Joint Modeling of Genetic Linkage and Association

by

©Haiyan Yang

A thesis submitted to the School of

Graduate Studies in partial fulfilment

of the requirements for degree of

Doctor of Philosophy

Department of Mathematics and Statistics

Memorial University of Newfoundland

May 2014

St. John's, Newfoundland and Labrador

Abstract

Understanding the complexities involved in identifying disease causing genes is still a monumental task. As we know, genetic variants and environmental factors can influence the risk of disease outcomes. Epidemiological studies have identified that age is one of a number of environmental risk factors for Familial Pulmonary Fibrosis (FPF), but the genetic risk factors involved identification of disease causing genes still are a problem largely unsolved. An inherited disease-causing locus occurs in the same genomic position as an ancestor who has the disease trait, and the disease genotype may be associated with a marker genotype. A joint modeling of genetic linkage and association within families having a remote common ancestor or at population level is presented in this thesis. This joint modeling uses a likelihood approach that allows the inclusion of other covariates into the model for quantitative traits and binary traits with multivariate random effects. Power studies via simulation compare the new proposed procedure with standard linkage or association procedures. The joint test is more powerful than linkage or association test alone where both sources of variation of linkage or association are present. Furthermore, the proposed method also allows testing against specific alternatives - for example, against the significance of linkage where there is no association, significance of association where there is no linkage, and significance of both linkage and association. By utilizing data from five FPF families in Newfoundland, four candidate loci were identified for the linkage or/and association with age-at-onset gene and FPF (rs4605929 in chromosome 6, rs11078200 in chromosome 7, rs1941686 in chromosome 18 and rs114682 in chromosome 22).

Acknowledgement

I would like to express my sincere gratitude to my supervisors, J C. Loredo-Osti and Michael Woods, for granting me the privilege of writing this thesis under their supervision and for their support, financial and otherwise, throughout my program. They have been very generous with their ideas and time, and seemingly have an infinite amount of patience and devotion. They patiently read through my work several times and offered their very valuable suggestions. This work would have been impossible without their guidance and help.

Many thanks go to my wife, Dr. Yuan Yuan, for her love and support, and for making available to me high performance computing facilities and applied mathematical knowledge.

I would like to thank Dr. Zhaozhi Fan, who is not listed as a supervisor but has been most helpful throughout: he has opened my eyes to new methods and opportunities to make a meaningful contribution to the field. I would like to thank all the professors who have shared their wealth of statistical knowledge at various points in the course of my program with me. A big thanks goes to Dr. Terry Young and Dr. Michael Woods for giving me the work opportunity on such interesting topics and exposing me to such influential minds in this field. I thank Dr. Michael Woods' lab members for the data used in my thesis. I would also like to thank the staff of the Mathematics and Statistics for providing the friendly atmosphere and necessary facilities to finish my program.

My sincere thanks to my families, friends, and classmates whose assistance and encouragement have seen me to the completion of my program.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Due to the immense academic and commercial effort in mapping the total human genome, it has become feasible to conduct large genome-wide linkage or association studies for complex behavior and disease, using measured genes or genetic markers (micro-satellites and/or single-nucleotide polymorphisms(SNPs)). Genetic mapping procedures are used to locate and identify the gene or genetic markers associated or linked to a particular inherited trait. Genetic mapping approaches - such as linkage analysis, association studies, or joint linkage and association studies - enable researchers to sample a large pool of genetic markers from each subject in a genome-wide manner, capture variation uniformly across an individual's genome. Such variation is used to explore how the genes and alleles contribute to susceptibility to a particular disease as well as the way they interact with each other as well as with environmental and other stochastic factors to produce phenotypes. The aim of this thesis is to explore and develop statistical approaches based on the joint modeling of linkage

and association for mapping and/or identification of genes of complex diseases. The methods proposed here will be applied to study the genetics underlying the inheritance of familial pulmonary fibrosis (FPF).

1.1 Linkage Analysis

Genetic linkage is the tendency of genetic loci that are located proximal to each other to be inherited together during meiosis. Loci within a small chromosome neighborhood are less likely to be separated onto different chromatids during crossover, and are therefore said to be genetically linked. The main idea of linkage studies is that the loci which are found in a vicinity on the chromosome have a tendency to stick together when passed on to offsprings. The goal of linkage analysis is to "infer relative position of two or more loci by examining transmission from parent to offspring or allele sharing patterns of relatives" (Sham 1998). Linkage analysis is used to infer locations on chromosomes where disease genes lie with respect to a set of genetic markers. In linkage studies, the relatives, who have similar phenotypes, will likely have identical alleles at the genetic markers only if the disease gene controlling that phenotype is linked to these markers. Therefore, it is of interest to find which markers are tightly linked to the transmission patterns of a putative disease gene.

In human genetics, much progress in statistical and computational methodology has been achieved for linkage analysis. The key approaches to human linkage analysis include the segregation (or co-segregation) analysis (Morton, 1955; Kruglyak et al.,

1996), regression methods (Haseman and Elston, 1972; Allison et al., 2000), variance components (Amos, 1994; Fulker and Cherny, 1996; Diao and Lin, 2010).

At the most basic level, linkage analysis tests for co-segregation between the location of a putative gene for a given trait and a genetic marker. The major groups of linkage statistics are classified as "model-based" (also termed "parametric") and "model-free" (also termed "nonparametric"). Model-based linkage requires specification of the model of inheritance (additive, dominant, or recessive). It is a three-step procedure: (i) genotype the collection of markers along the genome; (ii) calculate the appropriate linkage statistic between the putative locus and each marker or group of markers; (iii) identity the regions where the statistic analysis shows "significant" evidence of linkage.

An early approach to human linkage analysis was segregation analysis that used the "log of the odds ratio" (LOD) as the test statistic (Morton, 1955). Under this approach, the LOD score is calculated on a grid of locations for the putative gene determined by the markers and used to determine where the strongest evidence of co-segregation between markers and the putative gene associated to the phenotype come from. If the likelihood was maximized over a single recombination fraction alone, a LOD score of 3 was taken as significant. Lander and Schork (1994) synthesized the linkage methods to highlight some enlightening examples of the genetic dissection of complex traits. Kruglyak et al. (1996) performed multipoint parametric LOD-score calculations that use all available inheritance information about segregation at every point in the genome from general pedigrees of moderate size, based on genotypes at large number of mark-

ers considered simultaneously. Whittemore (1996) developed a score test for linkage analysis that differentiates the retrospective log likelihood of marker data when phenotypes are given with respect to model parameters. McPeek (1999) described the direct connection between the affected relative methods and traditional parametric linkage analysis for dominant, additive, and recessive models, and used this connection to produce explicit formulae for the optimal sharing statistics and weights that are applicable to all pedigree types.

Linkage analysis is motivated by the phenomenon of recombination. If we went to examine two loci which are close together, we would expect the number of recombinations between them to be close to 0. On the other hand, if we went to examine two loci which are on different chromosomes or far apart on a chromosome, then we would expect that half or nearly half of them recombine. So, testing for linkage between two loci is done by estimating if the recombination fraction differs significantly from 1/2.

Two-point linkage test is a test of linkage between two loci. The common method is to test if the recombination fraction between two loci is less than 0.5. If the number of recombinant individuals is k, then the probability of getting k recombinants is r^k. Likewise, the probability of getting $n-k$ nonrecombinants (n is the total number of people examined) is $(1-r)^{n-k}$. Remember that the recombination frequency between two unlinked markers is always 0.50. The probability of getting n individuals with any genotype is just 0.50^n. Therefore, the general formula for the LOD score is defined as

$$LOD = log_{10}\frac{L(r)}{L(r=0.5)} = log_{10}\frac{r^k(1-r)^{n-k}}{0.50^n} \tag{1.1}$$

A LOD score of 3 has been used as the threshold for linkage testing.

Therefore, for each marker, a LOD score is calculated to test the probability that the genetic marker and trait co-segregate. If the likelihood was maximized over a single recombination fraction alone, a LOD score of 3 was taken as significant.

Multipoint linkage test is commonly used to evaluate linkage of a disease to a small region by using multiple markers (Kruglyak et al. 1996; Goring and Terwilliger, 2000; Kong et al., 2004). Linkage analysis can be more efficient if data for more than two markers are analyzed simultaneously. Experimental geneticists have long used three-point crosses for linkage analysis. Suppose that data are available for three linked loci A, D, and B in the same families, and denote the three recombination rates as r_{AD}, r_{AB}, and r_{DB}. The classical approach consists of analyzing each pairwise combination of loci by computing a LOD-score, and takes the estimation of recombination fraction value for which the lod-score is maximum. The gene order may be inferred by inspection of the estimated recombination rates. If the given order is ADB, then $r_{AB} = 1 - (1 - r_{AD})(1 - r_{DB})$

Multipoint linkage test is more efficient than estimating the recombination fraction for intervals in a series of two-point crosses. A second advantage of multipoint linkage test in humans is that it helps overcome problems caused by the limited informativeness of markers. Some meioses in a family might be informative with marker A, and others uninformative for A but informative with the nearby marker B. Simultaneous linkage analysis of the disease with markers A and B extracts the full information.

Model-free approaches have been developed to account for between-family and

within-family variation without the specification of a genetic model. The power of model-free methods is based on knowing or estimating the proportion of sharing of marker alleles that are identical by descent (IBD); that is, alleles are direct copies of the same ancestral alleles. The usefulness of sib pairs for quantitative trait loci linkage analysis (QTL) is well established and is based on the use of IBD relationships among genotypes (Haseman and Elston 1972). Hössjer (2003) proposed a score test that is conditional on observed phenotypes within a unified framework, investigated the asymptotic behavior of disease locus estimators under perfect marker information, corresponding to a dense set of markers when all (or a sufficient number of) pedigree members are being typed. Later, Hössjer (2005a) developed a general strategy for linkage analysis which is applicable to arbitrary pedigree structures and genetic models, with major gene and environmental effects that require disease allele frequencies and penetrance parameters of the causal gene. Hössjer's score tests make generalized linear models for linkage analyses. Lemire (2005) proposed some nonparametric methods based on the estimation of inheritance vectors to test for linkage.

Another type of linkage analysis is referred to as the "variance components" (VC) methods. VC methods in genetic studies are a powerful tool for modeling continuous response variable in families (Lange et al.,1976; Hopper and Mathews, 1982; Goldgar, 1990; Schork, 1993; Amos, 1994; Amos et al., 1996; Falconer and Mackay, 1996; Kruglyak et al., 1996; Blangero and Almasy, 1997; Williams et al., 1997; Abecasis et al., 2002; Sullivan et al., 2003; Evans and Medland, 2003; Diao and Lin, 2010).

VC methods offer a powerful and flexible approach to model-free linkage analysis.

As in any linkage procedure that the purpose is to determine whether genetic variation at a specific marker locus can explain the variation in the phenotype and to estimate the location of a putative QTL. When a locus is associated to a trait, variation at its position increases the variance as well as induces correlation among relatives who share the same alleles by descent. VC linkage attempts to estimate proportion of these variance by exploring the relationship among relatives. It can be used to test the significance of a QTL effect through the use of a likelihood ratio test. The variance of the phenotype can be broken down into components due to genes of large effect linked to few marker locations and variation. The modeling under VC linkage is quite simple, instead of specifying the allele frequencies and penetrances for a trait locus. VC method examines the phenotypes co-variation of related individuals given the relationship between the individuals and the proportion of IBD genes shared at a specific marker locus.

