of observations within the same cluster is given by

$$\text{corr}(Y_{ij}, Y_{i'j}) = \frac{\sigma_b^2}{\sigma_e^2+\sigma_b^2}.$$

Therefore, the larger the variance of the random effect relative to the variance of the error term, the stronger the correlation or dependence. In the GLMM, a conditional mean is modeled as linear in the predictors through a monotonic link function, and so a similar relationship regarding dependency and variance continues to apply.

Nested designs like the one just described are fairly ubiquitous within survey methodology where it could be assumed that the strongest dependency occurs in the final sampling unit where interaction between the observational units is greatest. One such case of a nested design is seen in the American National Youth Tobacco Survey (NYTS), which is conducted on a biennial basis. In the NYTS it is possible to treat county, school and class as random effects whose variance components are ordered because dependency appears more significantly according to cluster size in this example. This thesis will propose, in detail, a method for order restricted testing of random effects with an application towards a portion of the 2011 NYTS data set to illustrate the methodology.

1.2 Overview of Thesis

The remainder of the thesis will be organised in the following manner: In Chapter 2, results concerning GLMMs will be described in some detail along with unconstrained estimation of their parameters. A review of the chi-bar-squared distribution, as it relates to inequality constrained testing problems, will be given thorough treatment. In Chapter 3, a testing procedure for making inequality constrained inferences on random effects in GLMMs will be expounded upon along with the relevant asymptotics. Chapter 4 summarises the National Youth Tobacco Survey data set which will

be used as an example in applying the proposed order restricted test statistic. The NYTS data set will be modeled using a logistic mixed model and complete formula derivations will be provided. Finally, the results and a discussion of potential avenues for future studies will be given in Chapter 5. A list of references along with some relevant R code can be found in the appendix at the end of the thesis.

Chapter 2

Review of Related Studies

In this chapter, concepts related to the generalized linear mixed model (GLMM) are reviewed along with the estimation and testing of parameters under linear inequality constraints.

2.1 Generalized Linear Mixed Models

Generalized linear mixed models allow us to model not only non-normal data but go one step further in providing a way to account for the correlation observed in clustered or longitudinal data. More generally, GLMMs may provide a way to model multiple sources of variation in the form of random effects. In this section we discuss the general theory behind the class of generalized linear mixed models.

2.1.1 Model Structure

Suppose the data are composed of n observations with responses y_i, $p \times 1$ covariate vectors $\mathbf{x}_i$ associated with the fixed effects and $q \times 1$ known vectors $\mathbf{z}_i$ associated with random effects. Then it is typical to assume that the responses, Y_i, given the random effects, $\mathbf{u}$, are conditionally independent and follow a distribution from an exponential family. In other words,

$$
\begin{aligned}
y_i|\mathbf{u} &\sim \text{independent} f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}), \\
f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}) &= \exp\{[y_i\gamma_i - b(\gamma_i)]/\tau^2 - c(y_i,\tau)\}, \\
\mathbf{u} &\sim f_{\mathbf{u}}(\mathbf{u}|\mathbf{D}).
\end{aligned} \tag{1}
$$

The conditional mean cannot be modeled directly as a linear function of the predictors unless the mean can take any value on the real line. For this reason a differentiable and monotone function $g(\cdot)$ is used to link the mean of the conditional distribution of y_i with the linear form of predictors. This function is thusly referred to as the link function.

$$
\begin{aligned}
E(y_i|\mathbf{u}) &= b'(\gamma_i) = \mu_i, \\
g(\mu_i) &= x_i^T\boldsymbol{\beta} + z_i^T\mathbf{u} = \eta_i.
\end{aligned} \tag{2}
$$

In many cases, the distribution of the random effects is assumed to be $N(0,\sigma^2)$ whenever $\mathbf{u}$ is univariate and $N_k(\mathbf{0},\boldsymbol{\Sigma})$ if $\mathbf{u}$ is a k-dimensional random vector. Here, we assume that random components are independent so that $\boldsymbol{\Sigma} = \text{diag}\{\mathbf{D}\} \equiv \text{diag}\{\sigma_1^2,\sigma_2^2,...,\sigma_k^2\}$.

2.1.2 Estimation in GLMMs

Estimation of the parameters for the generalized linear mixed model begins with writing down the log of the marginal likelihood of the observations, $\mathbf{y}$. The log-likelihood is obtained by integrating out the random effects from the joint distribution of y_i and $\mathbf{u}$:

$$
l = \log \int \prod_i f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) \mathrm{d}\mathbf{u}.
$$

This is in no way a trivial matter as the dimension of the integral, which depends on the dimension of the random effects, $\mathbf{u}$, can greatly complicate the calculation. When the dimension is relatively low, methods like numerical quadrature are available to tackle the integration. The method is described in detail in a number of numerical analysis books including one by Davis and Rabinowitz (1975). For GLMMs that are not amenable to numerical quadrature due to the dimensionality of the integral, Breslow and Clayton (1993) developed what they've referred to as the penalized quasi-likelihood (PQL) as an approximating method to solving for the MLE's. It utilizes the Laplace approximation as well as Taylor series expansions in circumventing the intractable integral. The Laplace method is used to approximate integrals of the form

$$\int \exp\{-q(x)\}\mathrm{d}x$$

by expanding $q(x)$ to the quadratic term. A multivariate analog of the Laplace approximation is used in the penalized quasi-likelihood method. Unfortunately, Breslow and Clayton found that this method only performed well when the conditional distribution of the response given the random effects was nearly normal. PQL did not work well, for example, with binary data. One reason that has been suggested for this is that there are just too many approximations. Breslow and Lin (1995) and Lin and Breslow (1996) made attempts to improve the performance of PQL by using higher order Laplace approximations but only saw marginal improvements.

With technological advances over the past decade, particularly with respect to computing power, Markov Chain Monte Carlo (MCMC) methods have certainly become very popular for solving integrals of all kinds. As such, MCEM and MCNR algorithms that combine the EM algorithm and the Newton-Raphson method, respectively, with Monte Carlo methods are very often employed in the estimation of parameters within the GLMM framework. An MCNR algorithm will be described in detail in Chapter 3, but first, the maximum likelihood estimating equations must be

presented.

As alluded to earlier, the log-likehood of $\mathbf{y}$ is given by the following integral:

$$\begin{aligned} l(\boldsymbol{\beta}, \tau, \mathbf{D}) &= \log \int \prod_{i=1}^{n} f_{Y_i|u}(y_i|\mathbf{u}, \boldsymbol{\beta}, \tau) f_{\mathbf{u}}(\mathbf{u}|\mathbf{D}) \mathrm{d}\mathbf{u} \\ &= \log \int f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \tau) f_{\mathbf{u}}(\mathbf{u}|\mathbf{D}) \mathrm{d}\mathbf{u} \\ &= \log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\beta}, \tau, \mathbf{D}). \end{aligned} \tag{3}$$

Now let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \tau^T)^T$. Then differentiating (3) with respect to $\boldsymbol{\theta}$ gives

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\theta}} &= \frac{1}{f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D})} \frac{\partial}{\partial \boldsymbol{\theta}} \int f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) f_{\mathbf{u}}(\mathbf{u}|\mathbf{D}) \mathrm{d}\mathbf{u} \\ &= \int \Big[\frac{\partial}{\partial \boldsymbol{\theta}} f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) \Big] f_{\mathbf{u}}(\mathbf{u}|\mathbf{D}) \mathrm{d}\mathbf{u} / f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}). \end{aligned} \tag{4}$$

Note that the partial derivative of the conditional distribution of $\mathbf{Y}$ given $\mathbf{u}$ may be written as

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) &= \Big[\frac{1}{f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) \Big] f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) \\ &= \Big[\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) \Big] f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}). \end{aligned} \tag{5}$$

Substituting (5) into (4), the derivative can be rewritten as

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\theta}} &= \int \Big[\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) \Big] f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) f_{\mathbf{u}}(\mathbf{u}|\mathbf{D}) \mathrm{d}\mathbf{u} / f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{D}) \\ &= \int \Big[\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) \Big] f_{\mathbf{u}|\mathbf{Y}}(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}, \mathbf{D}) \mathrm{d}\mathbf{u} \\ &= E\Big[\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta}) \Big| \mathbf{y} \Big]. \end{aligned}$$

Similarly, an expression for the partial derivative of the log-likelihood with respect to $\mathbf{D}$ can be derived:

$$\frac{\partial l}{\partial \mathbf{D}} = E\Big[\frac{\partial}{\partial \mathbf{D}} \log f_{\mathbf{u}}(\mathbf{u}|\mathbf{D})\Big|\mathbf{y}\Big].$$

Thus, the maximum likelihood estimating equations are given by

$$E\Big[\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\theta})\Big|\mathbf{y}\Big] = 0, \tag{6}$$

$$E\Big[\frac{\partial}{\partial \mathbf{D}} \log f_{\mathbf{u}}(\mathbf{u}|\mathbf{D})\Big|\mathbf{y}\Big] = 0. \tag{7}$$

In order to solve the first estimating equation, refer back to equation (1) and write the log-likelihood of the conditional distribution of $\mathbf{Y}$ given $\mathbf{u}$ as

$$l_{\mathbf{y}|\mathbf{u}} = \sum_{i=1}^{n} [y_i \gamma_i - b(\gamma_i)]/\tau^2 - \sum_{i=1}^{n} c(y_i, \tau). \tag{8}$$

Further simplification of this estimating equation requires the application of a number of results regarding the conditional distribution of y_i given the random effects $\mathbf{u}$. This conditional distribution, as defined in the preceding section, belongs to an exponential family. The useful results, for which derivations may be found in McCullagh and Nelder (1989) as well as McCulloch and Searle (2001), are given next.

$$\mu_i = \frac{\partial b(\gamma_i)}{\partial \gamma_i} \tag{9}$$

$$\begin{aligned} \text{var}(y_i|\mathbf{u}) &= \tau^2 \frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2} \\ &= \tau^2 v(\mu_i) \end{aligned} \tag{10}$$

$$\frac{\partial \gamma_i}{\partial \mu_i} = \frac{1}{v(\mu_i)} \tag{11}$$

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \left(\frac{\partial g(\mu_i)}{\mu_i} \right)^{-1} \mathbf{x}_i^T \tag{12}$$

The term $v(\mu_i)$ is often referred to as the variance function as it relates the conditional variance to the conditional mean. Furthermore, τ is known as the dispersion parameter which may be estimated along with the other parameters in the model. When the distribution of the response conditioned on the random effects is assumed to follow a binomial or poisson distribution, however, the dispersion parameter takes on a value of unity.

The results from these maximum likelihood equations will be used to analyze a case study in Chapter 4 where the response data is assumed to follow a Bernoulli distribution when conditioned on the random effects, and so τ will be assigned a value of 1 in these next derivations. For other assumed distributions, τ may be estimated in much the same way as the regression parameters.

Begin by differentiating equation (8) with respect to $\boldsymbol{\beta}$.

$$\begin{aligned}
\frac{\partial l_{\mathbf{y}|\mathbf{u}}}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left[y_i \frac{\partial \gamma_i}{\partial \boldsymbol{\beta}} - \frac{\partial b(\gamma_i)}{\partial \gamma_i} \frac{\partial \gamma_i}{\partial \boldsymbol{\beta}} \right] \\
&= \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \gamma_i}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \gamma_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{v(\mu_i)} \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} \mathbf{x}_i^T \quad \text{(upon using (11) and (12))}
\end{aligned}$$

$$= \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{v(\mu_i)} \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-2} \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right) \mathbf{x}_i^T$$
$$= \sum_{i=1}^{n} (y_i - \mu_i) w_i \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right) \mathbf{x}_i^T,$$

where $w_i = \left[v(\mu_i) \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^2 \right]^{-1}$ or $\left[v(\mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 \right]^{-1}$.

In matrix form, this can be written as

$$\frac{\partial l_{\mathbf{y}|\mathbf{u}}}{\partial \boldsymbol{\beta}} = \mathbf{X^T W \Delta (y - \boldsymbol{\mu})}, \tag{13}$$

where $\mathbf{W}$ is the diagonal matrix with elements w_i and $\mathbf{\Delta}$ is the diagonal matrix with elements $\frac{\partial g(\mu_i)}{\partial \mu_i}$ or $\frac{\partial \eta_i}{\partial \mu_i}$. Next, the second derivative of the conditional log-likelihood with respect to the regression parameters is found. This will be useful for deriving the asymptotic variance of $\hat{\boldsymbol{\beta}}$ as well as for finding $\hat{\boldsymbol{\beta}}$ itself.

$$\frac{\partial^2 l_{\mathbf{y}|\mathbf{u}}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\mathbf{X^T W \Delta} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta^T}} + \mathbf{X^T} \frac{\partial \mathbf{W \Delta}}{\partial \boldsymbol{\beta^T}} (\mathbf{y} - \boldsymbol{\mu})$$

Then taking the expectation of this last quantity gives

$$E \left[\frac{\partial^2 l_{\mathbf{y}|\mathbf{u}}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right] = -\mathbf{X^T W \Delta} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta^T}}$$
$$= -\mathbf{X^T W X}. \tag{14}$$

A second order Taylor expansion of the log-likelihood of $\mathbf{Y}|\mathbf{u}$ about a value $\boldsymbol{\beta}_0$ produces the following approximation, which is the basis for the Newton-Raphson iterative equation (or Fisher scoring algorithm if the expectation of the Hessian is used):

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}) \approx \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

where equations (13) and (14) may then be used to simplify the first and second terms on the right. Maximum likelihood estimates for the regression parameters are computed by inserting this equation into estimating equation (6) and solving for $\boldsymbol{\beta}$.

In order to solve estimating equation (7) for the variance components, a distribution for the random effects must first be specified. It is often the case that this distribution is assumed to be normal. Consider the event that $\mathbf{u} \sim N(0, \sigma^2)$. Then, since the log-likelihood of $\mathbf{u}$ is

$$\log f_{\mathbf{u}}(\mathbf{u}|\sigma^2) = -\frac{q}{2}\log(2\pi\sigma^2) - \sum_{j=1}^{q} u_j^2/(2\sigma^2),$$

the derivative with respect to σ^2 is given by

$$\frac{\partial}{\partial\sigma^2}\log f_{\mathbf{u}}(\mathbf{u}|\sigma^2) = -\frac{q}{2\sigma^2} + \frac{1}{2}\sum_{j=1}^{q} u_j^2/(\sigma^2)^2.$$

Inserting this equation into estimating equation (7) and solving for σ^2 will provide the MLE for the variance component.

Based on the shortcomings of other estimation methods as described at the beginning of this subsection, it was decided that the MCNR algorithm as defined in McCulloch (1997) will be the method of choice for the estimation of the GLMM parameters in this thesis. McCulloch also provides steps for an MCEM algorithm but authors like Neath (2006) have found that MCNR typically converges faster than MCEM.

Estimates of both the regression parameters and random components can be obtained by solving equations (6) and (7) in the manner that was just described. However, direct computation of these expectations requires knowledge of the distribution of $\mathbf{u}|\mathbf{y}$, which in turn requires knowledge of the marginal distribution of $\mathbf{y}$ - the distribution that is being avoided in the first place because it involves an intractable

integral. The MCMC solution to this problem is to produce a Monte Carlo sample from the distribution of $\mathbf{u}|\mathbf{y}$ via the Metropolis algorithm or other sampling methods. The Gibbs sampler is another popular choice for obtaining draws from an otherwise inaccessible distribution. Once draws have been made from this target distribution, Monte Carlo estimates can be used to compute the expectations found in estimating equations (6) and (7) and solve for the parameters in cohesion with the Newton-Raphson method. A full description of this MCNR algorithm will be given in Chapter 3.

2.2 Inference Under Linear Inequality Constraints

In the case of a one-dimensional parameter hypothesis test, a statistician might be interested in the following one-sided testing problem: $H_0 : \mu = 0$ against $H_1 : \mu > 0$. Testing under linear inequality constraints can be thought of as a multi-parameter extension to this single-parameter case. One example of a constrained test for population means could be, $H_0 : \mu_1 = \mu_2 = \mu_3$ against $H_1 : \mu_1 \leq \mu_2 \leq \mu_3$. In order to derive a likelihood ratio test statistic and its null distribution under this type of ordering, a few concepts from linear algebra must first be introduced as they form a central role in the estimation and testing procedure for problems considered throughout the rest of this thesis.

2.2.1 Norms, Convex Cones and Projections

Definition 2.1 A real inner product space, $\mathcal{V}$, is a real vector space equipped with an inner product. An inner product on $\mathcal{V}$ is a function that takes each ordered pair (x, y) of elements in $\mathcal{V}$ and maps them to a number, $\langle x, y \rangle$, with the following properties:

1. $\langle x, x \rangle \geq 0$ for all $x \in \mathcal{V}$

2. $\langle x, x\rangle = 0$ if and only if $x = 0$

3. $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$ for all $x, y, z \in \mathcal{V}$

4. $\langle ay, z\rangle = a\langle y, z\rangle$ for any scalar a and all $y, z \in \mathcal{V}$

Definition 2.2 A norm is a real-valued function $\|\cdot\|$ defined on $\mathcal{V}$ that satisfies these 3 properties:

1. $\|x\| \geq 0\ \forall x \in \mathcal{V}$ and $\|x\| = 0$ iff $x = 0$

2. $\|x + y\| \leq \|x\| + \|y\|\ \forall x, y \in \mathcal{V}$

3. $\|ax\| = |a|\|x\|$ for any scalar a and all $x \in \mathcal{V}$

An inner product on a vector space induces an associated norm and so an inner product space is also a normed vector space. The Euclidean space $\mathbb{R}^m$ is naturally equipped with the inner product $\langle x, y\rangle = x^T y$ as well as norm $\|x\| = (x^T x)^{1/2}$ $\forall x, y \in \mathbb{R}^m$. Furthermore, given a positive definite symmetric matrix $\mathbf{\Sigma}$, we can define the inner product with respect to $\mathbf{\Sigma}$ as $\langle x, y\rangle_{\mathbf{\Sigma}} = x^T \mathbf{\Sigma}^{-1} y$, which induces the corresponding norm $\|x\|_{\mathbf{\Sigma}} = \langle x, x\rangle_{\mathbf{\Sigma}}^{1/2}$. Note that the norm of a vector $x \in \mathbb{R}^p$ can be thought of as a measure of magnitude or distance from the origin to the endpoint of x.

Definition 2.3 A subset $\mathcal{C}$ of $\mathbb{R}^m$ is called a cone or a positively homogeneous set if $ax \in \mathcal{C}$ for any positive scalar a and $x \in \mathcal{C}$.

The linear inequality constraints of the alternative hypothesis suggested in the introduction to this section constitutes a cone. If $\mathcal{C}$ is a closed set and any two points $x, y \in \mathcal{C}$ can be connected by a line that is contained entirely within $\mathcal{C}$ then it is referred to as a closed convex cone. An associated polar cone, $\mathcal{C}^0$, may then be defined as follows.

Definition 2.4 The polar cone, $\mathcal{C}^0$, associated with the cone, $\mathcal{C}$, is the set, $\{x : x^T\boldsymbol{\Sigma}^{-1}y \leq 0 \ \forall y \in \mathcal{C}\}$. The polar cone is therefore the set of vectors that form an obtuse angle with every vector in $\mathcal{C}$.

Definition 2.5 The projection of a vector x onto a vector space $\mathcal{V}$ is the vector in $\mathcal{V}$ that is closest to x with respect to the metric $\|\cdot\|_{\boldsymbol{\Sigma}}$. It is denoted by $P_{\boldsymbol{\Sigma}}(x|\mathcal{V})$.

If it is well understood, based on the context, that the metric is $\|\cdot\|_{\boldsymbol{\Sigma}}$ then $P(x|\mathcal{V})$ will suffice.

As an example, consider the following minimization problem: $\min_{\boldsymbol{\theta}\in\mathcal{C}}(\mathbf{x}-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\theta})$. Since $(\mathbf{x}-\boldsymbol{\theta})^T\boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\theta}) = \|\mathbf{x}-\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}}$, the solution, $\mathbf{x}^*$, is the projection of $\mathbf{x}$ onto $\mathcal{C}$. We say that $\mathbf{x}^*$ is $\boldsymbol{\Sigma}$-closest to $\mathbf{x}$, or closest to $\mathbf{x}$ with respect to the distance $\|\cdot\|_{\boldsymbol{\Sigma}}$. Again, if the context is clear then simply $\|\cdot\|$ will do.

A couple of useful properties to note regarding norms and projections are

$$\|\mathbf{x}\|^2 = \|P(\mathbf{x}|\mathcal{C})\|^2 + \|\mathbf{x} - P(\mathbf{x}|\mathcal{C})\|^2 \tag{15}$$

and

$$\mathbf{x} - P(\mathbf{x}|\mathcal{C}) = P(\mathbf{x}|\mathcal{C}^0). \tag{16}$$

2.2.2 The Chi-bar-squared statistic

The chi-bar-squared statistic has a history that really dates back to Bartholomew (1959), in which an order restricted test for a set of means was constructed. The

statistic and its distributional result plays a central role in tests involving linear inequality constraints, and it will now be defined. Let $\mathbf{y} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ be a $p \times 1$ random variable and $\mathcal{C}$ a convex cone. Then the statistic

$$\begin{aligned}\bar{\chi}^2 &= \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} - \min_{\boldsymbol{\theta}\in\mathcal{C}}(\mathbf{y}-\boldsymbol{\theta})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta}) \\ &= \|\mathbf{y}\|^2 - \|\mathbf{y} - P(\mathbf{y}|\mathcal{C})\|^2 \\ &= \|P(\mathbf{y}|\mathcal{C})\|^2 \end{aligned} \tag{17}$$

is distributed as a mixture of chi-squared distributions, which is referred as the chi-bar-squared distribution. In other words,

$$Pr\{\bar{\chi}^2 \geq t\} = \sum_{i=0}^{p} w_i(p, \mathbf{\Sigma}, \mathcal{C}) Pr\{\chi_i^2 \geq t\} \ \forall t > 0, \tag{18}$$

where w_i are the chi-bar-squared weights that depend on $\mathbf{\Sigma}$ and $\mathcal{C}$ and that sum up to 1. It can be seen to be an average of chi-squared distributions. More information on the weights will be given in the following sections.

The chi-bar-squared statistic in (17) may also be written as

$$\begin{aligned}\bar{\chi}^2 &= \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} - \min_{\boldsymbol{\theta}\in\mathcal{C}}(\mathbf{y}-\boldsymbol{\theta})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta}) \\ &= \|P(\mathbf{y}|\mathcal{C})\|^2 \\ &= \|\mathbf{y} - P(\mathbf{y}|\mathcal{C}^0)\|^2 \\ &= \min_{\boldsymbol{\theta}\in\mathcal{C}^0}(\mathbf{y}-\boldsymbol{\theta})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta}). \end{aligned} \tag{19}$$

Hence, the statistics $\min_{\boldsymbol{\theta}\in\mathcal{C}^0}(\mathbf{y}-\boldsymbol{\theta})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta})$ and $\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} - \min_{\boldsymbol{\theta}\in\mathcal{C}}(\mathbf{y}-\boldsymbol{\theta})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta})$ have exactly the same chi-bar-squared distribution. This is a powerful result because often the statistics of interest will be of the form, $\min_{\boldsymbol{\theta}\in\mathcal{C}^0}(\mathbf{y}-\boldsymbol{\theta})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta})$.

2.2.3 Estimation Under Linear Inequality Constraints

While interest in this thesis lies in the estimation of parameters under a set of constraints by minimizing a quadratic form, Jamshidian (2004) considered the case of maximizing a given likelihood function. He proposes a gradient projection algorithm (GP) for maximum likelihood estimation under linear equality and linear inequality constraints on parameters.

In general, the goal is to maximize a given objective function, $Q(\boldsymbol{\theta})$ subject to $\boldsymbol{\theta} \in \mathcal{C}$ where $\mathcal{C}$ is a convex cone. In maximum likelihood estimation this would be the likelihood function. Written another way, one can say that the goal is to maximize $Q(\boldsymbol{\theta})$ subject to the following constraints:

$$\begin{aligned} \mathbf{a}_i^T \boldsymbol{\theta} &= \mathbf{b}_i, i \in I_1, \\ \mathbf{a}_i^T \boldsymbol{\theta} &\leq \mathbf{b}_i, i \in I_2, \end{aligned} \tag{20}$$

where I_1 and I_2 are index sets. Constraints that hold under equality are known as *active constraints*.

Jamshidian (2004) employs an active set gradient projection method that is described in detail in Fletcher (1987). Let $\mathcal{W}$ be an initial working set of active constraints. Let $\bar{A}$ be an $\bar{m} \times p$ matrix whose rows consist of $\mathbf{a}_i^T$ for all $i \in \mathcal{W}$, and let $\bar{\mathbf{b}}$ denote the corresponding vectors of $\mathbf{b}_i's$. Let W be a positive definite matrix, I be the identity matrix with the same order as W, and $\mathbf{g}(\boldsymbol{\theta})$ be the gradient of the objective function $Q(\boldsymbol{\theta})$. Also, let $\tilde{\mathbf{g}}(\boldsymbol{\theta}) = W^{-1}\mathbf{g}(\boldsymbol{\theta})$ be the generalized gradient of $Q(\boldsymbol{\theta})$ in the metric of W and let $\mathbf{d}$ denote the direction along which a move is made from one point in $\mathcal{C}$ to another point in $\mathcal{C}$. Points in $\mathcal{C}$ are called feasible points. Then Jamshidian's active set gradient projection algorithm proceeds as follows:

Starting with an initial point $\boldsymbol{\theta}_r$ that satisfies $\bar{A}\boldsymbol{\theta}_r = \bar{\mathbf{b}}$, the GP algorithm iterates

through steps 1 to 4 until convergence is achieved, resulting in an estimate that satisfies the necessary constraints.

1. Compute $\mathbf{d} = P_W \tilde{g}(\boldsymbol{\theta}_r)$, where $P_w = I - W^{-1}\bar{A}^T(\bar{A}W^{-1}\bar{A}^T)^{-1}\bar{A}$.

2. If $\mathbf{d} = 0$, compute the Lagrange multipliers $\lambda = (\bar{A}W^{-1}\bar{A}^T)^{-1}\bar{A}\tilde{\mathbf{g}}(\boldsymbol{\theta}_r)$. Let λ_i denote the i^{th} component of λ.
   If $\lambda_i \geq 0$ for all $i \in \mathcal{W} \cap I_2$, stop.

   If there is at least one negative λ_i for $i \in \mathcal{W} \cap I_2$, determine the index associated with the smallest λ_i and remove it from the set $\mathcal{W}$. Modify $\bar{A}$ and $\bar{\mathbf{b}}$ as well by dropping the corresponding row from each. Go to step 1.