Many authors considered a phenotype of interest, measured in a set of pedigrees, each including one or more related individuals. Denote Y_{ij} and $\mathbf{x}$ as the observed trait and covariates, respectively, for individual j in family i; G_{ijm} as the observed genotype at marker m for individual j in family i. For each of the genotyped SNP markers, researchers are interested in using VC model to testing whether the marker locus and the disease locus are linked. For the SNP being tested, label the two alleles "A" and "a", and define a genotype score, g_{ijm} , as 0, 1, or 2, depending on whether G_{ijm} is aa, Aa, or AA, respectively.

In a basic VC model, it is assume that the vector of phenotype in the jth family, $\mathbf{Y}$, has a multivariate normal distribution with mean $E(\mathbf{y}) = \boldsymbol{\mu} + \alpha\mathbf{g} + \boldsymbol{\gamma}\mathbf{x}$ and covariance

matrix:

$$\mathbf{\Sigma} = \sigma_a^2 \mathbf{\Sigma}_a + 2\sigma_g^2 \mathbf{\Sigma}_g + \sigma_e^2 \mathbf{I}$$

where $\boldsymbol{\mu}$ is the population mean, **g** is the genotype score, and **x** is the covariate of interest; σ_a^2, σ_g^2 and σ_e^2 are the polygenic, major gene and residuals variance components, respectively; $\mathbf{\Sigma}_a$ is a matrix that depends on the IBD status at the tested locus; $\mathbf{\Sigma}_g$ is the kinship coefficient matrix; and $\mathbf{I}$ is identity matrix. Self and Liang (1987) tested the null hypothesis $H_0 : \sigma_a^2 = 0$ against the alternative hypothesis $H_1 : \sigma_a^2 > 0$ by likelihood ratio test which is approximately a half-and-half mixture of a χ_1^2 variable and a point mass at 0.

The modeling framework used in VC analysis is remarkably general. Lange et al. (1976) suggested the likelihood ratio for testing linkage based on the maximum likelihood estimates of the VC. Hopper and Mathews (1982) introduced a VC linkage analysis procedure to estimate the effect of measured genetic markers and the effect of shared family environments. Goldgar (1990) presented a linkage test based on estimating the proportion of genetic material shared IBD by sibling pairs in a specified chromosomal region. Schork (1993) proposed a similar procedure based on specifying the expected genetic covariances in arbitrary relatives as a function of the IBD relationships at a QTL. Amos (1994) developed a linkage method based on variance components to estimate the genetic variance attributable to the region around a specific genetic marker.

Linkage analysis has been extremely successful at identifying genetic variations for many diseases that underline single-gene disorders following Mendelian inheritance

patterns, including Huntington's disease (Gusella et al. 1983), Duchenne muscular dystrophy (Koenig et al. 1987), cystic fibrosis (Kerem et al., 1989; Riordan et al., 1989; Rommens et al., 1989) and neurofibromatosis type-1 (Xu et al., 1990). However, for some complex traits that influence by multiple genetic, environmental factors and interaction, linkage analysis is limited.

1.2 Association Studies

Association refers to a correlation between a particular marker allele and a disease trait. Association studies are useful for assessing potential candidate genes, either in targeted regions (Xie and Ott, 1993; Zhao et al., 2002; Zaykin et al., 2002; Sham et al., 2004; Van Steen and Lange, 2005; Curtis et al., 2006) or in genome wide analyses (Farrer et al., 1997; Klein, et al., 2005; Barrett and Cardon, 2006; Duerr et. al., 2006; Sladek, et al., 2007).

Linkage disequilibrium (LD) is the non-random association of alleles at two or more loci. In other words, LD is the occurrence of some combinations of alleles or genetic markers in a population more often or less often than expected from a random formation of haplotypes. LD is not the same as linkage, it is the association of two or more loci on a chromosome with limited recombination between them. If two populations with different allele frequencies are mixed then overall population can display disequilibrium, even if the loci are unlinked. Non random mating can induce disequilibrium in the absence of linkage as well. If both loci A and B are under directional

selection, then there will be a negative correlation between the alleles at these loci in progeny of selected parents, even when the loci are unlinked. In all these cases association between genotypes frequencies of unlinked loci may be evident.

Linkage disequilibrium is often quantified using statistics of association between the allelic states at pairs of loci. Chakraborty and Weiss (1988) referred a gametic association as "mixture disequilibrium" when two populations with different allele frequencies at two loci will produce a gametic association between these loci in any admixed population. Lander and Schork (1994) developed a non-random association test when a case-control sample is ethnically mixed or is derived from a population that experienced admixture during the past few generations at markers completely unlinked to a disease locus. Rabinowitz (1997) introduced family-based LD tests for quantitative traits by using parental genotypes to construct well-matched controls in simplex families.

Association mapping is a method to find a statistical association between genetic markers and a trait. Genetic markers may be in LD with the causative gene or lie within candidate genes suspected to contribute to the variation in the trait, the goal of association mapping is to identify the actual genes affecting that trait. Since population genetic structure (genetic differences that accumulate between isolated populations) can cause LD, association analyses must account for population genetic structure whenever it is present in the population from which the sample has been drawn (Pritchard et al., 2000; Thornsberry et al., 2001).

Association mapping is most often performed by scanning the entire genome for

significant associations between a panel of SNPs and a particular phenotype. These associations must then be independently verified in order to show that they either a) contribute to the trait of interest directly, or b) are linked or in linkage disequilibrium with a locus that contributes to the trait of interest. The advantage of association mapping is that it can map quantitative traits with high resolution in a way that is statistically very powerful. Association mapping is already widely used in candidate gene studies when trying to detect or localize the active variants at a fine scale. To date, genome wide associations studies (GWAS) have been performed on the human genome in attempt to identify SNPs associated with a wide variety of complex human diseases (e.g. cancer, Alzheimers disease, and obesity), and the generalized linear regression model (or some other statistical techniques depending upon the nature of phenotype) could be used to test whether the regression coefficients are significant (Allison, 1997; Tsai et al. 2001, 2003; Tikhonoff et al. 2003).

Association studies aim to identify genetic variants related to diseases by examining the associations between phenotypes and hundreds of thousands of markers. There are three popular study designs for association: random sampling from the population, case-control, and family-based association.

Many genetic studies of complex disorders are performed with samples from ethnically stratified populations. A case-control study is an analytical epidemiological research method that works to identify the factors that contribute to a particular disease or condition. Researchers select two groups of people from a common population: the ones with a particular disease (the cases) and the group without the disease (the

controls). Case-control studies compare allele frequencies between a group of unrelated, affected individuals and an unrelated group of matched controls (Owerbach et al., 1997; Bain et al., 2005; Zhao et al., 2006).

Case-control studies have specific advantages compared to other study designs. They are comparatively quick, inexpensive, and easy. They have been advocated to be particularly appropriate for (1) investigating outbreaks, and (2) studying rare diseases. Since case-control studies start with people known to have the outcome, it is prone to stratification and, consequently, it may lead to high number of false positive associations. On the other hand, it may make it possible to enroll a sufficient number of patients with a rare disease. As with any epidemiological study, greater numbers in the study will increase the power of the study. Case-control studies are a relatively inexpensive and frequently used type of epidemiological study that can be carried out by small teams or individual researchers in single facilities. The case-control study design is often used in the study of diseases where little is known about the association between the risk factor and disease of interest.

If genetic variants are more frequent in people with the disease, the variants are said to be "associated" with the disease. Association analysis of genetic polymorphisms has been mostly performed in those case-control settings where unrelated affected subjects are compared to unrelated, unaffected subjects. Significant differences in allele frequencies between cases and controls are taken as evidence for the involvement of an allele in disease susceptibility. Alternatively, genotype frequencies rather than allele frequencies can be compared in cases and controls. Self et al. (1991) extended case-

control studies to incorporate information from a matched control series to estimate disease and environmental risk factor effects simultaneously. Case-control association analyses are sensitive to population heterogeneity of disease etiology and marker allele frequencies (Curtis and Sham 1996; Deng 2001). Chapman et al. (2003) treated disease gene alleles as hidden variables and mainly focused on population based case-control studies, and the relation between marker and disease causing alleles was modeled in terms of linear regression. They propose an "indirect" method that presumes the existence of one or more causal variants in the region. In the absence of migration, mutation, natural selection, and assortative mating, genotype frequencies at any locus are a simple function of allele frequencies. A population is in equilibrium, termed "Hardy-Weinberg equilibrium" (HWE), if the gene and genotypic frequencies are constant from generation to generation. The Hardy-Weinberg Disequilibrium (HWD) at a marker locus in affected patients can be interpreted as evidence for association with a disease (Nielsen et al. 1998; Lee 2003). The analysis consists of either a Pearson χ^2, likelihood ratio test, Fisher's exact test, or logistic regression to test association in the case-control design.

A strong association between two variables does not necessarily imply a cause-effect relationship between them. (1) The association can be due to chance. Tests of statistical significance are important in determining the probability that the association is due to chance. (2) The association can be due to a bias such as non-comparable criteria, or non-comparable information. (3) The association can be due to a mixing of effects between the exposure, the disease, and a confounding factors. Thus, caution

should be used when interpreting results from case control studies.

One problem with the case-control design is that genotype and haplotype frequencies vary between ethnic or geographic populations. If the case and control populations are not well matched for ethnicity or geographic origin then false positive association can occur because of the confounding effects of population stratification.

Family based association designs aim to avoid the potential confounding effects of population stratification by using the parents as controls for the case, which is their affected offspring. The advantage of family-based studies has received much attention because spurious associations caused by population structure can be controlled, and marker genotype information on diseased cases and their parents can be used to test the compound hypothesis of both linkage and linkage disequilibrium.

Association mapping based on family studies can identify genes that influence complex human traits while providing protection against population stratification. Family-based studies test for equality between the transmission and nontransmission of a given allele to affected children from heterozygous parents (Terwilliger et al., 1992; Spielman et al., 1993; Boomsma et al., 2000; Nash et al., 2005; Gosso et al., 2006). The transmission/disequilibrium test (TDT, Spielman et al. 1993) considers parents who are heterozygous for an allele associated with disease and evaluates the frequency with which that allele or its alternate is transmitted to affected offspring. It does not require data either on multiple affected family members or on unaffected sibs. The TDT is a simple means of detecting associations that should only be positive if the marker allele is linked to the disease locus when the parents of affected subjects are available. Curtis

(1997) extended TDT to use unaffected siblings rather than parents as controls, and like the TDT, it is robust against bias due to population stratification and other sources; it is expected to produce only positive results when a marker is both associated and linked with the disease locus. Boehnke and Langefeld (1998) developed family-based tests of association that use sib pairs where one sib is affected with a disease and the other is not. These tests are based on statistics that compare counts of alleles or genotypes or for symmetry in tables of alleles or genotypes. Ideally, TDT tests should use parental genotypes when available, and sibling genotypes otherwise, to consider all available information in the most efficient manner possible. Diao and Lin (2006) have constructed a most flexible and powerful quantitative transmission-disequilibrium tests (QTDT) based on the variance-components model and family-based tests of association for quantitative traits.