3. Obtain $\alpha_1 = argmax_\alpha\{\alpha : \boldsymbol{\theta}_r + \alpha\mathbf{d} \text{ is feasible }\}$. Then search for $\alpha_2 = argmax_\alpha\{Q(\boldsymbol{\theta}_r + \alpha\mathbf{d}) : 0 \leq \alpha \leq \alpha_1\}$. Set $\tilde{\boldsymbol{\theta}}_r = \boldsymbol{\theta}_r + \alpha_2\mathbf{d}$. Add indices of new coordinates, if any, of $\tilde{\boldsymbol{\theta}}_r$ that are newly on the boundary of the working set $\mathcal{W}$. Modify $\bar{A}$ and $\bar{\mathbf{b}}$ by adding new rows accordingly.

4. Replace $\boldsymbol{\theta}_r$ by $\tilde{\boldsymbol{\theta}}_r$ and go to step 1.

2.2.4 Hypothesis Testing Under Linear Inequality Constraints

Suppose that X is a $p \times 1$ random variable from a $N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ distribution and one wishes to test the null hypothesis that $\boldsymbol{\theta}$ satisfies a set of linear equality constraints against the alternative in which it instead satisfies a set of linear inequality constraints. That

is, interest lies in testing the following hypotheses:

$$H_0 : \boldsymbol{\theta} \in \mathcal{M} = \{\boldsymbol{\theta} : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}\},$$
$$H_1 : \boldsymbol{\theta} \in \mathcal{C} = \{\boldsymbol{\theta} : \mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}\},$$
$$H_2 : \boldsymbol{\theta} \in \mathbb{R}^p.$$

Defined in this way, $\mathcal{M}$ represents a linear space and $\mathcal{C}$ a convex cone. Let the matrix $\mathbf{R}$ be of full row rank. Then any set of linear inequalities to be defined on a parameter may be written in this manner. Take for example, the convex cone $\mathcal{C} = \{\boldsymbol{\theta} : \theta_1 \geq \theta_2 \geq \theta_3\}$. This cone can be written as $\mathcal{C} = \{\boldsymbol{\theta} : \mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}\}$, where $\mathbf{R}$ is given by

$$\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix},$$

and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$.

Now consider the test of H_0 against $H_1 - H_0$ based on a single realization of X. Here $H_1 - H_0$ indicates that $\boldsymbol{\theta}$ belongs to H_1 but not H_0. Then the likelihood ratio is given by

$$\begin{aligned} L_{01} &= \min\left\{(X-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(X-\boldsymbol{\theta}) : \boldsymbol{\theta} = \mathbf{0}\right\} - \min\left\{(X-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(X-\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{C}\right\} \\ &= X^T\boldsymbol{\Sigma}^{-1}X - \min_{\boldsymbol{\theta}\in\mathcal{C}}(X-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(X-\boldsymbol{\theta}) \\ &= \|X\|^2 - \|X - P(X|\mathcal{C})\|^2 \\ &= \|P(X|\mathcal{C})\|^2. \end{aligned}$$

Hence the null hypothesis is rejected for large values of the test statistic, L_{01}. Note that the projection is taken with respect to the metric $\boldsymbol{\Sigma}$. Although the null hypothesis is composite, the p-value for the test statistic is constant over H_0. That

is,

$$\begin{aligned} p\text{-}value &= \sup_{\boldsymbol{\theta}\in H_0} pr_{\theta}(L_{01} \geq l_{01}) \\ &= pr_{\boldsymbol{\theta}_0}(L_{01} \geq l_{01}) \text{ for any } \boldsymbol{\theta} \text{ with } \mathbf{R}\boldsymbol{\theta} = \mathbf{0}. \end{aligned}$$

However, one might also be interested in testing H_1 against $H_2 - H_1$, where $\boldsymbol{\theta}$ is unrestricted under H_2, and $H_2 - H_1$ indicates that $\boldsymbol{\theta}$ belongs to H_2 but not H_1. In this case, the likelihood ratio test statistic is

$$\begin{aligned} L_{12} &= \min\left\{(X-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(X-\boldsymbol{\theta}) : \boldsymbol{\theta}\in\mathcal{C}\right\} - \min\left\{(X-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(X-\boldsymbol{\theta}) : \boldsymbol{\theta}\in\mathbb{R}\right\} \\ &= \min_{\boldsymbol{\theta}\in\mathcal{C}}(X-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(X-\boldsymbol{\theta}) \\ &= \|X - P(X|C)\|^2. \end{aligned}$$

The null value for $\boldsymbol{\theta}$ for this test can lie anywhere in the null parameter space, $\mathcal{C} = \{\boldsymbol{\theta} : \mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}\}$, and choice of a null parameter value determines distributional properties of L_{12}. Fortunately, it has been shown by Robertson and Wegman (1978) that the least favourable null value occurs when the constraints in the null hypothesis are active, that is, when $\mathbf{R}\boldsymbol{\theta} = \mathbf{0}$.

The p-value for a test of H_1 against $H_2 - H_1$ is, accordingly, given by

$$\begin{aligned} p\text{-}value &= \sup_{\boldsymbol{\theta}\in\mathcal{C}} pr_{\boldsymbol{\theta}}(L_{12} \geq l_{12}) \\ &= pr_{\mathbf{R}\boldsymbol{\theta}=\mathbf{0}}(L_{12} \geq l_{12}). \end{aligned}$$

Formal statement of a theorem regarding the distribution of these particular test statistics can now be made. For details of these results, refer to Shapiro (1988).

Theorem 2.1 Let $\mathbf{X} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a known $p \times p$ positive definite

matrix, $\mathbf{R}$ be a matrix of order $s \times p$, $rank(\mathbf{R}) = s \leq p$, and let $\mathbf{R}_1$ be a submatrix of $\mathbf{R}$ of order $t \times p$. Consider hypotheses

$$H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{0} \text{ and } H_1 : \mathbf{R}_1\boldsymbol{\theta} \geq \mathbf{0}.$$

Then the null distribution of the LRT statistic, T_{01}, for a test of H_0 against $H_1 - H_0$ is given by

$$pr\{T_{01} \geq c\} = \sum_{i=0}^{t} w_i(t, \mathbf{R}_1\boldsymbol{\Sigma}\mathbf{R}_1, \mathbb{R}^{+t}) pr(\chi^2_{s-t+i} \geq c),$$

where $w_i(t, \mathbf{V}, \mathcal{C})$ are the probabilities that $P(\mathbf{Z}|\mathcal{C})$ with $\mathbf{Z} \sim N_t(\mathbf{0}, \mathbf{V})$ lies on exactly i-dimensional faces of $\mathcal{C}$, and sum to 1.

The chi-bar-squared distribution is a weighted mean of several chi-squared distributions. A testing problem involving inequality constraints under the null hypothesis against an unrestricted alternative is given in the next theorem.

Theorem 2.2 Let $X \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, $\mathbf{R}_1$ be a $s \times p$ matrix, $\mathbf{R}_2$ be a $t \times p$ matrix, and the rank of $[\mathbf{R}_1^T, \mathbf{R}_2^T]$ is $s + t$. Let the null and alternative hypotheses be

$$H_1 : \mathbf{R}_1\boldsymbol{\theta} \geq \mathbf{0}, \mathbf{R}_2\boldsymbol{\theta} = \mathbf{0} \text{ and } H_2 : \boldsymbol{\theta} \text{ is unrestricted.}$$

Then the least favourable null distribution of the LRT statistic, T_{12}, for a test of H_1 against $H_2 - H_1$ is obtained when the constraints are active: $H_1 : \mathbf{R}_1\boldsymbol{\theta} = \mathbf{0}, \mathbf{R}_2\boldsymbol{\theta} = \mathbf{0}$. The distribution is then given by

$$pr\{T_{12} \geq c\} = \sum_{i=0}^{s} w_{s-i}(s, \mathbf{A}, \mathbb{R}^{+s}) pr(\chi^2_{t+i} \geq c),$$

where $\mathbf{A} = \mathbf{R}_1\mathbf{\Sigma}\mathbf{R}_1^T - (\mathbf{R}_1\mathbf{\Sigma}\mathbf{R}_2^T)(\mathbf{R}_2\mathbf{\Sigma}\mathbf{R}_2^T)^{-1}(\mathbf{R}_2\mathbf{\Sigma}\mathbf{R}_1^T)$.

A third theorem summarizes some important results concerning chi-bar-squared weights.

Theorem 2.3 Let $\mathcal{C}$ be a closed convex cone in $\mathbb{R}^p$ and $\mathbf{\Sigma}$ be a $p \times p$ positive definite covariance matrix. Then the following are true:

1. Let $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$. If $\mathcal{C} = \mathbb{R}^{+p}$, then
   $w_i(p, \mathbf{\Sigma}, \mathcal{C}) = pr\{P(\mathbf{X}|\mathcal{C}) \text{ has exactly } i \text{ positive components}\}$.

2. $\sum_{i=0}^{p}(-1)^1 w_i(p, \mathbf{\Sigma}, \mathcal{C}) = 0$

3. $0 \leq w_i(p, \mathbf{\Sigma}, \mathcal{C}) \leq 0.5$.

4. Let $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^p : \mathbf{R}\boldsymbol{\theta} \geq 0\}$ where $\mathbf{R}$ is a $p \times p$ nonsingular matrix. Then $\bar{\chi}^2(\mathbf{\Sigma}, \mathcal{C}) = \bar{\chi}^2(\mathbf{R}\mathbf{\Sigma}\mathbf{R}^T, \mathbb{R}^{+p})$ and $w_i(p, \mathbf{\Sigma}, \mathcal{C}) = w_i(p, \mathbf{R}\mathbf{\Sigma}\mathbf{R}^T)$.

5. $w_i(p, \mathbf{\Sigma}) = w_{p-i}(p, \mathbf{\Sigma}^{-1})$

6. Let $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^p : \mathbf{R}\boldsymbol{\theta} \geq 0\}$ where $\mathbf{R}$ is a $k \times p$ of rank $k \leq p$. Then $w_{p-k+i}(p, \mathbf{\Sigma}, \mathcal{C}) = w_i(k, \mathbf{R}\mathbf{\Sigma}\mathbf{R}^T)$ for $i = 0, ..., k$ and 0 otherwise.

7. Let $\mathcal{C} = \{\boldsymbol{\theta} \in \mathbb{R}^p : \mathbf{R}_1\boldsymbol{\theta} \geq \mathbf{0}, \mathbf{R}_2\boldsymbol{\theta} = \mathbf{0}\}$ where $\mathbf{R}_1$ is $s \times p$, $\mathbf{R}_2$ is $t \times p$, $s + t \leq p$, $[\mathbf{R}_1^T, \mathbf{R}_2^T]$ is of full rank, and $\mathbf{A} = \mathbf{R}_1\mathbf{\Sigma}\mathbf{R}_1^T - (\mathbf{R}_1\mathbf{\Sigma}\mathbf{R}_2^T)(\mathbf{R}_2\mathbf{\Sigma}\mathbf{R}_2^T)^{-1}(\mathbf{R}_2\mathbf{\Sigma}\mathbf{R}_1^T)$. Then $w_{p-s-t+j}(p, \mathbf{\Sigma}, \mathcal{C}) = w_j(s, \mathbf{A}, \mathbb{R}^{+s})$ for $j = 0, ..., s$ and 0 otherwise.

Note that the number of chi-squares in the sum depends only on the number of inequality constraints involved in the test.

In order to compute the critical values and/or the p-values for a chi-bar-squared statistic, the weights will have to be calculated. Exact values for the weights, however,

are generally difficult to obtain for $p > 4$. The first point of Theorem 2.3 gives a nice geometric interpretation of the chi-bar-squared weights. Take the case where $p = 2$ as an example with $\mathbf{\Sigma} = I_{2\times 2}$ and $\mathcal{C}$ being the nonnegative orthant cone, which would coincide with the first quadrant of the Cartesian plane. The projection of any vector in quadrants II or IV onto $\mathcal{C}$ would be a vector on either the positive y axis or the positive x axis, and therefore have only a single positive component. The projection of any vector in the third quadrant onto $\mathcal{C}$ would lie on the origin, giving it 0 positive components. And finally, any vector in the first quadrant will already be contained in $\mathcal{C}$, giving the projection exactly 2 positive components. Hence, the probabilities associated with this cone would be as follows: $w_0 = 0.25$, $w_1 = 0.5$ and $w_2 = 0.25$. But increases in p along with different covariance matrices complicate the geometry such that finding closed form solutions for the weights become increasingly difficult. Fortunately, computing them via simulation is a viable option. Section 3.5 of Silvapulle and Sen (2005) give instructions on how to run such simulations. An algorithm to compute $w_i(p, \mathbf{\Sigma}, \mathbb{R}^{+p})$ for $i = 1, ..., p$ is given next.