Linear regression can be used to test for association between alleles and phenotypic outcomes. Abecasis et al. (2000) have built an identification of complex disease genes association methods for linkage-disequilibrium mapping of quantitative traits, to construct a general approach that can accommodate nuclear families of any size, with or without parental information by using variance components to construct a test that utilizes information from all available offspring. Laird and Lange (2006) treated the phenotype as the random response and the genotype as the fixed predictor and used the ordinary linear regression for association test. Baksh et al. (2007) presented an alternative likelihood-based method of analysis for ordered categorical phenotypes in nuclear families that permits straightforward inclusion of covariate, gene-gene, and

gene-covariate interaction terms in the likelihood, incorporating a simple model for ascertainment that allows for family-specific effects in the hypothesis test. Diao and Lin (2010) proposed a generalized linear VC model for the association analysis of ordinal traits.

1.3 Association in the Presence of Linkage

Association between complex traits and a series of closely linked SNPs is of central importance in modern human genetics. If association is due to LD between markers and causal loci, which in general acts over very short distances in the genome, this not only allows for fine mapping of disease susceptibility genes indicated by linkage studies, but it also offers an opportunity for discovering genes by association studies. The traditional route in gene discovering has linkage and association as two stages of the process. Once genetic linkage has been identified for a complex disease, the next step is an association analysis in which SNPs within the linkage region are genotyped. Genetic mapping studies reveal a region of linkage containing a number of associated variants. A marker may be genetically associated with the disease either because it has direct influence on disease susceptibility ("causal"), or because it is in linkage disequilibrium with a causal variants. Identifying the variant(s) that potentially 'explain' an observed linkage result is a routinely part of modern gene discovering methods. If a particular locus is the only causal variant in the region, then association with this locus should be able to explain all the linkage in the region. If the variant is not the

causal variant, or is not the only causal variant in the region, evidence of linkage should exist to explain the remaining variation. To localize the susceptibility allele more precisely, disease-marker association analysis with additional genetic markers specific to the linked region can be performed.

The TDT also fits to test for association in the presence of linkage when the marker locus and the hypothetical disease locus are linked and in linkage disequilibrium. Sham and Curtis (1995) derived the transmission probabilities for a multi-allele marker locus and a generalized single locus disease model that consists of two genotyped parents and an affected child in a random sample of affected families from a randomly mating population. The form of these transmission probabilities suggests an extension of the TDT to multi-allele marker loci, in which the alternative hypothesis is restricted to take account of the likely pattern of unequal transmission. However, family-based tests require information of parental marker genotypes, but for late-onset diseases parental data are often not available. Curtis (1997) proposed an alternative approach for analyzing larger sibships, resulting in a test of linkage and association by reducing each sibship to two siblings via two steps; first, randomly choose an affected individual, second, choose the unaffected sibling whose marker genotype is maximally different from that of the affected sibling in first step. Boehnke and Langefeld (1998) developed family-based tests of association of late-onset diseases that use discordant sib pairs in which one sib is affected with a disease and the other sib is not. Horvath and Laird (1998) introduced a discordant-sibship test that uses the data of all the affected and unaffected siblings. Monks et al. (1998) proposed an extension of family-based tests

of association and linkage, that utilizes unaffected siblings as surrogates for untyped parents in the context of a complex disease for both biallelic and multiallelic markers as well as for sibships of different sizes. Clayton (1999) proposed a score test for association in the presence of linkage in sibship data for situations in which transmission is uncertain, which means one or both parents are missing. Weinberg (1999) described a likelihood-based method for testing linkage disequilibrium for inclusion of genetic information from incomplete triads when one or both parents are missing. Rabinowitz and Laird (2000) proposed a family-based examination of linkage disequilibrium between marker alleles and traits, based on computing p-values. Martin et al. (2003) presented a test for association in the presence of linkage that incorporates IBD relationships to adjust for linkage when inferring missing parental genotypes in nuclear families. Lemire (2004) described a simple allele-sharing test statistic for discordant pairs (one affected and one unaffected individual) that share alleles less often than expected under Mendelian inheritance, and provide additional information about the segregation of the putative disease gene.

Statistical geneticists have devoted valuable thought to the problem of detecting association in the presence of linkage for quantitative traits. The available statistical procedures for the analysis of continuous traits have been proved to be very effective in sib-pair and related linkage procedures. Cardon et al. (2000) presented a systematic approach to the use of sib pairs for the analysis of both association and linkage for quantitative traits within the variance-components framework. Lake et al. (2000) performed an association test in the presence of linkage using the mean of the test statistic

and an empirical variance-covariance estimator that adjusts for the correlation among sibling marker genotypes. This provides a convenient means for testing allelic association in the presence of linkage that can be used with a wide range of test statistics and any pedigree configuration. Sun et al. (2002) showed that when the candidate SNPs is the sole causal site in the region, IBD sharing of affected sib pairs (ASPs) at the candidate SNP, is independent of their affected status and depends only on their genotypes at the SNP. Fan et al. (2003) investigated variance components models of both linkage analysis and high resolution LD mapping for QTL. The model simultaneously takes care of the linkage, LD or association, and the effects of the putative trait locus in the prior suggestive linkage region. Li et al. (2005) described a statistical framework that identifies candidate SNPs that can fully or partly explain the observed linkage signal based on joint modelling of linkage and association. Assuming one causal SNP in the region of linkage, they modelled the likelihood of the marker data conditional on the trait data for a sample of ASPs, with disease penetrances and disease-SNP haplotype frequencies as parameters, proposing likelihood-ratio tests to characterize the LD between the candidate and disease SNPs. Biernacka and Cordell (2007) tested a particular variant that can explain all of the observed linkage versus those that cannot.

Some authors have tested the linkage in the presence of association. Spielman and Ewens (1998) described a method, called the "sib TDT" (or "S-TDT"), that uses marker data from unaffected sibs instead of parents, thus allowing the application of the principle of TDT to sibships without parental data and allowing all the data to be used jointly in one overall TDT-type procedure to test for linkage in the presence of

association. Knapp (1999) proposed a method for testing a marker for linkage with a disease, which employs parental-genotype reconstruction information from affected and unaffected sibs. Zollner and Pritchard (2005) outlined a general coalescent framework that uses genotype data in linkage disequilibrium-based mapping studies to detect association and to estimate the location of the causative variation.

1.4 Joint Linkage and Association Analysis

Linkage and association methods are widely used in the genetic analysis in family studies, but the study of joint linkage and associated is not that common. It must be pointed out that joint modeling of linkage and association is not the same as performing a linkage study with a small set of markers followed by an association study with a denser set of markers on some "regions of interest" whenever "a linkage peak" is found, nor the other way around. Instead, in the study of joint linkage and association, we take into consideration both forms to develop the testing and estimation procedures to account for linkage and association simultaneously.

Linkage and association are different phenomena. Linkage describes the relationship phenotype/loci while association describes the relationship phenotype/alleles. Linkage is a consequence of co-segregation, a fundamental genetic principle, while association is simply a statistical statement about the co-occurrence of alleles. In contrast with association, linkage is a phenomenon to be studied within families, but not amongst unrelated people. Nonetheless, whenever two supposedly unrelated people

with disease D actually inherited their disease from a distant common ancestor, they may very well share particular ancestral alleles at loci closely linked to D. In so far as a population that can be seen as a large extended family, with families descending from a common ancestor, population level association due to linkage disequilibrium should exist between ancestral disease susceptibility genes and closely linked markers. In a situation like this, jointly model linkage and association methods are desirable which have greater efficiency than either method considered alone.

The methodological literature on genetic analysis of joint linkage and association analysis is very limited. Zhao et al. (1998) defined a semi-parametric estimating equations method, with one linkage and one association component in the score vector. Fulker et al. (1999) developed a method to test linkage while simultaneously modeling allelic association by using of the variance-components framework of means and variances for sib-pair data. Sham et al. (2000) introduced a joint likelihood ratio test for linkage and family-based association in which the association and linkage parameters are contained in mean vector and covariance matrix respectively. Hössjer (2005b) focused on family-based association studies and used the joint distribution of marker and disease alleles to introduce a combined score test for association and linkage analysis based on a biologically plausible model. His test is based on a retrospective likelihood of marker data given phenotypes, treating the alleles of the causal gene as hidden data. The score vector has one association and one linkage component, which can be used to define separate tests for association and linkage. Except for small pedigrees with very simple structures, the distribution of test statistic may be difficult to find. The

combined test is a robust alternative; it does not substantially under-perform relative to either linkage or association test, and sometimes significantly out-perform both tests. It should be considered particularly useful when little is known about the genetic model.

In linkage analysis, genome regions are sought for where marker allele transmissions from parents to children are correlated with phenotypes. Underlying linkage is the occurrence of crossovers in meioses and occurs for all markers associated to the disease locus. In association analysis, one searches regions of non-independence between phenotypes and marker alleles at the population level. Since both association and linkage tests use marker and phenotype data from a number of families, a combined linkage and association test optimally extracts information from data and hence should have greater power in detecting a disease susceptibility locus. If there is no linkage between the marker and the disease loci and no association between any particular allele variant at the marker with a variant at the putative locus, then for each sib, regardless of the size of the sibship, the affection status and the alleles at the marker are independent. Alternatively, if there is linkage but no association between any particular allele variant at the marker with a variant at the putative locus, linkage can cause excess sharing of marker alleles among affected siblings. When there is no evidence for linkage between a marker and disease locus but there is an association between alleles at the two loci, the affection status depends on alleles among affected siblings. A joint association and linkage test may have significant power even when there is no association between marker and disease genotypes or no evidence for linkage between two loci. This is not the case for methods based on transmitted and non-transmitted

founder alleles, such as the TDT-tests. Furthermore, linkage is modeled through the covariance structure, while the association along with other covariates are modeled via the regression parameter.

The aim of this thesis is to develop statistical methods for the joint modeling of genetic linkage and association, then apply the methods to family data that are computationally feasible and biologically sound. There are two distinct kinds of statistical goals: testing for association and testing for mapping. When testing the hypothesis for association, we try to explore how does genetic variation contributes to the phenotype. When mapping, our purpose is to determine the location of the variant(s). We have developed joint modeling of linkage and association for pedigrees by using a conditional likelihood approach for the phenotype functions. One of the objectives of this research is to extend linear regression or logistic functions that include other covariates into the model and use the well-known variance components model to test for the joint effects of linkage and association analysis in family studies. This variance-components approach allows simultaneous testing of the linkage parameter and association parameter, implying that all the information in a set of individuals can be used to construct a test of joint linkage and association.

Our model has some advantages. First, the joint test can increase the power of detecting disease locus when the marker and the disease loci are linked, and association between any particular allele variant at the marker and a variant at the putative locus. Second, the joint test can be applied when the linkage or association evidence vanishes entirely, the marker locus may be linked to disease locus or the disease itself, or in very

strong linkage disequilibrium with the disease alleles. Third, our proposed variance components model provides a flexible and powerful maximum-likelihood framework for further generalizations and extensions, such as multiple phenotypes. These further developments should lead to a set of powerful tools for the detection of disease loci and the dissection of complex traits in humans.

In Chapter 2, we review some basic concepts and methods of genetics. This will provide the basic knowledge that is required to read this dissertation.

In Chapter 3, we propose a quantitative trait joint linkage and association test based on variance components model that considers the association between markers and phenotypes, as well as linkage between marker and disease respective loci within families when a common ancestor or in population level is present. This test is based on a likelihood ratio test to overcome the usual identifiability issues that affect most of the standard methods intended to address this joint test. The parameter estimations are obtained through an implementation of the EM algorithm.