1. Generate $\mathbf{Z}$ from $N_p(\mathbf{0}, \mathbf{\Sigma})$
2. Compute $\tilde{\mathbf{Z}}$, the point at which $(\mathbf{Z}-\boldsymbol{\theta})^T\mathbf{\Sigma}^{-1}(\mathbf{Z}-\boldsymbol{\theta})$ is the minimum over $\boldsymbol{\theta} \geq \mathbf{0}$
3. Count the number of positive components of $\tilde{\mathbf{Z}}$
4. Repeat the previous three steps N times
5. Estimate $w_i(p, \mathbf{\Sigma}, \mathbb{R}^{+p})$ by the proportion of times $\tilde{\mathbf{Z}}$ has exactly i positive components, $i = 1, ..., p$

Bartholomew (1959a, 1959b) considered tests involving the mean parameter of the type $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$ against ordered alternative hypotheses while Kudo (1963) advanced this earlier work by making the null hypothesis more general. In his

paper, a closed-form solution for the weights when $p = 3$ is given.

$$
\begin{aligned}
w_3(3, \mathbf{\Sigma}) &= (4\pi)^{-1}(2\pi - cos^{-1}\rho_{12} - cos^{-1}\rho_{13} - cos^{-1}\rho_{23}), \\
w_2(3, \mathbf{\Sigma}) &= (4\pi)^{-1}(3\pi - cos^{-1}\rho_{12.3} - cos^{-1}\rho_{13.2} - cos^{-1}\rho_{23.1}), \\
w_1(3, \mathbf{\Sigma}) &= \frac{1}{2} - w_3(3, \mathbf{\Sigma}), \\
w_0(3, \mathbf{\Sigma}) &= \frac{1}{2} - w_2(3, \mathbf{\Sigma}),
\end{aligned}
$$

where ρ_{ij} is the correlation coefficient $\sigma_{ij}\{\sigma_{ii}\sigma_{jj}\}^{-1/2}$, and $\rho_{ij.k}$ is the partial correlation coefficient $(\rho_{ij} - \rho_{ik}\rho_{jk})/\{(1 - \rho_{ik}^2)(1 - \rho_{jk}^2)\}^{1/2}$.

For $p = 2$, the weights are found by

$$
\begin{aligned}
w_0(2, \mathbf{\Sigma}) &= \frac{1}{2}\pi^{-1}cos^{-1}\rho_{12}, \\
w_1(2, \mathbf{\Sigma}) &= \frac{1}{2}, \\
w_2(2, \mathbf{\Sigma}) &= \frac{1}{2} - w_0(2, \mathbf{\Sigma}).
\end{aligned}
$$

Chapter 3

Testing Random Components Under Order Restrictions

In Chapter 2, it was stated that in testing for inequalities on the mean parameter of a multivariate normal random variable, the null distribution of the LRT statistic was chi-bar-squared where the weights depended on the covariance matrix of the variable as well as the convex cone defined in the test. However, the LRT certainly isn't the only test available for order restricted hypotheses. It will be shown in this chapter that versions of the Wald statistic as well as score statistic also have multivariate analogs of the single parameter one-sided test. Particularly in constrained inference, the Wald statistic has an advantage over the LRT statistic due to its computational simplicity. We will use this method in developing an order restricted test for comparing random effects. We note here that higher dependency is expected if observations share a random component with larger variability. Since smaller clusters result in higher dependence, it is natural to assume that random effects are reversely ordered according to their corresponding cluster sizes. Thus, we are interested in testing problems based on the following hypotheses:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2,$$
$$H_1 : \sigma_1^2 \leq \sigma_2^2 \leq \cdots \leq \sigma_k^2,$$

H_2 : Unconstrained,

where random components $u_i's$ are independently distributed with $N(0, \sigma_i^2)$, $i = 1, ..., k$, respectively.

3.1 Asymptotic Tests with Linear Inequalities

This section will begin with the definition of some notation that will be used often throughout the rest of the chapter. Let $Y_1, Y_2, ..., Y_n$ be iid with common density function $f(y; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$. Also let, $l(\boldsymbol{\theta}) = \sum \log f(Y_i; \boldsymbol{\theta})$, $\mathcal{S}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ be the score vector, and $\mathcal{I}_{\boldsymbol{\theta}} = \mathcal{I}(\boldsymbol{\theta}) = -\lim_{n\to\infty} \frac{1}{n} E\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]$ be the Fisher information matrix in the more general sense.

3.1.1 Test Statistics

Much has been said about the LRT statistic in chapter 2 where its null distribution under inequalities was described as being chi-bar-squared, but other popular methods do exist for testing these hypotheses. In the literature, Wald-type tests and score tests are also very common in statistical inference.

The score statistic, also called the Langrange multiplier statistic, for a test of $K_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $K_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ is based on the idea that when H_0 is true, $n^{-1/2} S(\boldsymbol{\theta}_0)$ converges in distribution to a normal random variable with mean $\mathbf{0}$ and variance $\mathcal{I}_{\boldsymbol{\theta}_0}$. It consequently holds that the statistic, $n^{-1} S(\boldsymbol{\theta}_0)^T \mathcal{I}_{\boldsymbol{\theta}_0}^{-1} S(\boldsymbol{\theta}_0)$, is also chi-squared. One way to extend this test to a problem involving $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}$ and $H_1 : \mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}$ is to consider the difference between $S(\bar{\boldsymbol{\theta}})$ and $S(\boldsymbol{\theta}^*)$, where $\bar{\boldsymbol{\theta}}$ is the estimate under H_0 and $\boldsymbol{\theta}^*$ is the estimate under H_1. If H_0 is true then $[S(\bar{\boldsymbol{\theta}}) - S(\boldsymbol{\theta}^*)]$ is expected to be close to 0. However, if H_1 is true then the difference is not expected to be close to 0. Robertson et al. (1988) has suggested the use of the following score statistic to test H_0 against $H_1 - H_0$:

$$\mathbf{S}_{01} = n^{-1}\{S(\bar{\boldsymbol{\theta}}) - S(\boldsymbol{\theta}^*)\}^T \mathcal{I}(\bar{\boldsymbol{\theta}})^{-1}\{S(\bar{\boldsymbol{\theta}}) - S(\boldsymbol{\theta}^*)\}.$$

The Wald statistic is based on $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx N_p\{\mathbf{0}, \mathcal{I}(\boldsymbol{\theta})^{-1}\}$. If $\hat{\mathcal{I}}$ is a consistent estimator of the Fisher information, then the usual Wald statistic for a test of $K_0^* : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}$ against $K_1^* : \mathbf{R}\boldsymbol{\theta} \neq \mathbf{0}$ is given by $W = n(\mathbf{R}\hat{\boldsymbol{\theta}})^T \mathbf{R}\hat{\mathcal{I}}^{-1}\mathbf{R}^T(\mathbf{R}\hat{\boldsymbol{\theta}})$. The statistic is chi-squared distributed with r degrees of freedom. One interpretation of the Wald statistic is that it is a measure of the squared distance between the MLE's of parameters under the null and alternative hypotheses. This concept can be applied to constrained tests. Consider $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}$, $H_1 : \mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}$ and H_2 with no constraints and let $\bar{\boldsymbol{\theta}}$, $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$ denote estimators under those hypotheses, respectively. Then test statistics T_{01} and T_{12} can now be defined for tests of H_0 against $H_1 - H_0$ and H_1 against $H_2 - H_1$, respectively, as follows:

$$\begin{aligned} T_{01} &= n(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}})^T \hat{\mathcal{I}}(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}) \\ &= n(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})^T \hat{\mathcal{I}}(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) - n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \hat{\mathcal{I}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \qquad (21) \\ T_{12} &= n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \hat{\mathcal{I}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &= n(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})^T \hat{\mathcal{I}}(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) - n(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}})^T \hat{\mathcal{I}}(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}). \qquad (22) \end{aligned}$$

Note that the second equality comes from the Pythagorean theorem.

Alternatively, the notation can be condensed by writing the test statistics using norms:

$$T_{01} = n\|\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}\|^2_{\hat{\mathcal{I}}^{-1}} - n\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2_{\hat{\mathcal{I}}^{-1}}, \qquad (23)$$

$$T_{12} = n\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2_{\hat{\mathcal{I}}^{-1}}. \qquad (24)$$

One might also be interested in testing H_0 against $H_2 - H_0$. In this case, the test

statistic is given by summing T_{01} and T_{12} to give

$$\begin{aligned} T_{02} &= T_{01} + T_{12} \\ &= n\|\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}\|^2_{\hat{\mathcal{I}}^{-1}}. \end{aligned} \tag{25}$$

We note here that T_{02} is a usual unconstrained test for H_0.

3.1.2 The Null Distribution

Throughout these derivations it is assumed that the following regularity conditions, originally formulated by Cramer (1946), are valid.

Regularity Conditions $\mathcal{R}$.

1. The first 3 derivatives of $logf(x;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist almost surely.
2. $\left|\frac{\partial f}{\partial \theta_i}\right| < F(x)$, $\left|\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}\right| < F(x)$, $\left|\frac{\partial^3 \log f}{\partial \theta_i \partial \theta_j \partial \theta_k}\right| < H(x)$
   where F is finitely integrable and $E\{H(x)\} = M < \infty$.
3. The Fisher information matrix is finite and positive definite.

Given conditions, $\mathcal{R}$, the following standard results can be obtained:

Lemma 3.1 Under the true value, $\boldsymbol{\theta}_0$, of $\boldsymbol{\theta}$, we have:

1. $n^{-1/2}\mathbf{S}(\boldsymbol{\theta}_0) \xrightarrow{d} N_p\{\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)\}$
2. $n^{-1/2}\mathcal{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{S}(\boldsymbol{\theta}_0) = n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1)$
3. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N_p\{\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}\}$

Moreover, when the regularity conditions in $\mathcal{R}$ are satisfied, a pair of useful quadratic approximations may also be obtained:

Lemma 3.2 Let $\mathbf{u} = \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ and $K > 0$ be given. Then

1. $l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}_0) + (1/2)n^{-1}\mathbf{S}(\boldsymbol{\theta}_0)^T\mathcal{I}_{\boldsymbol{\theta}_0}^{-1}\mathbf{S}(\boldsymbol{\theta}_0) - (1/2)(\mathbf{Z}_n - \mathbf{u})^T\mathcal{I}_{\boldsymbol{\theta}_0}(\mathbf{Z}_n - \mathbf{u}) + \delta_n(\mathbf{u})$
   where $\mathbf{Z}_n = n^{-1/2}\mathcal{I}_{\boldsymbol{\theta}_0}^{-1}\mathbf{S}(\boldsymbol{\theta}_0)$ and $\sup_{\|\mathbf{u}\|<\mathbf{K}} |\delta_n(\mathbf{u})| = o_p(1)$
2. $l(\boldsymbol{\theta}) = l(\hat{\boldsymbol{\theta}}) - (1/2)(\mathbf{Z}_n - \mathbf{u})^T\mathcal{I}_{\boldsymbol{\theta}_0}(\mathbf{Z}_n - \mathbf{u}) + \delta_n(\mathbf{u})$
   where $\mathbf{Z}_n = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and $\sup_{\|\mathbf{u}\|<\mathbf{K}} |\delta_n(\mathbf{u})| = o_p(1)$

The first approximation is based on the Taylor series expansion of $l(\boldsymbol{\theta})$ about $\boldsymbol{\theta}_0$, and a re-expression in terms of $\mathbf{Z}_n$. Derivation of the second approximation can be found in Silvapulle (1994).

Using the above conditions and Lemmas, the asymptotic distribution for the distance statistic, T_{01} and T_{12}, as defined in the previous section can now be derived. If the conditions in $\mathcal{R}$ hold, then the second quadratic approximation for the log-likelihood given in Lemma 3.2 can be applied towards the likelihood ratio test to give the following result:

$$L_{01} = 2[\sup\{l(\boldsymbol{\theta}) : \mathbf{R}\boldsymbol{\theta} \geq \mathbf{0}\} - \sup\{l(\boldsymbol{\theta}) : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}\}].$$

Now under the null hypothesis H_0, $\mathbf{R}\boldsymbol{\theta}_0 = \mathbf{0}$, which implies that $\mathbf{Ru} = \sqrt{n}(\mathbf{R}\boldsymbol{\theta} - \mathbf{R}\boldsymbol{\theta}_0) = \sqrt{n}\mathbf{R}\boldsymbol{\theta}$. Hence, $\mathbf{R}\boldsymbol{\theta} \geq \mathbf{0} \iff \mathbf{Ru} \geq \mathbf{0}$ and $\mathbf{R}\boldsymbol{\theta} = \mathbf{0} \iff \mathbf{Ru} = \mathbf{0}$. Therefore, the LRT statistic becomes

$$\begin{aligned} L_{01} = 2[&\sup_{\mathbf{Ru}\geq\mathbf{0}} \{l(\hat{\boldsymbol{\theta}}) - (1/2)(\mathbf{Z}_n - \mathbf{u})^T\mathcal{I}_{\boldsymbol{\theta}_0}(\mathbf{Z}_n - \mathbf{u}) + \delta_n(\mathbf{u})\} \\ &- \sup_{\mathbf{Ru}=\mathbf{0}} \{l(\hat{\boldsymbol{\theta}}) - (1/2)(\mathbf{Z}_n - \mathbf{u})^T\mathcal{I}_{\boldsymbol{\theta}_0}(\mathbf{Z}_n - \mathbf{u}) + \delta_n(\mathbf{u})\}]. \end{aligned}$$

Since the first term in the quadratic approximation is a function of the unrestricted MLE and not of $\boldsymbol{\theta}$, this simplifies to

$$\begin{aligned}
L_{01} &= 2[\sup_{\mathbf{Ru}\geq\mathbf{0}}\{-(1/2)(\mathbf{Z}_n-\mathbf{u})^T\mathcal{I}_{\boldsymbol{\theta}_0}(\mathbf{Z}_n-\mathbf{u})+\delta_n(\mathbf{u})\} \\
&\quad - \sup_{\mathbf{Ru}=\mathbf{0}}\{-(1/2)(\mathbf{Z}_n-\mathbf{u})^T\mathcal{I}_{\boldsymbol{\theta}_0}(\mathbf{Z}_n-\mathbf{u})+\delta_n(\mathbf{u})\}] \\
&= \inf_{\mathbf{Ru}=\mathbf{0}}\{(\mathbf{Z}_n-\mathbf{u})^T\mathcal{I}_{\boldsymbol{\theta}_0}(\mathbf{Z}_n-\mathbf{u})\} - \inf_{\mathbf{Ru}\geq\mathbf{0}}\{(\mathbf{Z}_n-\mathbf{u})^T\mathcal{I}_{\boldsymbol{\theta}_0}(\mathbf{Z}_n-\mathbf{u})\}+o_p(1) \\
&= \inf_{\mathbf{R}\boldsymbol{\theta}=\mathbf{0}}\{[\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)-\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)]^T\mathcal{I}_{\boldsymbol{\theta}_0}[\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)-\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)]\} \\
&\quad - \inf_{\mathbf{R}\boldsymbol{\theta}\geq\mathbf{0}}\{[\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)-\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)]^T\mathcal{I}_{\boldsymbol{\theta}_0}[\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)-\sqrt{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)]\}+o_p(1) \\
&= \inf_{\mathbf{R}\boldsymbol{\theta}=\mathbf{0}}\{n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\mathcal{I}_{\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\} - \inf_{\mathbf{R}\boldsymbol{\theta}\geq\mathbf{0}}\{n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\mathcal{I}_{\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\}+o_p(1).
\end{aligned}$$

Replacing $\mathcal{I}_{\boldsymbol{\theta}_0}$ with its consistent estimator, $\hat{\mathcal{I}}_{\boldsymbol{\theta}}$, we can show

$$L_{01} = T_{01} + o_p(1). \tag{26}$$

Furthermore, since $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \xrightarrow{d} N_p(\mathbf{0}, \mathcal{I}_{\boldsymbol{\theta}_0}^{-1})$ by Lemma 3.1, it follows by Theorem 2.1 that for any true value, $\boldsymbol{\theta}_0$, of $\boldsymbol{\theta}$ in $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{0}$, we have

$$\lim_{n\to\infty} pr\{T_{01} \geq t_{01}\} = \sum_{i=0}^{r} w_i(r, \mathbf{R}\mathcal{I}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}^T, \mathbb{R}^{+r}) pr(\chi_i^2 \geq t_{01}). \tag{27}$$

Similar arguments apply to the LRT statistic, L_{12}, for testing H_1 against $H_2 - H_1$. That is, $L_{12} = T_{12} + o_p(1)$ and the null distribution for T_{12} is given by

$$\lim_{n\to\infty} pr\{T_{01} \geq t_{01}\} = \sum_{i=0}^{r} w_{r-i}(r, \mathbf{R}\mathcal{I}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}^T, \mathbb{R}^{+r}) pr(\chi_i^2 \geq t_{12}). \tag{28}$$

As we discussed earlier, the LRT statistic and Rao's score-based statistic are asymptotically equivalent to their corresponding Wald's statistic and have the same limiting

null distributions as given in (27) and (28).

3.1.3 Testing Ordered Random Effects in GLMMs

The parameters to be estimated in the generalized linear mixed model of (1) and (2) include the regression parameters, $\boldsymbol{\beta} = (\beta_0, ..., \beta_s)^T$, and the variance components, $\mathbf{D} = (\sigma_1^2, ..., \sigma_k^2)^T$, when we assume $\tau = 1$. However, in this situation the regression parameters represent a set of nuisance parameters in the testing procedure because our interest is centred on the random components. Thus, considering a $(k-1) \times k$ matrix $\mathbf{R} = [r_{ij}]$ where

$$r_{ij} = \begin{cases} -1 & \text{if } 1 \leq i = j \leq k-1 \\ 1 & \text{if } j = i+1,\ 1 \leq i \leq k-1 \\ 0 & \text{otherwise,} \end{cases}$$

we can rewrite the testing hypotheses, H_0 and H_1, as

$$H_0 : \{\mathbf{D} : \mathbf{RD} = \mathbf{0}\} \text{ and } H_1 : \{\mathbf{D} : \mathbf{RD} \geq \mathbf{0}\}.$$

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \mathbf{D}^T)^T$ and denote its information matrix by $\mathcal{I}_{\boldsymbol{\theta}}$, as defined at the beginning of Section 3.1. If we introduce an augmented matrix

$$\mathbf{Q} = \left[\mathbf{O} \vdots \mathbf{R} \right],$$

where $\mathbf{O}$ is a $(k-1) \times s$ null matrix, then the hypotheses are equivalent to

$$H_0 : \mathbf{Q}\boldsymbol{\theta} = \mathbf{0} \text{ and } H_1 : \mathbf{Q}\boldsymbol{\theta} \geq \mathbf{0}$$

because $\mathbf{Q}\boldsymbol{\theta} = \mathbf{RD}$. By applying the Schur complement to a block partitioned matrix, we can show that

$$\mathbf{Q}\mathcal{I}_{\boldsymbol{\theta}}^{-\mathbf{1}}\mathbf{Q}^{\mathbf{T}} = \mathbf{R}(\mathbf{I}_{\mathbf{DD}} - \mathbf{I}_{\mathbf{D}\boldsymbol{\beta}}\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-\mathbf{1}}\mathbf{I}_{\boldsymbol{\beta}\mathbf{D}})^{-\mathbf{1}}\mathbf{R}^{\mathbf{T}},$$

where I_{ab} are submatrices of

$$\mathcal{I}_{\boldsymbol{\theta}} = \begin{bmatrix} I_{\boldsymbol{\beta}\boldsymbol{\beta}} & I_{\boldsymbol{\beta}\mathbf{D}} \\ I_{\mathbf{D}\boldsymbol{\beta}} & I_{\mathbf{DD}} \end{bmatrix}.$$

Let $\mathcal{I}^{\mathbf{DD}} = I_{\mathbf{DD}} - I_{\mathbf{D}\boldsymbol{\beta}}I_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}I_{\boldsymbol{\beta}\mathbf{D}}$ and $\hat{\mathcal{I}}^{\mathbf{DD}}$ denote the same expression obtained from the observed Fisher information matrix, $\hat{\mathcal{I}}_{\boldsymbol{\theta}}$. Then by Lemma 3.1 and the results in (27) and (28), the asymptotic null distributions for T_{01} and T_{12} can be derived. The following theorem summarizes the results.

Theorem 3.1 Let $\bar{\boldsymbol{\theta}}$, $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$ be the MLE's of $\boldsymbol{\theta}$ under H_0, H_1 and H_2, respectively. Then, for any true value, $\boldsymbol{\theta}_0$, of $\boldsymbol{\theta}$ in H_0, we have the asymptotic null distributions for $T_{01} = n\|\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}\|^2_{\hat{\mathcal{I}}_{\boldsymbol{\theta}}^{-1}}$ and $T_{12} = n\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2_{\hat{\mathcal{I}}_{\boldsymbol{\theta}}^{-1}}$ given by

$$\lim_{n\to\infty} pr\{T_{01} \geq t_{01}\} = \textstyle\sum_{i=0}^{k-1} w_i(k-1, \mathbf{R}(\mathcal{I}^{\mathbf{DD}})^{-1}\mathbf{R}^T, \mathbb{R}^{+(k-1)})pr(\chi_i^2 \geq t_{01})$$

and

$$\lim_{n\to\infty} pr\{T_{12} \geq t_{12}\} = \textstyle\sum_{i=0}^{k-1} w_{k-1-i}(k-1, \mathbf{R}(\mathcal{I}^{\mathbf{DD}})^{-1}\mathbf{R}^T, \mathbb{R}^{+(k-1)})pr(\chi_i^2 \geq t_{12}).$$

These asymptotic null distributions depend on unknown parameters, $\boldsymbol{\theta}_0$, through $\mathcal{I}^{\mathbf{DD}}$ but in a practical situation we may replace it by its consistent estimator, $\hat{\mathcal{I}}^{\mathbf{DD}}$. Another point is that finding the constrained MLE's of parameters in GLMMs is computationally very demanding as will be discussed in Section 3.2. To avoid this problem, we may use least squares estimators that can be obtained by minimizing $\mathbf{G}(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\mathcal{I}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ under H_0 and H_1, respectively. Denoting these estimators by $\bar{\boldsymbol{\theta}}_0$ and $\boldsymbol{\theta}_0^*$, we can approximate Wald's test statistics as $T_{01}^0 = n[\mathbf{G}(\bar{\boldsymbol{\theta}}_0) - \mathbf{G}(\boldsymbol{\theta}_0^*)]$ and $T_{12}^0 = n\mathbf{G}(\boldsymbol{\theta}_0^*)$.

3.2 Estimation

It was mentioned in Section 2.1.2 that there are a number of ways of calculating maximum likelihood estimates under the framework of a GLMM. In the next section, one method in particular will be presented that is to be applied towards a case study in Chapter 4. Estimates of both the regression parameters and random components can be obtained using a method suggested by McCulloch (1997) that involves a Monte Carlo simulation of the random effects via a Metropolis algorithm and Newton-Raphson iteration. It will be seen that simulating the distribution of $\mathbf{u}$ given $\mathbf{y}$ is required to calculate expectations involving this conditional distribution.

3.2.1 Unconstrained Maximum Likelihood Estimation

Monte Carlo Newton-Raphson

Obtaining unconstrained estimates will begin with the ML estimating equations $\frac{\partial l}{\partial \boldsymbol{\beta}} = E\left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}) \Big| \mathbf{y}\right]$ and $\frac{\partial l}{\partial \mathbf{D}} = E\left[\frac{\partial}{\partial \mathbf{D}} \log f_{\mathbf{u}}(\mathbf{u}|\mathbf{D}) \Big| \mathbf{y}\right]$. The second equation involving the variance components has a closed-form solution when the distribution is assumed to be normal. If the conditional distribution of the random effects, $\mathbf{u}|\mathbf{y}$, can be simulated, then the maximum likelihood estimates can be obtained by computing Monte Carlo averages for these otherwise intractable expectations. The estimating equation associated with the regression parameter is a little more complicated than the one for the variance components, but it is amenable to Newton-Rapshon. The MLE's for $\boldsymbol{\beta}$ can be found by first expanding the term within the expectation as a second order Taylor series around the true value $\boldsymbol{\beta}_0$.

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u},\boldsymbol{\beta}) \approx \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u},\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u},\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\boldsymbol{\beta}-\boldsymbol{\beta}_0). \tag{29}$$

Given the results from Section 2.1, the right-hand side of this equation can be written as

$$\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta}_0,\mathbf{U})\boldsymbol{\Delta}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}(\mathbf{Y}-\boldsymbol{\mu}(\boldsymbol{\beta}_0,\mathbf{U})) + \mathbf{X}^T\mathbf{W}(\boldsymbol{\beta_0},\mathbf{U})\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0), \tag{30}$$

where $\mu_i(\boldsymbol{\beta},\mathbf{u}) = E[Y_i|\mathbf{u}]$, $\mathbf{W}(\boldsymbol{\beta},\mathbf{u})^{-1} = diag\{(\frac{\partial \eta_i}{\partial \mu_i})^2 \text{var}(Y_i|\mathbf{u})\}$, and $\boldsymbol{\Delta} = diag\{\frac{\partial \eta_i}{\partial \mu_i}\}$. Inserting (30) into the ML estimating equation for $\boldsymbol{\beta}$ produces the following iterative equation:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \text{E}[\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta}^{(m)},\mathbf{U})\mathbf{X}|\mathbf{y}]^{-1} \times E\left[\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta}^{(m)},\mathbf{U})\boldsymbol{\Delta}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}} \times (\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^{(m)},\mathbf{U}))|\mathbf{y}\right]. \tag{31}$$

Now because the expectations in (31) generally cannot be computed in closed form, a Monte Carlo approximation may be used instead. McCulloch (1997) suggests obtaining a simulated sample from the conditional distribution of $\mathbf{u}$ given $\mathbf{y}$ by using a Metropolis algorithm.

The candidate distribution for the Metropolis algorithm from which potential new draws are to be made will be $f_{\mathbf{u}}$. The advantage of selecting this as the candidate distribution is that it will simplify the acceptance function in a convenient way. Now suppose that $\mathbf{u}$ is a previous selection from the conditional distribution

$\mathbf{u}|\mathbf{y}$. A new value, u_k^* for the k^{th} element of $\mathbf{u}$, is generated from $f_{\mathbf{u}}$. The candidate, $\mathbf{u}^*$ can then be inserted into the acceptance function along with $\mathbf{u}$, where, $\mathbf{u} = (u_1, u_2, ..., u_{k-1}, u_k, u_{k+1}, ..., u_q)$ and $\mathbf{u}^* = (u_1, u_2, ..., u_{k-1}, u_k^*, u_{k+1}, ..., u_q)$. The acceptance function is specified as

$$A_k(\mathbf{u}, \mathbf{u}^*) = min\left\{1, \frac{f_{u|y}(\mathbf{u}^*|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D})h_u(\mathbf{u})}{f_{u|y}(\mathbf{u}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D})h_u(\mathbf{u})}\right\}. \tag{32}$$

Upon selecting $f_{\mathbf{u}}$ as the candidate distribution, the right side of the term in braces simplifies to

$$\begin{aligned}\frac{f_{u|y}(\mathbf{u}^*|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D})f_u(\mathbf{u}|\mathbf{D})}{f_{u|y}(\mathbf{u}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{D})f_u(\mathbf{u}|\mathbf{D})} &= \frac{f_{y|u}(\mathbf{y}|\mathbf{u}^*, \boldsymbol{\beta})}{f_{y|u}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta})} \\ &= \frac{\prod_{i=1}^n f_{y_i|u}(y_i|\mathbf{u}^*, \boldsymbol{\beta})}{\prod_{i=1}^n f_{y_i|u}(y_i|\mathbf{u}, \boldsymbol{\beta})}.\end{aligned} \tag{33}$$

Therefore, one needs only to specify the conditional distribution of $\mathbf{y}|\mathbf{u}$ in order to proceed. The new draw, $\mathbf{u}^*$, is accepted with probability $A_k(\mathbf{u}, \mathbf{u}^*)$. If it is rejected then the previously accepted value $\mathbf{u}$ is retained instead. Continually sampling from the distribution of $f_{\mathbf{u}}$ and accepting new values based on the specified acceptance function creates a sample that eventually stabilizes towards realizations from the desired conditional distribution of $\mathbf{u}|\mathbf{y}$. This sample can then be used to calculate Monte Carlo estimates of the expectations in (31).