In Chapter 4, we consider LRT, Wald, and score test on testing the joint linkage and association components for binary phenotypes through a mixed model with multivariate normal random variables that are computationally feasible and biologically sound. We develop joint modeling of linkage and association for pedigrees that uses a conditional likelihood approach for the phenotype functions. These methods to general forms of logistic functions allow the inclusion of other covariates into the model and use those three tests to test for the joint effects of linkage and association analysis in family studies when the true parameter values may be on the boundary of parameter

space.

In Chapter 5, we derive a test statistic based on the score function for a general sampling distribution where an alternative hypotheses is based on a set of dependent unknown distribution random variables. This test requires an estimation of the model only under the null hypothesis, and the functional form of the test statistic is independent of the form of the mixing distribution.

In Chapter 6, we discuss the simulation results of the joint linkage and association test compared with linkage test and association test. The simulation shows that the joint test approach for treating both linkage and association provides rigorous inference, that is more accurate and more robust than linkage or association test alone.

In Chapter 7, we apply score statistical method to study of joint linkage and association to Familial Pulmonary Fibrosis, in addition, use two-point and multi-point linkage analysis by Merlin program to find the significant LOD scores that these loci may be linked with Familial Pulmonary Fibrosis.

Finally, in Chapter 8, the results obtained in this thesis are summarized, and the future research work are given.

Chapter 2

Review of Genetic Principles

2.1 Basic Concepts

A **gene** is a functional DNA unit which often encodes for a protein. Genes hold the information to build and maintain the cells in an organism. To introduce the basic genetic concepts, we give a diagram in Figure 2.1. The position of a gene on a

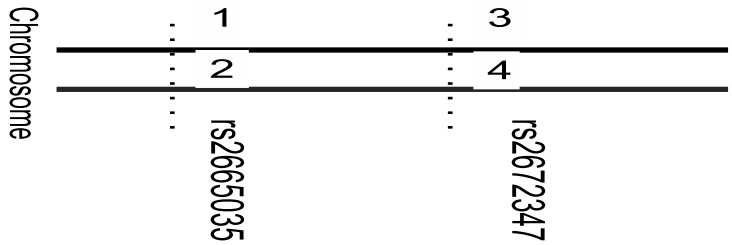


Figure 2.1: Diagram showing the basic genetic concepts

chromosome is known as its **locus** (rs2665035 and rs2672347). Variants of a DNA sequence at a locus among individuals are called **alleles**. In Figure 2.1, the two forms, 1 and 2, are alleles of locus rs2665035. **Allele frequency** is the number of copies of

a particular allele divided by the number of copies of all alleles at a particular locus in a population. A **genotype** (12 or 34 in Figure 2.1) is the combination the maternal and paternal inherited alleles at a particular locus. Also, broadly speaking, it is the genetic makeup of a cell, an organism, or an individual usually with reference to a specific character under consideration. The physical expression of the genotype is called the **phenotype**. Phenotype is an organism's actual observed properties, such as morphology, development, or behavior. Phenotypes result from the expression of an organism's genes, as well as the influence of environmental factors and the interactions between the two. It is generally accepted that inherited genotypes, epigenetic factors, and environmental variation contribute to the phenotype of an individual.

Genetic disease is a condition or state caused by the expression of one or more genes in a person, which results in a clinical phenotype. The goal of gene discovery is to locate these genes, usually called disease susceptibility loci, so that we can diagnose and/or develop treatments for these diseases. In this thesis, it is assumed that one or several genes cause a single disease. **Disease alleles** are passed from parents to offspring, but do not always result in a disease phenotype. The probability that a certain genotype causes a particular phenotype is called the **penetrance** of the genotype. In epidemiology, the penetrance of a disease-causing gene is the proportion of individuals with the disease-causing variant who exhibits clinical symptoms. For example, if a disease-causing gene responsible for a particular autosomal dominant disorder has 95% penetrance, then 95% of those individuals with one copy of the disease-causing variant will develop the disease, while 5% will not. For many hereditary diseases, the

onset of symptoms is age related and, in addition to the genetic determinants, it may also be affected by environmental factors such as nutrition and smoking, as well as epigenetic regulation of expression.

A **haplotype** in genetics is a combination of alleles (DNA sequences) at different places (loci) on the same chromosome that are transmitted together. A haplotype may be one locus, several loci, or an entire chromosome. In Figure 2.1, the first genotype has alleles 1 and 2, and the second genotype has alleles 3 and 4. The four possible haplotypes for these two genotypes are 13, 14, 23, and 24.

The amount of **linkage disequilibrium** (LD) is the difference between observed and expected allelic frequencies. In a diploid population, two alleles A_1 and A_2 are segregating at locus t_i, and alleles B_1 and B_2 are segregating at locus t_{i+1}. There are four possible gametes A_1B_1, A_1B_2, A_2B_1 and A_2B_2 with probilities $p_{A_1B_1}, p_{A_1B_2}, p_{A_2B_1}$ and $p_{A_2B_2}$. The expression of measures of LD value by Lewontin and Kojima (1960) is:

$$D = p_{A_1B_1}p_{A_2B_2} - p_{A_1B_2}p_{A_2B_1}$$

In case with $D = 0$ is called **linkage equilibrium**.

For biallelic markers, another useful and common measure (Hill and Robertson 1968) is the squared correlation between the presence and absence of alleles at different loci,

$$r^2 = D^2/[p_{A_1}p_{A_2}p_{B_1}p_{B_2}] \tag{2.1}$$

All these measures are closely related to each other and to the standard χ^2-statistic for a 2×2 contingency table. When "significant LD" is discussed, it is usually in the sense of a simple contingency table test of association even between unlinked loci. When the genotype of one of the loci perfectly predicts the other locus, $r^2 = 1$ implies that two cells in the 2×2 table are 0, and is referred to as perfect LD. r^2 also ranges from 0 to 1, and is the percentage of noncentrality parameter for an association test conducted at a marker in LD with the disease locus (Sham et al. 2000).

2.2 Genetic Recombination and Genetic Maps

Meiosis is the type of cell division by which germ cells (eggs and sperm) are produced. Meiosis involves a reduction in the amount of genetic material. At the beginning of meiosis, during the prophase, it occurs the phenomenon known as crossing over on which homologous chromosomes pair up, intertwines and exchanges sections of DNA material. The end result of this process are gametes with a new combination of genes that differs from the chromosomes found in the parents. Through this process of recombining genes, organisms can produce offspring with new combinations of maternal and paternal traits. Thus recombination can cause alleles previously on the same chromosome to be separated and end up in different daughter cells.

During meiosis, the maternal and paternal homologs of each chromosome pair together. Each chromosome consists of two sister chromatids. Whole two homologous chromosomes remain paired, they can exchange segments in a random way through a

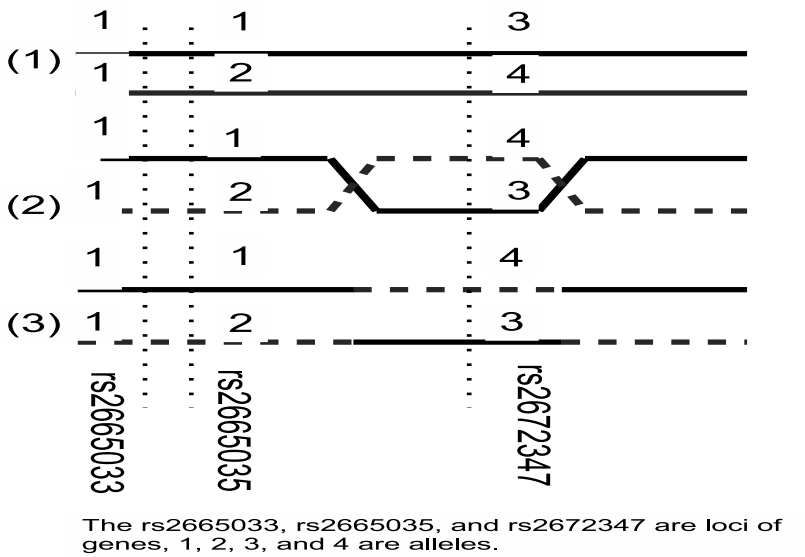


Figure 2.2: Basic symbols and terminologies in meiosis

process known as recombination (crossover). Recombination involves physical breakage of the double helix in one paternal and one maternal chromatid, and rejoining of maternal with paternal ends (Figure 2.2). (1) part of two chromatids of the two homologous chromosomes in a parent's cell, rs2665033, rs2665035, and rs2672347 are loci of genes. Two alleles at the same locus are denoted by numbers. For example, at locus rs2665033 the two alleles are 1 and 1 which inherited from one paternal and one maternal chromatid. (2) during meiosis, the two chromosomes may tangle together and exchange material. (3) after meiosis, the resulting gametes are formed. If there is a large distance between two loci, such as rs2665035 and rs2672347, there is a good chance that recombination will occur between them. However, if the two loci (rs2665033 and rs2665035) are close together recombination will rarely occur (the two loci will tend to stay together rather than being split apart by recombination).

Recombination frequency (r) is the proportion of progeny being recombinant with respect to a pair of loci on the same chromosome. A centiMorgan (cM) is a unit

that describes a recombination frequency. In this way we can measure the genetic distance between two loci, based upon their recombination frequency. If the loci lie on different chromosomes, in absence of interference, the recombinant fraction would be the half, because during meiosis the chromosomes assort randomly into gametes, such that the segregation of alleles on locus is independent of the segregation on the other, as stated in the Mendel's Second Law. For example, consider the crossing of the homozygote parental strain with genotype AABB with a different strain with genotype aabb, A and a and B and b represent the alleles of genes A and B assumed to be on different chromosomes. Crossing these homozygous parental strains will result in F1 generation offspring with genotype AaBb. The F1 offspring AaBb produces gametes that are AB, Ab, aB, and ab with equal frequencies (25%) because the alleles of gene A assort independently of the alleles for gene B during meiosis. Note that 2 of the 4 gametes (50 %) Ab and aB that represent recombination were not present in the parental generation, and they are the sole consequence of independent assortment. When two genes are on the same chromosome, they do not assort independently, a recombination frequency is less than 50%. The lower the recombination frequency between two loci, the more likely that they will segregate together and thus be closely linked.

The greater the frequency of recombination between two genetic loci, the farther apart they lie. Conversely, the lower the frequency of recombination between the markers, the smaller the genetic distance between them.

2.3 Identity by Descent(IBD) Estimation

When two individuals have the same allele (gene-variant) at a specific location, the alleles are considered to be **identical by state** (IBS). A pair of individuals can have zero, one or two alleles IBS. When the allele at the specific location is inherited from a common ancestor, the alleles are said to be **identical by descent** (IBD). Phenotypes of relatives are often similar because they may have similar genotypes and may share a common environment, and could have identical copies of a IBD gene segregating from a common ancestor. A **founder** within a set of pedigree data is defined as an individual whose neither the mother nor father is known. Such an individual may truly be a "founding ancestor" of a breed or "population" in the sense that it is not related to any other founder, or it may be related - possibly closely - to other members of the group but the details of this are not known. Because of the lack of information, we usually assume the genes in founders are not IBD. The calculation of IBD is based on the probability for the two alleles to be IBD. Mendel's first law states that: a diploid individual receives, at any given locus, a copy of a randomly chosen one of the two genes in his father and (independently) a copy of a randomly chosen one of the two genes in his mother, and will pass on a copy of a randomly and independently chosen one of these two genes to each of his offspring. The probability of IBD between more distant relatives are obtained by transmitting the parent-offspring information along the path between the relatives. When the information relies only on markers, it is rapidly eroded along the pedigree path due to recombination events during meiosis. Kinship and inbreeding are best thought of as relationships between gametes rather

than between individuals. The **coefficient of kinship** between two individuals B and C, denoted by $\psi(B, C)$, is the probability that homologous genes on gametes segregating from B and from C are IBD, while the **inbreeding coefficient** of an individual B, denoted by

$$f_B = \psi(M_B, F_B),$$

is the probability that homologous genes on the two gametes unite to form individual B are IBD. Where M_B and F_B are the parents of B. An individual is inbred if his parents are related. The process of observing probability of IBD starts with coefficients of inbreeding and kinship, since these provide an introduction to the ideas of gene identity by descent, to alternative computational approaches, and to Monte Carlo estimation of expectations.

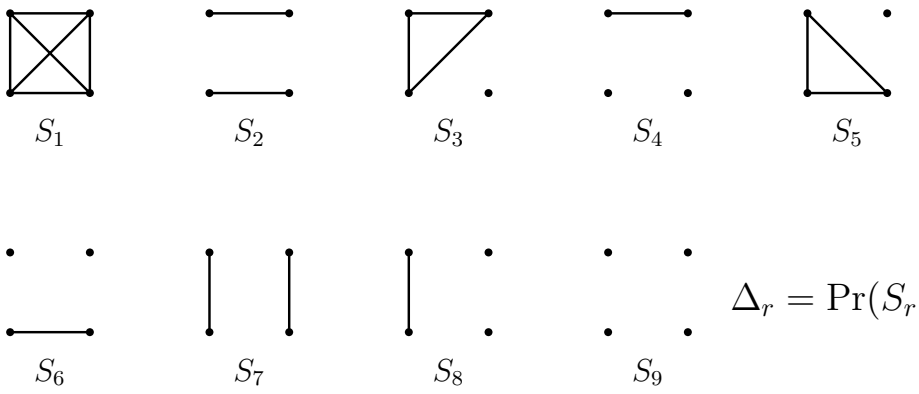
The early approach of path-counting (Wright 1922) for computing kinship coefficients simply enumerates all the possibilities (in an efficient way). Each path from the individual, B, to common ancestor, A, of its parents, descending via a disjoint set of individuals to B again contributes a term $2^{-(n_M+n_F+1)}(1+f_A)$ to the inbreeding coefficient f_B, where n_M and n_F are the number of segregations in the maternal and paternal lines of the path. If the common ancestor is inbred itself, its coefficient of inbreeding f_A must be worked out from its pedigree.

A parent and its offspring always have exactly one allele each that are IBD (the other allele in the offspring comes from the other parent.) Full sibs (a sibling with whom an individual shares the same biological parents), may have 0, 1, or 2 pairs of alleles that are IBD. Half sibs (shares the same mother but different father, or one

that shares the same father but different mother) and first cousins (share the same grandparents in common) may have 0 or 1 pairs of alleles that are identical by descent if the parents are not related.

To describe the relationship between two individuals there are nine condensed identity states ($S_r,\ r = 1, 2, \ldots, 9$) and the probabilities of these states are known as condensed identity coefficients (Lange 2002), which are denoted by Δ_r, $r = 1, 2, \ldots, 9$, then the kinship coefficient φ between these two individuals can be written as

$$\varphi_{jj'} = \Delta_{1,jj'} + \frac{1}{2}\left(\Delta_{3,jj'} + \Delta_{5,jj'} + \Delta_{7,jj'}\right) + \frac{1}{4}\Delta_{8,jj'}$$



There are many programs to compute the Δ coefficients, for example, 'parente' is a good C++ program to carry on the task that wrote by K. Morgan and J C. Loredo-Osti. The computation of these coefficients only requires knowledge of the pedigree. However, such a computation given the marker information at a fixed marker m, also requires the allele frequencies at such a locus, i.e.,

$$\varphi^{(m)}_{jj'} = \Delta^{(m)}_{1,jj'} + \frac{1}{2}\left(\Delta^{(m)}_{3,jj'} + \Delta^{(m)}_{5,jj'} + \Delta^{(m)}_{7,jj'}\right) + \frac{1}{4}\Delta^{(m)}_{8,jj'}$$

where

$$\Delta_{r,jj'}^{(m)} = \Pr\left(S_r \mid g_j^{(m)}, g_{j'}^{(m)}\right)$$

and $g_j^{(m)}, g_{j'}^{(m)}$ are the genotypes of the jth and j'th individuals at the marker m. For example with bi-allelic markers there are six distinguishable genotype pairs and their conditional probabilities given the identity state, $\Pr(g_j, g_{j'} \mid S_r)$, are presented in the following table.

	Condensed identity states								
Genotype pair	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
aa, aa	p_a	p_a^2	p_a^2	p_a^3	p_a^2	p_a^3	p_a^2	p_a^3	p_a^4
aa, ab	0	0	$p_a p_b$	$2p_a^2 p_b$	$p_a p_b$	$2p_a^2 p_b$	0	$2p_a^2 p_b$	$4p_a^3 p_b$
aa, bb	0	$2p_a p_b$	0	$p_a p_b$	0	$p_a p_b$	0	0	$2p_a^2 p_b^2$
ab, ab	0	0	0	0	0	0	$2p_a p_b$	$p_a p_b$	$4p_a^2 p_b^2$
ab, bb	0	0	$p_a p_b$	$2p_a p_b^2$	$p_a p_b$	$2p_a p_b^2$	0	$2p_a p_b^2$	$4p_a p_b^3$
bb, bb	p_b	p_b^2	p_b^2	p_b^3	p_b^2	p_b^3	p_b^2	p_b^3	p_b^4

Thus, an application of the Bayes Theorem yields

$$\Pr\left(S_r \mid g_j^{(m)}, g_{j'}^{(m)}\right) = \frac{\Delta_{r,jj'} \Pr\left(g_j^{(m)}, g_{j'}^{(m)} \mid S_r\right)}{\sum_{s=1}^{9} \Delta_{s,jj'} \Pr\left(g_j^{(m)}, g_{j'}^{(m)} \mid S_s\right)}$$

For polymorphic markers, the following table contains the relevant extensions to allow the computation of this conditional probability.

	Condensed identity states								
Genotype pair	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
aa, aa	p_a	p_a^2	p_a^2	p_a^3	p_a^2	p_a^3	p_a^2	p_a^3	p_a^4
aa, ab	0	0	$p_a p_b$	$2p_a^2 p_b$	$p_a p_b$	$2p_a^2 p_b$	0	$2p_a^2 p_b$	$4p_a^3 p_b$
aa, bb	0	$2p_a p_b$	0	$p_a p_b(p_a + p_b)$	0	$p_a p_b(p_a + p_b)$	0	0	$2p_a^2 p_b^2$
ab, ab	0	0	0	0	0	0	$2p_a p_b p_a p_b(p_a + p_b)$		$4p_a^2 p_b^2$
aa, bc	0	0	0	$2p_a p_b p_c$	0	$2p_a p_b p_c$	0	0	$4p_a^2 p_b p_c$
ab, ac	0	0	0	0	0	0	0	$2p_a p_b p_c$	$8p_a^2 p_b p_c$
ab, cd	0	0	0	0	0	0	0	0	$8p_a p_b p_c p_d$

2.4 Genome-Wide Association Study

In contrast to the methods that specifically test one or a few genetic regions, the **genome-wide association studies** (GWAS) investigates the entire genome based on single-marker analysis. It is a genetic association study design in which a sample of cases and controls, is genotyped for a large number of genetic markers. The ultimate aim of the GWAS design is to capture all common genetic variation across the genome and relate this variation to disease risk by case-control cohorts (Sullivan et al. 2001). Evidence for association is typically based on a simple statistical test of single SNPs, such as the chi-square test based on genotype counts with two degrees of freedom, or based on allele counts with 1 degrees of freedom. A standard linear or logistic regression is widely applied to the analysis of quantitative or binary outcomes

in population-based GWAS.

GWAS can also be used for family-based design. The advantage of this design is that it provides protection against spurious findings due to population stratification and other biases. Its significant disadvantage is inefficiency, as a large proportion of markers will have low power to detect association. One approach to the analysis of GWAS data is to compute power to detect association for each SNP and rank the SNPs by power with the primary analysis consisting of some number of SNPs with the greatest power (Sham 1998; Herbert et al. 2006). Aulchenko et al (2010) have designed genome-wide regression under linear, and logistic models for family-based association studies and genetically-isolated human populations.

As of present, over 1,200 human GWASs have been examined over 200 diseases and traits, and almost 4,000 SNP associations have been found throughout the human genome.

Chapter 3

Joint Testing for Quantitative Traits

3.1 Introduction

Linkage or association methods are widely used in the genetic analysis of quantitative traits in family studies, but using them jointly is not often done. Sham et al. (2000) derived analytical formulas for the noncentrality parameters for the linkage and association tests under a variance-components approach and showed empirically that the power of association is directly related to the QTL heritability and the power of linkage is related more closely to the square of the QTL heritability. However, their model makes no allowance for any correlation or interaction between the candidate gene and the environment. They consider six parameters: additive effect and dominance deviation for mean part, additive component and dominance component of QTL variance,

residual shared and nonshared variance for variance part. Linkage test is conducted by testing additive effect and dominance deviation. Overall association test is conducted by testing additive component and dominance component of QTL variance. Those testing parameters may reach the boundary. Therefore, the distribution of the likelihood ratio statistics of linkage, association, or both are complicated. Hössjer (2005b) introduced a combined score test for association and linkage analysis for quantitative traits based on a retrospective likelihood of marker data when given phenotypes, treating the alleles of the causal gene as hidden data with association between markers and causal genes, and penetrance between phenotypes and the causal gene. It is common to use a multivariate distribution of phenotypes for giving genotypes of family members. This mixed model incorporates effects of the major gene, G, only in the mean vector; the covariance matrix is independent of G. This method performs well for small pedigrees with a very simple structure, but the test statistic may be hard to calculate in large and complicated pedigrees.

In this chapter, we use the well-known variance-components model to test the joint effects of linkage and association analysis of quantitative traits in relatively large and complicated pedigrees that may have a remote common ancestor, or in population level. An EM algorithm implementation for parameter estimation is proposed. A likelihood ratio test is constructed to decide the significance of the hypothesis induced by the model of no association and no linkage, the association test is two-sided, the linkage test is a one-sided test, and those parameters may reach the boundary for joint linkage and association.