This Metropolis algorithm can be incorporated into the Newton-Raphson equation to estimate the regression parameters, $\boldsymbol{\beta}$, by using the method outlined next:

1. Set $m = 0$ and choose initial values for $\boldsymbol{\beta}^{(0)}$ and $\mathbf{D}^{(\mathbf{0})}$.

2. Generate N values, $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, ..., \mathbf{u}^{(N)}$ from the distribution $f_{u|y}(\mathbf{u}|\mathbf{y}, \boldsymbol{\beta}^{(\mathbf{m})}, \mathbf{D}^{(\mathbf{m})})$. Use them to calculate the following quantities:

(a) Calculate $\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \mathrm{E}[\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta}^{(m)}, \mathbf{U})\mathbf{X}|\mathbf{y}]^{-1}$

$$\times E\left[\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta}^{(m)},\mathbf{U})\boldsymbol{\Delta}\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}} \times (\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^{(m)},\mathbf{U}))|\mathbf{y}\right]$$

(b) Calculate $\mathbf{D}^{(m+1)}$ by maximizing $\frac{1}{N}\sum_{k=1}^{N} lnf_u(\mathbf{u}^{(k)}|\mathbf{D})$

(c) Set $m = m+1$

3. If convergence is achieved then declare $\boldsymbol{\beta}^{(m+1)}$ and $\mathbf{D}^{(m+1)}$ to be the maximum likelihood estimates. If convergence has not been achieved then begin again at step 2.

The Monte Carlo Newton-Rapshon method that has just been described has been shown by McCulloch (1997), via simulation studies, to be very effective in finding the unconstrained MLE's for both the regression parameters and the variance components.

Observed Fisher Information Matrix

It is well known that the inverse of the Fisher information provides an asymptotic estimate for the variance of the MLE, which makes it an important quantity for estimation. One of the arguments against the MCEM method is that the Fisher information is not calculated during estimation as is typically the case with MCNR. Using McCulloch's MCNR algorithm, however, does not give us the Fisher information for $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\mathbf{D}}^T)^T$ either, since the mixed partial derivatives are never obtained. In fact, finding the Hessian matrix is often a difficult if not overly tedious task, which is one reason researchers do opt for the EM algorithm over Newton-Raphson. With that said, it has already been stated that MCNR is often chosen for its relative speed advantage. One alternative then is to use Louis' method for finding the observed Fisher information matrix, which has been shown by Efron and Hinkley (1978) to be a superior measure of information than the expected Fisher Information. Sinha (2004) developed a robust estimation method for GLMMs by building upon McCulloch's

MCNR algorithm where he also recommends use of the observed Fisher information for finding the variance of MLE's.

The observed Fisher information is the negative of the Hessian evaluated at the MLE. Louis (1982) derived his method for finding the observed Fisher under the framework of the EM algorithm and can be calculated in the following manner.

First define the complete-data, $\mathbf{c}$, in the context of the EM algorithm as the observed data taken jointly with the missing data, $\mathbf{c^T} = (\mathbf{y^T}, \mathbf{u^T})$, where the observed data is $\mathbf{y}$ and the missing data are the unobservable random effects, $\mathbf{u}$. Then the complete-data log-likelihood is given by $\log L_c = \log f_{\mathbf{Y|u}} + \log f_{\mathbf{U}}$. One advantage of this specification is the fact that the regression parameters, $\boldsymbol{\beta}$, enter only through the GLM portion of the equation while the variance components, $\mathbf{D}$, enters only through the random effects.

Louis' formula for the observed Fisher information is

$$\begin{aligned} I_Y(\boldsymbol{\beta}, \mathbf{D}) =& \mathcal{I}(\boldsymbol{\beta}, \mathbf{D}) - \mathrm{E}\{\mathcal{S}_{\mathbf{c}}(\boldsymbol{\beta}, \mathbf{D}; \mathbf{Y}, \mathbf{u}) \mathcal{S}_{\mathbf{c}}^{\mathbf{T}}(\boldsymbol{\beta}, \mathbf{D}; \mathbf{Y}, \mathbf{u}) | \mathbf{Y}\} \\ & + \mathrm{E}\{\mathcal{S}_c(\boldsymbol{\beta}, \mathbf{D}; \mathbf{Y}, \mathbf{u}) | \mathbf{Y}\} \mathrm{E}\{\mathcal{S}_{\mathbf{c}}^{\mathbf{T}}(\boldsymbol{\beta}, \mathbf{D}; \mathbf{Y}, \mathbf{u}) | \mathbf{Y}\}, \end{aligned} \tag{34}$$

where $\mathcal{S}_c(\boldsymbol{\beta}, \mathbf{D}; \mathbf{Y}, \mathbf{U})$ is the first derivative of the complete-data log-likelihood and $\mathcal{I}(\boldsymbol{\beta}, \mathbf{D}) = -\mathrm{E}\left(\frac{\partial^2 \log \mathbf{L_c}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathbf{T}}} | \mathbf{Y}\right)$. Upon evaluation at the MLE, the last term in the equation is effectively zero. Hence, the observed Fisher information involves only two terms, both of which are relatively easy to find because the likelihood of the complete data factors nicely into a product of $f_{\mathbf{Y|u}}$ and $f_{\mathbf{U}}$, allowing for the separation of the regression parameters and the variance components upon taking logs.

3.2.2 Constrained Estimation

Section 2.2.3 described the gradient projection algorithm for the maximization of an arbitatry objective function under linear equality and inequality constraints. Jamshidian (2004) applied this method towards maximum likelihood estimation to obtain constrained MLE's. In Section 3.2.1 it was found that employing a distance statistic had a computational advantage over the LRT and score statistics because we can avoid constrained maximization in each cycle of the algorithm. In fact, with respect to the distance statistic, once an unrestricted MLE has been found, we need only to project it onto the parameter spaces defined in the hypothesis to compute a test statistic. That is, the test statistic can be found by first solving the following quadratic forms:

$$\underset{\boldsymbol{\theta}\in\{\boldsymbol{\theta}:\mathbf{R}\boldsymbol{\theta}=\mathbf{0}\}}{\operatorname{argmin}} \mathbf{G}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\{\boldsymbol{\theta}:\mathbf{R}\boldsymbol{\theta}=\mathbf{0}\}}{\operatorname{argmin}} (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\hat{\mathcal{I}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$$

and

$$\underset{\boldsymbol{\theta}\in\{\boldsymbol{\theta}:\mathbf{R}\boldsymbol{\theta}\geq\mathbf{0}\}}{\operatorname{argmin}} \mathbf{G}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\{\boldsymbol{\theta}:\mathbf{R}\boldsymbol{\theta}\geq\mathbf{0}\}}{\operatorname{argmin}} (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\hat{\mathcal{I}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}).$$

The solutions to these quadratic forms are also the solutions to the maximization problem involving the negative of $\mathbf{G}(\boldsymbol{\theta})$. Therefore, if the objective function is defined to be $\mathbf{G}(\boldsymbol{\theta}) = -\frac{1}{2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\hat{\mathcal{I}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$, then the Gradient Projection algorithm of Section 2.2.3 can be applied towards finding a solution to this quadratic maximization problem where the constraints are given by

$$\begin{aligned} a_i^T\boldsymbol{\theta} &= 0, i \in I_1, \\ a_i^T\boldsymbol{\theta} &\leq 0, i \in I_2. \end{aligned} \tag{35}$$

As outlined in Chapter 2, the algorithm begins with an initial value $\boldsymbol{\theta}_r$ that satisfies the constraints before cycling through the next steps.

1. Compute $\mathbf{d} = P_W(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_r)$, where $P_w = I - (\hat{\mathcal{I}}_{\boldsymbol{\theta}})^{-1}\bar{A}^T(\bar{A}(\hat{\mathcal{I}}_{\boldsymbol{\theta}})^{-1}\bar{A}^T)^{-1}\bar{A}$.

2. If $\mathbf{d} = 0$, compute the Lagrange multipliers $\lambda = (\bar{A}(\hat{\mathcal{I}}_{\boldsymbol{\theta}})^{-1}\bar{A}^T)^{-1}\bar{A}\tilde{g}(\theta_r)$. Let λ_i denote the i^{th} component of λ.

   If $\lambda_i \geq 0$ for all $i \in \mathcal{W} \cap I_2$, stop.

   If there is at least one negative λ_i for $i \in \mathcal{W} \cap I_2$, determine the index associated with the smallest λ_i and remove it from the set $\mathcal{W}$. Modify $\bar{A}$ and $\bar{\mathbf{b}}$ as well by dropping the corresponding row from each. Go to step 1.

3. If $\mathbf{d} \neq 0$, obtain $\alpha_1 = \max_i\{\frac{0-(\bar{A}\boldsymbol{\theta}_r)_i}{(\bar{A}\mathbf{d})_i}\}$ and $\alpha_2 = \min\{\alpha_1, \frac{\mathbf{d}^T\hat{\mathcal{I}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_r)}{\mathbf{d}^T\hat{\mathcal{I}}_{\boldsymbol{\theta}}\mathbf{d}}\}$. Set $\tilde{\boldsymbol{\theta}}_r = \boldsymbol{\theta}_r + \alpha_2\mathbf{d}$. Add indices of new coordinates, if any, of $\tilde{\boldsymbol{\theta}}_r$ that are newly on the boundary of the working set $\mathcal{W}$. Modify $\bar{A}$ and $\bar{\mathbf{b}}$ by adding new rows accordingly.

4. Replace $\boldsymbol{\theta}_r$ by $\tilde{\boldsymbol{\theta}}_r$ and go to step 1.

Chapter 4

Analysis of Youth Smoking

In this chapter, a model is specified and formulas are derived for analysing the 2011 National Youth Tobacco Survey data under the framework of a generalized linear mixed model. Parameter estimates will be presented along with the test statistics for an order restricted test on the variance components that were defined in Chapter 3.

4.1 The National Youth Tobacco Survey

The National Youth Tobacco Survey is a survey run by the American Centers for Disease Control and Prevention that was developed to provide the data necessary for the design and implementation of tobacco prevention and control programs targeting youth. Initially beginning in 1999 it has since been conducted on a nearly biennial basis with 2012 being the most recent version of the study. The NYTS provides one of the more comprehensive data sets on tobacco use for both middle school (grades 6-8) and high school (grades 9-12) students.

4.1.1 Sampling Design

The population from which subjects were sampled for the study included all public, Catholic, and other private school students enrolled in regular middle schools and

high schools in grades 6 through 12 across 50 states and the District of Columbia. Institutions such as alternative schools and vocational schools were excluded from the population.

The sampling strategy involved a three-stage cluster design where the primary sampling units (PSUs) were randomly selected counties. From each county, a sample of schools was selected at the second sampling stage, and then finally, whole classes were sampled randomly from within each school. In total there were 82 counties, 194 schools, and 3 to 8 classes per school that were sampled.

The respondent variable that is of main concern in this study is a binary outcome variable that indicates whether a student has ever smoked a cigarette in the past. The adoption of random effects as part of the model is based on the sampling design of the NYTS, which gives way to the presence of nested clusters within the data. The students are clustered within classes which are nested within schools, which are in turn, nested within counties. Dependency amongst students who share some or all of the same clusters may be justified by considering both behavioural and financial influences. It is assumed that students within the same class behave similarly due to such social interaction factors as peer pressure. A similar type of dependency might be reflected at the school level, albeit to a smaller degree because students within the same school are not as socially connected as those who share the same classroom. And finally, there may be some dependence in smoking behaviour amongst students of the same county as some counties are likely to provide more funding towards education and prevention than others. County to county comparisons potentially reflect the disparities in socioeconomic performance that are often written about at the state level in America. Hence this data can be modeled through a generalized linear mixed model with random effects associated with class, school and county where one might expect the level of dependency observed at the class level to be at least as large as the dependency produced by the school level, which is at least as great as the county

effect.

4.1.2 Summary Statistics

The total sample size of the 2011 NYTS study was 17746 students. To reduce this sample down to a more manageable size, 15 counties were randomly sampled from the original data set to reduce the sample down to 3576. In this way, the time required to conduct the data analysis was much more reasonable. Summary statistics for 3 covariates of interest taken from this reduced data set are given in Table 1.