3.2 Methods

Consider a phenotype of interest measured in N independent families, each one consistent of n_i related individuals, where $\sum n_i = n$ individuals. Let y_{ij} and $\mathbf{x}_{ij} = (x_{ij1}, ..., x_{ijs})$ denote the observed trait and an s-dimension covariates vector respectively for individual j in family i. Similarly, for each SNP in the data set, label the two alleles as "A" and "a" and define a genotype score, g_{ij} , as the counting of "A" alleles in the genotype. Based on phenotypes, marker data, and covariates from all families at each locus, we would like to test

H_0 : marker is not linked to disease locus nor associated to disease genotypes

H_1 : marker is linked to disease locus and/or associated to disease genotypes

Now consider the set of phenotypes $\mathbf{y}_i = \{y_{i1}, y_{i2}, ..., y_{in_i}\}$, genotypes $\mathbf{g}_i = \{g_{i1}, g_{i2}, ..., g_{in_i}\}$, and covatiates $\mathbf{x}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{in_i}\}$ for each individuals in the ith family. Given a fixed locus in family i, the particular model is:

$$\mathbf{y}_i = \mu\mathbf{1} + \alpha\mathbf{g}_i + \mathbf{x}_i\boldsymbol{\gamma} + \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i \tag{3.1}$$

Assume random variables in the ith family $\boldsymbol{\xi}_i \sim N(\mathbf{0}, \beta^2\boldsymbol{\Sigma}_{\xi_i})$, and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. μ is overall mean, and σ^2 is the residuals variance. In family i, $\boldsymbol{\Sigma}_{\xi_i}$ is a known positive definite identity-by-descent (IBD) matrix at tested locus and $\mathbf{I}$ is identity matrix. The hypotheses of interest involve parameter α and the variance component β^2. The param-

eter α quantifies association between $\mathbf{y}$ and $\mathbf{g}$, β^2 quantifies linkage between marker locus and disease locus, and $\boldsymbol{\gamma}$ is a nuisance parameter. If $\alpha = 0$, the phenotype and marker alleles are not associated. Otherwise, the phenotype are associated with the marker gene. If $\beta^2 > 0$, the marker locus and disease locus are linked together. If the estimation gives a negative estimate of β^2 due to random sampling, but we know that a variance component cannot be negative, we use zero instead of a negative number of β^2, the marker locus does not linked with disease locus, the only association effect α be tested. With the normality assumption, the model given in (3.1) can be expressed as follows:

$$\boldsymbol{y}_i \sim N(\mu\mathbf{1} + \alpha\, \boldsymbol{g}_i + \mathbf{x}_i\boldsymbol{\gamma}, \beta^2\, \boldsymbol{\Sigma}_{\xi_i} + \sigma^2\mathbf{I}) \tag{3.2}$$

and the log-likelihood can be written as:

$$\begin{aligned} l(\boldsymbol{\theta}) \;=\; & -\frac{1}{2}\sum_{i=1}^{N}\big[\ln|\beta^2\boldsymbol{\Sigma}_{\xi_i} + \sigma^2\mathbf{I}| + \\ & (\mathbf{y}_i - \mu\mathbf{1} - \alpha\mathbf{g}_i - \mathbf{x}_i\boldsymbol{\gamma})^T(\beta^2\boldsymbol{\Sigma}_{\xi_i} + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_i - \mu\mathbf{1} - \alpha\mathbf{g}_i - \mathbf{x}_i\boldsymbol{\gamma})\big] \end{aligned} \tag{3.3}$$

According to this parameterization, we rewrite the hypothesis test as:

$$H_0 : \alpha = 0 \text{ and } \beta^2 = 0$$

$$H_1 : \alpha \neq 0 \text{ and/or } \beta^2 > 0.$$

3.3 Estimation

Dempster et al. (1977) presented a general approach to compute the maximum-likelihood estimates iteratively when the observations can be viewed as incomplete data. This iterative algorithm consists of an expectation step followed by a maximization step and it is called the EM algorithm. First, in the E-step, find the expectation of the logarithm of the likelihood given the observed data and the current estimated value of the parameter. The second step of the EM algorithm, the M-step, maximize the expected log likelihood which yields the next value of the parameter. Using the new value of the parameter, compute the next E-step and continue. Dempster et al. (1977) proved that this iterative process converges to the maximum likelihood estimators. McLachlan and Krishnan (1997) derived the MLE of mixed model parameters. Sammel et al. (1997) discussed a general framework that EM algorithm was performed to find the estimates that maximize the likelihood. Harville (1977) applied maximum likelihood approaches to the estimation of variance components, and the estimation of the model's fixed and random effects. The problem of estimating variance components can be regarded as a special case of a general linear model problem in which the elements of the covariance matrix are known functions of a parameter vector to be estimated. Loredo−Osti (2014) proposed a bootstrapping procedure under a mixed model applied to quantitative trait locus mapping, implemented an application of ML theory to the estimation of variance components, and the fixed and random effects. We use a variance-components model and apply EM algorithm to estimate fixed and random parameters for several independent families that assumes a random vector $\boldsymbol{\xi}_i$ with a multinomial distribution in ith

family.

In order to apply the EM algorithm, first we assume the values of β^2 and σ^2 are known, and rewrite the model (3.2) as following:

$$\boldsymbol{y}_i \sim N(\mu\mathbf{1} + \alpha\, \boldsymbol{g}_i + \mathbf{x}_i\boldsymbol{\gamma}, \sigma^2\boldsymbol{\Sigma_i}) \tag{3.4}$$

where $\varsigma = \beta^2/\sigma^2, \boldsymbol{\Sigma_i} = \varsigma\, \boldsymbol{\Sigma}_{\xi_i} + \mathbf{I}$ is a known $n \times n$ positive definite matrix. ς represents the signal$-$to$-$noise ratio. By using the generalized least square method, we can obtain the following algorithm at iteration $m+1$ for all N families:

$$\begin{aligned} \boldsymbol{\theta}^{(m+1)} &= (\mu^{(m+1)}\, \alpha^{(m+1)}\, \boldsymbol{\gamma}^{(m+1)})^T \\ &= \left(\sum_{i=1}^{N} \mathbf{z}_i^T (\boldsymbol{\Sigma}_i^{(m)})^{-1}\mathbf{z}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i^T (\boldsymbol{\Sigma}_i^{(m)})^{-1}\mathbf{y}_i \end{aligned} \tag{3.5}$$

and the best unbiased predictor of $\boldsymbol{\xi}_i$ can be written as

$$\hat{\boldsymbol{\xi}}_i^{(m)} = \varsigma^{(m)}\boldsymbol{\Sigma}_{\xi_i}(\boldsymbol{\Sigma}_i^{(m)})^{-1}\left(\mathbf{y}_i - \mathbf{z}_i\boldsymbol{\theta}^{(m+1)}\right). \tag{3.6}$$

Also

$$\hat{\sigma}^{2(m+1)} = \sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{z}_i\boldsymbol{\theta}^{(m+1)})^T(\boldsymbol{\Sigma}_i^{(m)})^{-1}(\mathbf{y}_i - \mathbf{z}_i\boldsymbol{\theta}^{(m+1)})/(n-s-2) \tag{3.7}$$

$$\hat{\beta}^{2(m+1)} = \frac{1}{n}\sum_{i=1}^{N}\left(\hat{\boldsymbol{\xi}}_i^{(m)'}(\varsigma^{(m)}\boldsymbol{\Sigma}_{\xi_i})^{-1}\hat{\boldsymbol{\xi}}_i^{(m)} + \hat{\sigma}^{2(m+1)}\mathrm{tr}((\varsigma^{(m)}\boldsymbol{\Sigma}_{\xi_i})^{-1}\mathbf{C}_i^{(m)})\right) \tag{3.8}$$

with

$$\mathbf{C}_i^{(m)} = \left(\mathbf{I} - \mathbf{z}_i(\mathbf{z}_i'\mathbf{z}_i)^{-1}\mathbf{z}_i' + (\varsigma^{(m)}\boldsymbol{\Sigma}_{\xi_i})^{-1}\right)^{-1}$$

where $\mathbf{z}_i = (\mathbf{1}\ \boldsymbol{g}_i\ \mathbf{x}_i)$ is a known $n \times (s+2)$ matrix, and

$$\boldsymbol{\Sigma}_i^{(m)} \ = \ \varsigma^{(m)}\boldsymbol{\Sigma}_{\xi_i} + \mathbf{I}$$

Beginning with a reasonable initial guess about the parameters, the system of equations (3.5) to (3.8) provides an iterative algorithm that proceeds until the relative change in the estimated parameters is sufficiently small, such as 10^{-4} (Although in principle, the EM algorithm yield the overall maximum, the way we treat includes the possibility that any application of the procedure stops in a local maximum).

3.4 Test of Hypotheses on the Boundary of the Parameter Space

Berkhof and Snijders (2001) showed that the likelihood ratio test has the best power properties for a multilevel model with random coefficients if those correlations are large. The asymptotic distribution of likelihood ratio test statistic will be chi-squared when the null hypothesis values are interior points of the permissible parameter space. Under the following regularity conditions (Chernoff (1954)): (a) The parameter space Ω has finite dimension p, $\boldsymbol{\theta}$ is within of Ω; (b) it can take (up to third) derivatives of $\ln f(\boldsymbol{y}, \boldsymbol{\theta})$ with respect to all $\boldsymbol{\theta}$, $l'(\boldsymbol{\theta})$ denotes the p vector of first derivatives of $l(\boldsymbol{\theta})$,

$l''(\boldsymbol{\theta})$ denotes the $p \times p$ matrix of second derivatives of $l(\boldsymbol{\theta})$; (c) $E_{\boldsymbol{\theta}_0}[l'(\boldsymbol{\theta})] = 0$ and the Fisher information is positive and bounded; (d) simple algebra up to a second-order expansion of the log-likelihood is sufficient and valid; if the hypothesis that parameter $\boldsymbol{\theta}$ lies on a $p-s$ dimensional hyperplane of p dimensional space is true, the distribution of the likelihood ratio is asymptotically χ^2 with s degrees of freedom, while the value of the parameter is not a boundary point of both the set of $\boldsymbol{\theta}$ corresponding to the null and alternative hypothesis. However, it happens frequently that the population value of the parameter vector is a boundary point of the feasible region or at least is sufficiently close to the boundary of the region. If such a situation occurs, the asymptotic distribution of likelihood ratio test statistic will not be chi-squared. Moran (1971) studied the asymptotic behavior of maximum likelihood when the true parameter point in estimation problems lies on the boundary of the parameter space. Chant (1974) introduced the asymptotic tests when the parameter is on the boundary of a closed parameter space: the asymptotic distributional form of the maximum likelihood estimators is established under the null hypothesis. Shapiro (1985) presented the asymptotic distribution of the likelihood ratio test statistic that is a mixture of chi-squared distributions when the null hypothesis value is a boundary point of the feasible region. Self and Liang (1987) investigated the existence of a consistent maximum likelihood estimator, the large sample distribution of the estimator, and the large sample distribution of likelihood ratio statistic under regularity conditions, allowing the true parameter value to be on the boundary of the parameter space. The exact limiting distributions are complicated by the number of unknown parameters. In some relatively simple cases, the

limiting distributions of the maximum likelihood estimator and likelihood ratio statistics are mixtures of normals and mixtures of chi-squared distributions, respectively. Feng and McCulloch (1992) examined both the property of local maxima of the log-likelihood and asymptotic coverage probability using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space when the data are independent and identically distributed observations with a known density function.

In this chapter, we examine the statistical inference for variance components model with multivariate normal random effects, using maximum likelihood estimation and the generalized likelihood ratio when the true parameters of linkage and association are on the boundary of the parameter space.

3.4.1 Test of Joint Linkage and Association

Under H_0, the model is:

$$\mathbf{y} = \mu_0 \mathbf{1} + \mathbf{x}\boldsymbol{\gamma}_0 + \boldsymbol{\epsilon}_0 \tag{3.9}$$

where $\boldsymbol{\epsilon}_0 \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I})$. The estimators under the reduced model can be obtained through the procedure of linear regression.

Under H_a, the model is the same as model (3.1) and the parameter estimators $(\hat{\mu}, \hat{\alpha}, \hat{\boldsymbol{\gamma}}, \hat{\beta}^2, \hat{\sigma}^2)$ can be obtained through the procedure described in previous section. For our null hypothesis $H_0 : \alpha = 0$ and $\beta^2 = 0$, the alternative hypothesis has three

cases:

$$
\begin{aligned}
H_{A1} &: \alpha \neq 0,\ \beta^2 = 0 \\
H_{A2} &: \alpha = 0,\ \beta^2 > 0 \\
H_{A3} &: \alpha \neq 0,\ \beta^2 > 0
\end{aligned}
$$

Because of the three cases alternative hypothesis, the test statistic can be used for making inferences about signals arising from the linkage, the association, and both. Figure (3.1) identifies four regions indexed by estimated $|\alpha|$ and β^2; when both $|\alpha|$ and β^2 are approximately at their respective null values, there is apparently no signal. Otherwise, it yields a strong signal that can be detected by linkage or by association only, or both.

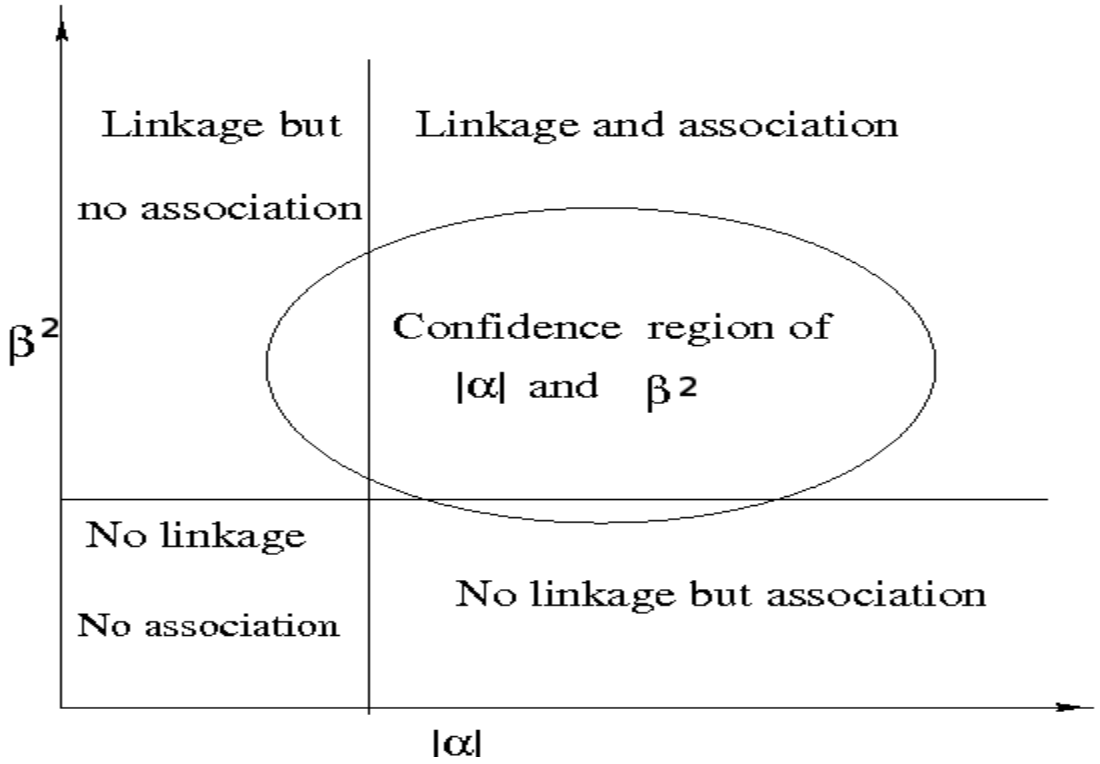


Figure 3.1: Illustration of a joint linkage/association analysis with one SNP marker

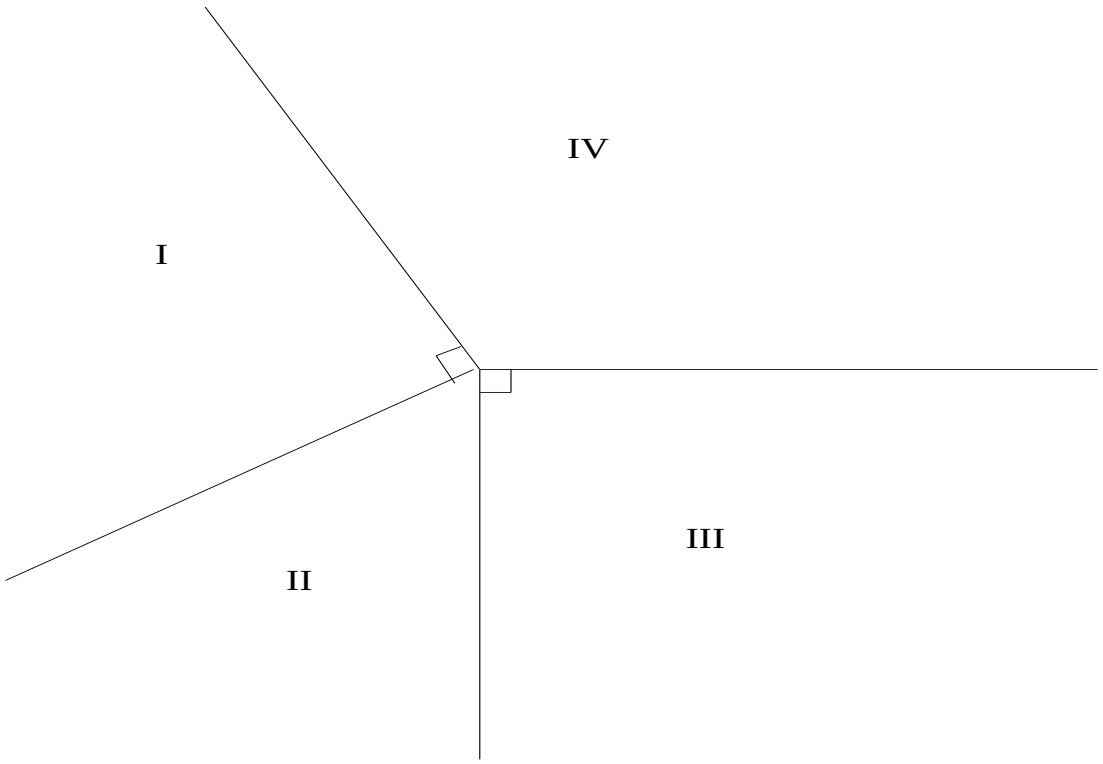


Figure 3.2: Diagram of the parameter space

Let $(\beta^2, \alpha, \mu, \boldsymbol{\gamma}, \sigma^2)$ be the dimension (1, 1, 1, s, 1) and let $(\mu, \boldsymbol{\gamma}, \sigma^2)$ represent nuisance parameters with true values not on the boundary. By the four tuple (Self and Liang 1987) of parameters of interest with true values on the boundary, parameters of interest with true values not on the boundary, nuisance parameters with true values on the boundary, and nuisance parameters with true values not on the boundary. For hypothesis test $H_0 : \beta^2 = 0, \alpha = 0$ vs. $\beta^2 > 0$ or/and $\alpha \neq 0$, since β^2 is nonnegative, the value of likelihood ratio test statistic is set equal to test case H_{A1} if the estimate of β^2 is not positive. We extend a result of Self and Liang (1987, case 7) regarding two boundary parameter $\{0\}$ of β^2 and α, where the parameter space is either $[0, \infty)$ or $(-\infty, \infty)$. So then, the parameter configuration will be $(2, 0, 0, s + 2)$. Figure (3.2) identifies the parameter space for this case. Region I with angle $\pi/2$ represents the likelihood ratio test for alternative case H_{A1}, which has a χ^2_1 distribution. Region

III with angle $\pi/2$ represents the likelihood ratio test for alternative case H_{A2}, which has a 50:50 mixture of χ^2_0 and χ^2_1 distribution. Region IV with angle ρ represents the likelihood ratio test for alternative case H_{A3}, which has a 50:50 mixture of χ^2_1 and χ^2_2 distribution. Self and Liang (1987) gave:

$$\rho = \arccos \frac{I_{12}}{\sqrt{I_{11} I_{22}}} \tag{3.10}$$

where the I'_{ij}s are the (i, j) entries of the information matrix under null hypothesis. From our model (3.1) and for the multivariate normal distribution assumption, α is independent of β^2, $\rho = \pi/2$. Finally, likelihood ratio test reduces to zero (or χ^2_0) in region II with angle $\pi - \rho = \pi/2$. So that the asymptotic distribution of the likelihood ratio test is as follows:

$$\begin{aligned} &\frac{1}{4}\chi^2_1 + \frac{1}{4}(\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1) + \frac{\rho}{2\pi}(\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2) + \frac{\pi - \rho}{2\pi}\chi^2_0 \\ &= \frac{3}{8}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{8}\chi^2_2 \end{aligned} \tag{3.11}$$

which is a mixture of χ^2_0, χ^2_1, and χ^2_2 distribution with mixing probabilities 3/8, 1/2, and 1/8 respectively.

3.4.2 Test of Association

When $\alpha \neq 0$ and β^2 is free, the model can be expressed same as (3.1) and the estimators under this model can be obtained through the procedure described in section 3.3.

When $\alpha = 0$ and β^2 is free, the reduced model can be expressed as

$$\mathbf{y}_i \;=\; \mu_0^\bullet \mathbf{1} + \mathbf{x}_i \boldsymbol{\gamma}_0^\bullet + \boldsymbol{\xi}_{i0}^\bullet + \boldsymbol{\epsilon}_0^\bullet \tag{3.12}$$

Similarly, the estimators can be obtained through the procedure described in the previous section by setting $\alpha = \alpha^{(m)} = 0$ for all m.

For testing association, the likelihood ratio test is approximately χ^2 distributed with 1 degree of freedom.

3.4.3 Test of Linkage

When α is free and $\beta^2 > 0$, the model and the parameter estimators are the same as before. Under $\beta^2 = 0$, the reduced model can be expressed as

$$\mathbf{y} \;=\; \mu^\star \mathbf{1} + \alpha^\star \mathbf{g} + \mathbf{x}\boldsymbol{\gamma}^\star + \boldsymbol{\epsilon}^\star \tag{3.13}$$

where $\boldsymbol{\epsilon}^\star \sim N(\mathbf{0}, \sigma^{\star 2}\mathbf{I})$. The estimators under the reduced model can be obtained through linear regression.

For testing linkage, the distribution of likelihood ratio test is approximately a half-and-half mixture of a χ_1^2 variable and a point mass at 0 (Self and Liang, 1987).

Chapter 4

Binary Phenotype with Multivariate Normal Random Effects

4.1 Introduction

In epidemiology and human molecular genetics, it is common to have the binary outcome variables. Typical binary variables express the disease statements through response alternatives such that the individual phenotype is either present or absent. An example is the testing of the family-based joint linkage and association in human genetical studies with binary phenotypes, SNPs, and covariates.