Table 1: 2011 NYTS Summary Statistics

			Response	
Covariate	Level	Sample Size	Never Smoked	Has Smoked
Gender	Female	1785	65.55 %	34.45 %
	Male	1791	69.24 %	30.76 %
Age Group	9-12	203	89.66 %	10.34 %
	13-14	1028	83.07 %	16.93 %
	15-16	2045	60.98 %	39.02 %
	17+	300	42.33 %	57.67 %
Race	White	1803	66.50 %	33.50 %
	Black	896	71.54 %	28.46 %
	Hispanic	720	68.69 %	31.31 %
	Asian	116	66.38 %	33.62 %
	Native American	41	68.29 %	31.71 %

4.2 Comparison of Cluster Effects on Youth Smoking

In this section, the model is specified and formulas are derived for analysing the 2011 National Youth Tobacco Survey data under the framework of a generalized linear mixed model.

4.2.1 Defining the Model

The response variable, y_{ijkl}, represents the smoking status of the l^{th} student of the k^{th} class of the j^{th} school in the i^{th} county. It is an indicator variable that takes a value of 1 if the student has ever smoked and 0 if they have never once tried it. There are fixed effects associated with the student's age, race and gender as well as 3 nested random effects for county, school and class. Since the response is binary, the logit link function will be applied to the conditional expectation of y_{ijkl} given the random effects so that it may be modeled as linear in the predictors. There are m counties in total, n_i schools in county i, n_{ij} classes in school j of county i, and n_{ijk} students in class k of school j of county i. Let $p_{ijkl} = E[Y_{ijkl}|\mathbf{u}]$. Then the data is modeled as

follows:

$$
\begin{aligned}
Y_{ijkl}|\mathbf{u} &\sim \text{independent Bernoulli}(p_{ijkl}) \\
\log\Big[\frac{p_{ijkl}}{1-p_{ijkl}}\Big] &= \beta_0 + \beta_1 Age2_{ijkl} + \beta_2 Age3_{ijkl} + \beta_3 Age4_{ijkl} \\
&\quad + \beta_4 Gender2_{ijkl} + \beta_5 Race2_{ijkl} + \beta_6 Race3_{ijkl} \\
&\quad + \beta_7 Race4_{ijkl} + \beta_8 Race5_{ijkl} \\
&\quad + a_i + b_{ij} + c_{ijk} \\
&= x_{ijkl}^T \boldsymbol{\beta} + a_i + b_{ij} + c_{ijk} \\
&= \eta_{ijkl} \qquad (36)
\end{aligned}
$$

$$
\begin{gathered}
i = 1, 2, ..., m \\
j = 1, 2, ..., n_i \\
k = 1, 2, ..., n_{ij} \\
l = 1, 2, ..., n_{ijk} \\
a_i \sim \mathrm{N}(0, \sigma^2_{county}), b_{ij} \sim \mathrm{N}(0, \sigma^2_{school}), c_{ijk} \sim \mathrm{N}(0, \sigma^2_{class}), \\
a_i, b_{ij} \text{ and } c_{ijk} \text{ are independent of each other.}
\end{gathered}
$$

4.2.2 Estimation Using MCNR

In estimating the regression paramaters and the variance components for the random effects, ML equations (6) and (7) must be solved. We begin by finding an expression for the regression parameters. This involves adapting equation (31) in the MCNR algorithm to the model above.

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}) \approx \sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}} \mathbf{x}_{ijk}^T \mathbf{W}(\boldsymbol{\beta_0}, \mathbf{U}) \frac{\partial \eta_{ijk}}{\partial \mu_{ijk}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\mathbf{Y}_{ijk} - \boldsymbol{\mu}_{ijk}(\boldsymbol{\beta}_0, \mathbf{U})) - \sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}} \mathbf{x}_{ijk}^T \mathbf{W}(\boldsymbol{\beta}_0, \mathbf{U}) \mathbf{x}_{ijk} (\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

Next, insert this expression into equation (6) to get the iterative equation:

$$\begin{aligned}
\boldsymbol{\beta}^{(m+1)} &= \boldsymbol{\beta}^{(m)} + \left\{ \sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}} E\Big[\mathbf{x}_{ijk}^T \mathbf{W}(\boldsymbol{\beta}_0, \mathbf{U}) \mathbf{x}_{ijk} | \mathbf{y}_{ijk}\Big] \right\}^{-1} \\
&\quad \left\{ \sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}} E\Big[\mathbf{x}_{ijk}^T \mathbf{W}(\boldsymbol{\beta}_0, \mathbf{U}) \frac{\partial \eta_{ijk}}{\partial \mu_{ijk}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\mathbf{Y}_{ijk} - \boldsymbol{\mu}_{ijk}(\boldsymbol{\beta}_0, \mathbf{U})) | \mathbf{y}_{ijk}\Big] \right\} \\
&= \boldsymbol{\beta}^{(m)} + \left\{ E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}} \mathbf{x}_{ijk}^T \mathbf{W}(\boldsymbol{\beta}_0, \mathbf{U}) \mathbf{x}_{ijk} | \mathbf{y}_{ijk}\Big] \right\}^{-1} \\
&\quad \left\{ E\Big[\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}} \mathbf{x}_{ijk}^T \mathbf{W}(\boldsymbol{\beta}_0, \mathbf{U}) \frac{\partial \eta_{ijk}}{\partial \mu_{ijk}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\mathbf{Y}_{ijk} - \boldsymbol{\mu}_{ijk}(\boldsymbol{\beta}_0, \mathbf{U})) | \mathbf{y}_{ijk}\Big] \right\},
\end{aligned}$$

where the components of this expression take the following forms based on the conditional distribution of $y_{ijkl}|\mathbf{u}$ as defined above in the logistic mixed model:

$$\mu_{ijkl}(\boldsymbol{\beta}, \mathbf{u}) = E[Y_{ijkl}|\mathbf{u}] = p_{ijkl} = \frac{e^{\eta_{ijkl}}}{1 + e^{\eta_{ijkl}}}. \tag{37}$$

For the matrix $\mathbf{W}(\boldsymbol{\beta_0}, \mathbf{U})$, the following expression is obtained:

$$\mathbf{W}(\boldsymbol{\beta}_0, \mathbf{U})^{-1} = \text{diag} \left\{ \left(\frac{\partial \eta_{ijkl}}{\partial \mu_{ijkl}} \right)^2 \right\} \text{var}\,(Y_{ijkl}|\mathbf{u})$$

where we may use

$$\eta_{ijkl} = \log \left(\frac{\mu_{ijkl}}{1 - \mu_{ijkl}} \right)$$

to get an expression for the partial derivative term

$$\left(\frac{\partial \eta_{ijkl}}{\partial \mu_{ijkl}}\right) = \frac{1-\mu_{ijkl}}{\mu_{ijkl}}\left[\frac{(1-\mu_{ijkl})+\mu_{ijkl}}{(1-\mu_{ijkl})^2}\right] = \frac{1}{\mu_{ijkl}(1-\mu_{ijkl})}. \tag{38}$$

And lastly, since the distribution of $Y_{ijkl}|\mathbf{u}$ is assumed to be Bernoulli(p_{ijkl}), the conditional variance is

$$\text{var}(Y_{ijkl}|\mathbf{u}) = p_{ijkl}(1-p_{ijkl}) = \mu_{ijkl}(1-\mu_{ijkl}). \tag{39}$$

Solving ML estimating equation (7) will give the estimates for the variance components. This is done in step 2(b) of the MCNR algorithm by maximizing $\log f_u(\mathbf{u}^{(k)}|\mathbf{D})$ with respect to $\mathbf{D}$. An update for each variance component in the iterative method is given by

$$\sigma^{2(m+1)}_{county} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{n_{county}}\mathbf{a}^{(k)T}\mathbf{a}^{(\mathbf{k})},$$
$$\sigma^{2(m+1)}_{school} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{n_{school}}\mathbf{b}^{(k)T}\mathbf{b}^{(\mathbf{k})},$$
$$\sigma^{2(m+1)}_{class} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{n_{class}}\mathbf{c}^{(k)T}\mathbf{c}^{(\mathbf{k})},$$

where n_{county}, n_{school} and n_{class} represent the total number of counties, schools and classes, respectively. The vectors $\mathbf{a}^{(k)}$, $\mathbf{b}^{(k)}$ and $\mathbf{c}^{(k)}$ contain the simulated random effects obtained from the Metropolis algorithm.

Now set $\mathbf{u} = (\mathbf{a}^T, \mathbf{b}^T, \mathbf{c}^T)^T$ and let the candidate distribution, $h_u(\mathbf{u})$, be $f_u(\mathbf{u}) \sim N(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is the diagonal matrix with components σ^2_{county}, σ^2_{school} and σ^2_{class}. Then the acceptance function for the Metropolis algorithm for the specified model is

given by

$$A_{(ijk)}(\mathbf{u},\mathbf{u}^*) = \min\Big[1, \exp\{y_{(ijk)+}(a_i + b_{ij} + c_{ijk} - a_i^* - b_{ij}^* - c_{ijk}^*)\} \prod_{l=1}^{n_{ijk}} \left(\frac{1 + e^{x_{ijkl}^T\boldsymbol{\beta} + a_i + b_{ij} + c_{ijk}}}{1 + e^{x_{ijkl}^T\boldsymbol{\beta} + a_i^* + b_{ij}^* + c_{ijk}^*}}\right)\Big].$$

4.2.3 Computation of Test Statistics and Results

A preliminary analysis of the data found that race was statistically insignificant based on the unconstrained Z-test. The variable was therefore dropped from the model while the gender and age group fixed effects were kept. The results for the unconstrained and constrained variance component estimates are given in Table 2 along with the unconstrained fixed effects parameters based on this new model. Monte Carlo Newton-Raphson used 50, 200, 1000, and 5000 Monte Carlo replications for Newton-Raphson iterations 1-20, 21-40, 41-45, and 46-48 before satisfactory convergence. The equality and inequality constrained estimates were then calculated based on the GP algorithm as outlined in Section 3.2.2.

With respect to the regression parameters, age seems to be the most significant factor. Estimates for $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_3$ associated with the age covariates are each positive and increasing which means that the odds ratios are all greater than one. For instance, the odds that a student in the 17+ age group has ever smoked is just over 14 times the odds that a child aged 9-12 has smoked, given that they belong to the same class, ie., that they have identical random effects. The variance components were not found to be linearly ordered in the way that was originally hypothesized, but a more thorough assessment can be made using the test statistics proposed in Chapter 3.

Tests of H_0 against $H_1 - H_0$ and H_1 against $H_2 - H_1$ were conducted with the hypotheses specified as

Table 2: 2011 NYTS Parameter Estimates

	Unconstrained		**Constrained**	
Covariate	Estimate	Standard Error	Under H_0	Under H_1
σ^2_{county}	0.0390	0.0147	0.0830	0.0388
σ^2_{school}	0.2980	0.0715	0.0830	0.2816
σ^2_{class}	0.2770	0.0360	0.0830	0.2816
Intercept	-2.2498	0.2428	-2.3600	-2.2578
Age2	0.5708	0.2551	0.6606	0.5771
Age3	1.8053	0.2458	1.9065	1.8149
Age4	2.6487	0.2790	2.6221	2.6680
Gender2	-0.1570	0.0787	-0.1435	-0.1561

Note: All covariates are significant at $\alpha = 0.05$ based on the usual unconstrained asymptotic Z-test.

$$H_0 : \sigma^2_{county} = \sigma^2_{school} = \sigma^2_{class},$$
$$H_1 : \sigma^2_{county} \leq \sigma^2_{school} \leq \sigma^2_{class},$$
$$H_2 : \sigma^2\text{'s unconstrained.}$$

Both tests involve a total of 2 inequality constraints and so k from Theorem 3.1 is set to 3. The chi-bar-squared weights $w_i(2, \mathbf{R}(\mathcal{I}^{\mathbf{DD}})^{-1}\mathbf{R}^T, \mathbb{R}^{+2})$, as given by Theorem 3.1, were found using the solutions at the end of Section 2.2.4. The results obtained for the test statistics and their associated p-values are shown in Table 3. For a test of H_0 against $H_1 - H_0$, the p-value was found to be highly significant. The null hypothesis is rejected in favour of the alternative which states that the variance components are linearly ordered. For the test of H_1 against $H_2 - H_1$, the p-value was 0.6409 indicating that there is not sufficient evidence to reject the idea that the variance component for the class effect is at least as large as that of the school effect, which is at least as large as the variance component associated with the county effect. This implies that with respect to the incidence of youth smoking, the level of dependency amongst

students in the same class is at least as strong as the dependency amongst students in the same school, which is stronger than the dependency amongst students in the same county based on this order restricted test.

Table 3: Test Statistics

Test Statistic	Estimate	P-Value
t_{01}	49.6272	< 0.0001
t_{12}	0.0647	0.6409

Chapter 5

Discussion

Order restricted tests have been well-studied in statistical inference when mean parameters are involved, and rightfully so. Mean values are easy to interpret and assumptions about the ordering of a multivariate mean parameter are easier to justify. The same might not be true when it comes to the subject of variance components for a set of random effects, but there are certainly cases where it is still possible to incorporate partial information with the goal of developing a more efficient testing method. In some cases, that extra information by way of a parameter constraint is naturally imposed. Take for example the test for the presence of a single random effect in a GLMM. The parameter space for the variance component is actually a closed convex cone since it must take a nonnegative value. Problems of this type are referred to as boundary value tests and have been proven by Chernoff (1954) and more recently but slightly more generally by Self and Liang (1987) to have asymptotically chi-bar-squared null distributions as well.

To illustrate a case of a multi-parameter, order restricted test on variance components, data from the National Youth Tobacco Survey was used. It should be stressed that a specific order was imposed on the random effects parameters due to the reasonable assumption that greater social interaction amongst students should lead to greater dependency in smoking response, and that the level of interaction is inversely

proportional to cluster size. The point to be made is that an ordering of the random effects cannot be based on cluster size alone. There must be a mechanism that can explain the dependency.

The intent of this thesis was to study order restricted tests on the variance components within a GLMM where an ordering might be imposed based on the level of dependency in the data associated with each random effect. Maximum likelihood estimation was done via MCNR to more appropriately deal with both the intractable integral as well as a binary response variable. In proposing the distance statistics, T_{01} and T_{12}, as our test statistics for inequality constrained problems involving random components, two large sample approximations were made. The first utilizes the result that the maximum likelihood estimator for any parameter, including a set of variance components, is asymptotically multivariate normal under certain regularity conditions. And the second approximates the log-likelihood function with a quadratic Taylor series expansion whose remainder term converges in probability to 0, given those same regularity conditions. With these two large sample results, we are able to show that the asymptotic null distributions for both the distance statistic and the LRT are equivalent and chi-bar-squared. Moreover, the proposed test statistics do not require the calculation of a constrained MLE, giving it a significant computational advantage over the LRT.

With that said, not much else is known regarding the advantages or disadvantages of one test statistic over the other. The score test can also be shown to have the same limiting distribution as these two tests. A thorough study of each of their performances in a side by side setting might be worthy of future study. Finally, power comparisons for the test statistics, T_{01} and T_{12}, using small to moderate sample sizes would also be a good idea. With two asymptotic approximations justifying its use, the proposed test is most likely best suited for large samples. Understanding how well it performs in different situations would be logical next step.

Chapter 6

Appendix

Monte Carlo Newton-Raphson Estimation. This portion only contains the estimation section of the program. The metropolis sampling section has been left out, as well as the data formatting steps.

```
proc.time()-ptm.ma
i<-NULL
j<-NULL
k<-NULL
#begin newton raphson
#####################
#########################################################
#set up random effects vectors for monte carlo loop
#sample random effects from the metropolis distribution
N<-3000
a.samp<-a.vec[,(mc.size-N+1):mc.size]
b.samp<-b.vec[,(mc.size-N+1):mc.size]
c.samp<-c.vec[,(mc.size-N+1):mc.size]
#add i,j,k identifiers to the random effects vectors
a<-cbind(n.i,a.samp)[c(-2)]
```

```
b<-cbind(n.ij,b.samp)[c(-3)]
c<-cbind(n.ijk,c.samp)[c(-4)]
##########################################################
score.alpha <- 0
fisher.alpha <- 0
score.beta <- 0
fisher.beta <- 0
F<-0 #fisher
S<-0    #score
F.sigma<-0
S.sigma<-0
SS<-0
S.A<-0 #variance component a
S.B<-0 #vc b
S.C<-0 #vc c

#monte carlo estimation loop
for (s in 1:N)
{
  F0<-0
  S0<-0
  for (it in 1:n.c.)
  {
  ni<-subset(n.i$ni,n.i$i==it)
    for (jt in 1:ni)
    {
      nij<-subset(n.ij$nij,n.ij$i==it & n.ij$j==jt)
```

```
        for (kt in 1:nij)
        {
          x.ijk<-as.matrix(subset(data,(data$county==it & data$school==jt &
data$class==kt))[c(2:10)])
          y.ijk<-subset(data,(data$county==it & data$school==jt &
data$class==kt))$ever_smoked
          a.i<-unlist(subset(a,a$i==it)[c(-1)])
          b.j<-unlist(subset(b,(b$i==it & b$j==jt))[c(-1,-2)])
          c.k<-unlist(subset(c,(c$i==it & c$j==jt & c$k==kt))[c(-1,-2,-3)])
          eta.ijk<-x.ijk%*%beta0 + a.i[s] + b.j[s] + c.k[s]
#(nxp)*(px1)=nx1 vector

          mu.ijk<-exp(eta.ijk)/(1+exp(eta.ijk)) #nx1 vector

          #create the w matrix
          if (length(mu.ijk)>1)
          {
            w<-diag(c(mu.ijk*(1-mu.ijk)))
          }
          else
          {
            w<-mu.ijk*(1-mu.ijk)
          }
          F00<-t(x.ijk)%*%w%*%x.ijk
          S00<-t(x.ijk)%*%(y.ijk-mu.ijk)
          F0<-F0+F00 #cluster summation
          S0<-S0+S00
```

```
        }
      }
    }
    #calculating score for one MC draw
    one.mc.var.a<-t(a[,s+1])%*%a[,s+1]/n.c.     #var est. for one MC draw
    one.mc.var.b<-t(b[,s+2])%*%b[,s+2]/n.c.s.
    one.mc.var.c<-t(c[,s+3])%*%c[,s+3]/n.c.s.c.
    one.mc.score.a<-(t(a[,s+1])%*%a[,s+1])/(2*sigma0[1]^2)-
n.c./(2*sigma0[1])
    one.mc.score.b<-(t(b[,s+2])%*%b[,s+2])/(2*sigma0[2]^2)-
n.c.s./(2*sigma0[2])
    one.mc.score.c<-(t(c[,s+3])%*%c[,s+3])/(2*sigma0[3]^2)-
n.c.s.c./(2*sigma0[3])
    S0.sigma<-c(one.mc.score.a,one.mc.score.b,one.mc.score.c)
    one.mc.fisher.a<-(t(a[,s+1])%*%a[,s+1])/(sigma0[1]^3)-
n.c./(2*sigma0[1]^2)
    one.mc.fisher.b<-(t(b[,s+2])%*%b[,s+2])/(sigma0[2]^3)-
n.c.s./(2*sigma0[2]^2)
    one.mc.fisher.c<-(t(c[,s+3])%*%c[,s+3])/(sigma0[3]^3)-
n.c.s.c./(2*sigma0[3]^2)
    F0.sigma<-c(one.mc.fisher.a,one.mc.fisher.b,one.mc.fisher.c)
    ################################################
    #calculate the second term in Louis' observed Fisher
    one.mc.beta.beta<-S0%*%t(S0)     #beta-beta block
    one.mc.beta.sigma<-S0%*%t(S0.sigma)     #beta-sigma block
    one.mc.sigma.beta<-S0.sigma%*%t(S0)     #sigma-beta block
    one.mc.sigma.sigma<-S0.sigma%*%t(S0.sigma)     #sigma-sigma block
```

```
    #second term in observed Fisher
    score.score<-rbind(cbind(one.mc.beta.beta,one.mc.beta.sigma),
cbind(one.mc.sigma.beta,one.mc.sigma.sigma))
    ########################################################
    ########################################################
    #monte carlo summation
    F<-F+F0   #Fisher for regression parameters
    S<-S+S0   #score for regression parameters
    F.sigma<-F.sigma+F0.sigma    #Fisher for variance components
    S.sigma<-S.sigma+S0.sigma    #score for sigma
    SS<-SS+score.score     #second term in Louis' formula
    S.A<-S.A+one.mc.var.a #variance component for a
    S.B<-S.B+one.mc.var.b     #variance component for b
    S.C<-S.C+one.mc.var.c     #variance component for c
    ########################################################
    cat("mc estimation loop:s loop, ", s, "\n")
  }#end of monte carlo estimation loop
  mc.fisher<-F/N
  mc.score<-S/N
  mc.fisher.d<-diag(F.sigma/N)
  mc.score.d<-S.sigma/N
  mc.score.score<-SS/N
  mc.sigma.a<-S.A/N
  mc.sigma.b<-S.B/N
  mc.sigma.c<-S.C/N
  mc.sigma.abc<-c(mc.sigma.a,mc.sigma.b,mc.sigma.c)
  #updata parameters for next iteration
```

```
  beta0<-beta0+solve(mc.fisher)%*%mc.score
  beta<-data.frame(cbind(beta,beta0))
  sigma<-data.frame(cbind(sigma,mc.sigma.abc))
  cat("entire loop:iter loop, ", nr, "\n")
} #end of iter loop
proc.time()-ptm
```

List of References

[1] Andrews, D.W.K. (1999), Estimation when a parameter is on a boundary, *Econometrica*, 67, 1341–1383.

[2] Bartholomew, D.J. (1959), A test of homogeneity for ordered alternatives, *Biometrika*, 46, 36–48.

[3] Bartholomew, D.J. (1959), A test of homogeneity for ordered alternatives II, *Biometrika*, 46, 328–335.

[4] Breslow, N.E. and Clayton, D.G. (1993), Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, 88, 9–25.

[5] Breslow, N.E. and Lin, X. (1995), Bias correction in generalised linear mixed models with a single component of dispersion, *Biometrika*, 82, 81–91.

[6] Casella, G. and Berger, R.L. (1990), *Statistical Inference*, Duxbury Press Belmont, CA.

[7] Chernoff, H. (1954), On the distribution of the likelihood ratio, *The Annals of Mathematical Statistics*, 25, 573–578.

[8] Cramér, H., (1946) *Methods of Mathematical Statistics*, Princeton University Press.

[9] Davis, K.A. (2011), *Constrained Statistical Inference in Generalized Linear, and Mixed Models with Incomplete Data*, Carleton University.

[10] Davis, P.J. and Rabinowitz, P. (1975), *Methods of Numerical Integration*, Academic Press New York.

[11] Dykstra, R., Kochar, S. and Robertson, T. (1991), Statistical inference for uniform stochastic ordering in several populations, *The Annals of Statistics*, 19, 870–888.

[12] Efron, B. and Hinkley, D.V. (1978), Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, *Biometrika*, 65, 457–483.

[13] Farebrother, R.W. (1986), Testing linear inequality constraints in the standard linear model, *Communications in Statistics-Theory and Methods*, 15, 07–31.

[14] Fletcher, R. (2013), *Practical methods of optimization*, John Wiley & Sons.

[15] Gourieroux, C., Holly, A. and Monfort, A. (1982), Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters, *Econometrica: Journal of the Econometric Society*, 50, 63–80.

[16] Green, P.J. (1987), Penalized likelihood for general semi-parametric regression models, *International Statistical Review/Revue Internationale de Statistique*, 55 245–259.

[17] Hall, D.B. and Præstgaard, J.T. (2001), Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models, *Biometrika*, 88, 739–751.

[18] Jamshidian, M. (2004), On algorithms for restricted maximum likelihood estimation, *Computational Statistics & Data Analysis*, 45, 137–157.

[19] Kudo, A. (1963), A multivariate analogue of the one-sided test, *Biometrika*, 50, 403–418.

[20] Kuk, A.Y.C. and Cheng, Y.W. (1997), The monte carlo newton-raphson algorithm, *Journal of Statistical Computation and Simulation*, 59, 233–250.

[21] Lin, X. (1997), Variance component testing in generalised linear models with random effects, *Biometrika*, 84, 309–326.

[22] Lin, X. and Breslow, N.E. (1996), Bias correction in generalized linear mixed models with multiple components of dispersion, *Journal of the American Statistical Association*, 91, 1007–1016.

[23] Louis, T.A. (1982), Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, 44, 226–233.

[24] McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models, 2nd Ed.*, Chapman and Hall.

[25] McCulloch, C.E., (1997), Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association*, 92, 162–170.

[26] McCulloch, C.E. and Searle, S.R. (2001), *Generalized, Linear, and Mixed Models*, John Wiley and Sons.

[27] McLachlan, G. and Krishnan, T. (2008), *The EM Algorithm and Extensions, 2nd Ed.*, John Wiley and Sons.

[28] Neath, R.C. (2006), *Monte Carlo Methods for Likelihood-based Inference in Hierarchical Models*, ProQuest.

[29] Perlman, M.D. (1969), One-sided testing problems in multivariate analysis, *The Annals of Mathematical Statistics*, 40, 549–567.

[30] Robert, C.P. and Casella, G. (2010), *Introducing Monte Carlo Methods with R*, Springer.

[31] Robertson, T. and Wegman, E.J. (1978), Likelihood ratio tests for order restrictions in exponential families, *The Annals of Statistics*, 6, 485–505.

[32] Robertson, T., Wright, F.T. and Dykstra, R.L. (1988), *Order Restricted Statistical Inference*, Wiley New York.

[33] Self, S.G. and Liang, K.Y. (1987), Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association*, 82, 605–610.

[34] Shapiro, A. (1985), Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints, *Biometrika*, 72, 133–144.

[35] Shapiro, A. (1988), Towards a unified theory of inequality constrained testing in multivariate analysis, *International Statistical Review/Revue Internationale de Statistique*, 56, 49–62.

[36] Silvapulle, M.J. (1994), On tests against one-sided hypotheses in some generalized linear models, *Biometrics*, 50, 853–858.

[37] Silvapulle, M.J. and Sen, P.K. (2005), *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*, John Wiley and Sons.

[38] Sinha, S.K. (2004), Robust analysis of generalized linear mixed models, *Journal of the American Statistical Association*, 99, 451–460.

[39] Wald, A. (1949), Note on the consistency of the maximum likelihood estimate, *The Annals of Mathematical Statistics*, 20, 595–601.

[40] Wand, M.P. (2007), Fisher information for generalised linear mixed models, *Journal of Multivariate Analysis*, 98, 1412–1416.

[41] Zhang, H. (2002), On estimation and prediction for spatial generalized linear mixed models, *Biometrics*, 58, 129–136.

[42] Zhang, D. and Lin, X. (2008), Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics, *Random Effect and Latent Variable Model Selection*, 192, 19–36.