As discussed in Chapter 3, models for continuous data that incorporate both fixed

and random effects (mixed models) are commonly used in genetic studies. It is usually assumed that the random variables have a multivariate normal distribution with mean vector zero and a covariance matrix depending on some variance components. Williams (1975) and Crowder (1978) hypothesized a mixing distribution directly based on the probability of success, but this approach does not easily generalize to multiple random effects. Zeger and Liang (1986) and Liang and Zeger (1986) have proposed an estimating equation approach, but their methods focus on the fixed effects and only estimate the variances and covariances as nuisance parameters. Prentice (1988) has considered extensions of the Zeger and Liang (1986) estimating equation approach, explicitly estimating the covariances as well. McCulloch (1994) presented a class of probit-normal models and described MLE of the parameters in the model by using EM algorithm (Dempster et al. 1977). This implementation of the EM algorithm is identical to the continuous case, which represents an unobserved continuous variable replaced by their expected values given the observed binary phenotypes.

Many authors consider a likelihood ratio test for binary variables with independent random variables in linear mixed models. Verbeke and Molenberghs (2000) used linear mixed models for independent random variables that has a known distribution. Baksh et al. (2007) presented an alternative likelihood-based method of analysis for ordered categorical phenotypes in nuclear families. Aulchenko et al (2010) have designed genome-wide regression that facilitates fast genome-wide association analysis under logistic model for family-based association studies and genetically-isolated human populations.

We apply LRT, Wald, and score test on testing the joint linkage and association components for binary phenotypes with dependent random variables in non-linear mixed model. We have developed a joint modeling of linkage and association for pedigrees that uses a conditional likelihood approach for the phenotype functions. One of the objectives of our study is to extend the method to general forms of logistic functions, allow the inclusion of other covariates into the model. When the true parameter values may be on the boundary of parameter space, we use the above mentioned three tests to test the joint effects of linkage and association analysis in family studies. This proposed mixed model with random effects due to linkage and/or polygenic factors provides a flexible and powerful framework for further generalizations and extensions, such as multiple phenotypes. These further developments should lead to a set of powerful tools for the detection of disease loci and the dissection of complex traits in humans.

4.2 Mixed Model without Linkage Effects

Suppose that we have a random sample $y_{i1}, y_{i2}, \ldots, y_{in_i}$, with $y_{ij} \in \{0, 1\}$, $j = 1, 2, \ldots, n_i$ in ith family and each y_{ij} has an associated vector of covariates $\mathbf{z}_{ij} = \{1, g_{ij}, \mathbf{x}_{ij}\} \in \mathbb{R}^{S+2}$, $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, n_i$. Additionally, assume that for each individual there is an independent random variable $\xi_{ij} \in \mathbb{R}$ such that, for given ξ_{ij} and $\mathbf{z}_{ij}$, the random vector variable y_{ij} has a binomial distribution with parameter

$\pi_{ij}(\xi_{ij})$, where

$$\begin{aligned}\pi_{ij}(\xi_{ij}) &= \Pr\left(y_{ij} = 1 \mid \mathbf{z}_{ij}, \xi_{ij}\right) \\ &= \frac{\mathrm{e}^{\mathbf{z}_{ij}\boldsymbol{\theta}+\xi_{ij}}}{1+\mathrm{e}^{\mathbf{z}_{ij}\boldsymbol{\theta}+\xi_{ij}}} \quad \text{for } i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, n_i.\end{aligned} \tag{4.1}$$

Define $\boldsymbol{\theta} = \{\mu, \alpha, \boldsymbol{\gamma}\}'$, $\zeta_{ij} = e^{\mathbf{z}_{ij}\boldsymbol{\theta}}$, $\zeta_{ij}(\xi) = \mathrm{e}^{\xi}\,\zeta_{ij}$ and $\lambda_{ij}(\xi) = (1+\zeta_{ij}(\xi))^{-1}$ so that $\pi_{ij}(\xi) = \zeta_{ij}(\xi)\lambda_{ij}(\xi)$. Thus, if each ξ_{ij} has a density parameterized by $\boldsymbol{\vartheta}$, say $\mathrm{f}_{\boldsymbol{\vartheta}}(\cdot)$, for $i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, n_i$

$$\mathrm{p}_{ij}\left(y_{ij} \mid \mathbf{z}_{ij}, \xi\right) = \zeta_{ij}^{y_{ij}}(\xi)\lambda_{ij}(\xi) \tag{4.2}$$

$$\begin{aligned}\Pr\left(y_{ij} \mid \mathbf{z}_{ij}\right) &= \int_{-\infty}^{\infty} \mathrm{p}_{ij}\left(y_{ij} \mid \mathbf{z}_{ij}, \xi\right)\ \mathrm{f}_{\boldsymbol{\vartheta}}(\xi)\ d\xi \\ &= \mathrm{E}_{\boldsymbol{\vartheta}}\left(\pi_{ij}(\xi) \,\middle|\, \mathbf{z}_{ij}\right)^{y_{ij}} \mathrm{E}_{\boldsymbol{\vartheta}}\left(\lambda_{ij}(\xi) \,\middle|\, \mathbf{z}_{ij}\right)^{1-y_{ij}} \\ &= y_{ij} + (-1)^{y_{ij}} \int_{-\infty}^{\infty} \lambda_{ij}(\xi)\, \mathrm{f}_{\boldsymbol{\vartheta}}(\xi)\ d\xi \\ &= \pi_{ij},\end{aligned} \tag{4.3}$$

i.e., π_{ij} is a function of y_{ij} and $\mathbf{z}_{ij}$ only. Consequently, conditional on $\{\mathbf{z}_{ij}\}$,

$$\Pr\left(y \mid \{\mathbf{z}_{ij}\}\right) \;=\; \prod_{i=1}^{N}\prod_{j=1}^{n_i} \pi_{ij} \tag{4.4}$$

and the log-likelihood can be written as

$$\ell(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log \pi_{ij}$$

with $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ being the parameters of the logistic model and over-dispersion distribution, respectively. $\ell_i(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is the ith family log-likelihood. Under mild assumptions regarding $\mathrm{f}_{\boldsymbol{\vartheta}}$, we have

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} &= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} \Big(\lambda_{ij}(\xi)\, \zeta_{ij}^{y_{ij}}\, \mathrm{e}^{\xi\, y_{ij}} \Big)\, \frac{\mathrm{f}_{\boldsymbol{\vartheta}}(\xi)}{\mathrm{f}\left(y_{ij} \mid \mathbf{z}_{ij}\right)}\, d\xi \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \int_{-\infty}^{\infty} \mathbf{z}_{ij} \Big(y_{ij} - \pi_{ij}(\xi) \Big) \frac{\mathrm{f}\left(y_{ij} \mid \mathbf{z}_{ij}, \xi\right) \mathrm{f}_{\boldsymbol{\vartheta}}(\xi)}{\mathrm{f}\left(y_{ij} \mid \mathbf{z}_{ij}\right)}\, d\xi \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \int_{-\infty}^{\infty} \mathbf{z}_{ij} \Big(y_{ij} - \pi_{ij}(\xi) \Big) \mathrm{f}\left(\xi \mid \mathbf{z}_{ij}, y_{ij}\right)\, d\xi \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{z}_{ij} \Big(y_{ij} - \mathrm{E}_{\xi}\left(\pi_{ij}(\xi) \,\middle|\, \mathbf{z}_{ij}, y_{ij}\right) \Big) \qquad (4.5) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} (-1)^{1-y_{ij}} \mathbf{z}_{ij}\, \frac{\mathrm{E}_{\xi}\big(\mathrm{Var}(y_{ij} \mid \mathbf{z}_{ij}, \xi_{ij})\big)}{\pi_{ij}} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{z}_{ij} \left(\frac{y_{ij} - \mu_{ij}}{\mu_{ij}(1-\mu_{ij})} \right) \mathrm{E}_{\xi}\big(\mathrm{Var}(y_{ij} \mid \mathbf{z}_{ij}, \xi_{ij})\big) \qquad (4.6)
\end{aligned}$$

where $\mu_{ij} = \mathrm{E}\left(y_{ij} \mid \mathbf{z}_{ij}\right)$ and $\mathrm{Var}(y_{ij}|\mathbf{z}_{ij}, \xi_{ij}) = \pi_{ij}(\xi_{ij})\big(1-\pi_{ij}(\xi_{ij})\big)$. To see this, from

$$\begin{aligned}
\frac{\partial \mathrm{p}_{ij}(y_{ij} \mid \mathbf{z}_{ij}, \xi)}{\partial \boldsymbol{\theta}} &= \mathbf{z}_{ij} \left(y_{ij} - \pi_{ij}(\xi) \right) \mathrm{p}_{ij}(y_{ij} \mid \mathbf{z}_{ij}, \xi) \qquad (4.7) \\
\frac{\partial \log\, \pi_{ij}}{\partial \boldsymbol{\theta}} &= \frac{1}{\pi_{ij}} \int_{-\infty}^{\infty} \frac{\partial \mathrm{p}_{ij}(y_{ij} \mid \mathbf{z}_{ij}, \xi)}{\partial \boldsymbol{\theta}}\, \mathrm{f}_{\boldsymbol{\vartheta}}(\xi)\, d\xi \qquad (4.8)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} &= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \int_{-\infty}^{\infty} \frac{\mathrm{f}\left(y_{ij} \mid \mathbf{z}_{ij}, \xi\right) \mathrm{f}'_{\boldsymbol{\vartheta}}(\xi)}{\mathrm{f}\left(y_{ij} \mid \mathbf{z}_{ij}\right)} \, d\xi \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \int_{-\infty}^{\infty} \frac{\mathrm{f}\left(y_{ij} \mid \mathbf{z}_{ij}, \xi\right) \mathrm{f}_{\boldsymbol{\vartheta}}(\xi)}{\mathrm{f}\left(y_{ij} \mid \mathbf{z}_{ij}\right)} \frac{\mathrm{f}'_{\boldsymbol{\vartheta}}(\xi)}{\mathrm{f}_{\boldsymbol{\vartheta}}(\xi)} \, d\xi \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \int_{-\infty}^{\infty} \frac{\partial \log \mathrm{f}_{\boldsymbol{\vartheta}}(\xi)}{\partial \boldsymbol{\vartheta}} \, \mathrm{f}\left(\xi \mid \mathbf{z}_{ij} \; y_{ij}\right) \; d\xi \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathrm{E}_{\xi} \left(\frac{\partial \log \mathrm{f}_{\boldsymbol{\vartheta}}(\xi)}{\partial \boldsymbol{\vartheta}} \,\middle|\, \mathbf{z}_{ij}, y_{ij} \right)
\end{aligned} \tag{4.9}$$

These expressions can be used to maximize $\ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ by iteratively solving

$$\frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \qquad \text{and} \qquad \frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = \mathbf{0}$$

In general, the above equations cannot be solved explicitly, instead, we can use a first order approximation to the score function

$$\frac{\partial \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}})}{\partial \hat{\boldsymbol{\theta}}} \approx \frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}'} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \tag{4.10}$$

or its delta method approximation

$$\frac{\partial \ell(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\vartheta}})}{\partial \hat{\boldsymbol{\theta}}} \approx \frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} - \mathrm{E}\left(- \frac{\partial^2 \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}'} \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \tag{4.11}$$

where

$$\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} &= \sum_{i=1}^{N} \frac{\partial^2 \ell_i(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{z}_{ij} \mathbf{z}'_{ij}\, \mathrm{E}_{\xi} \left(\big(y_{ij} - \pi_{ij}(\xi_{ij})\big)^2 - \pi_{ij}(\xi_{ij}) \left(1 - \pi_{ij}(\xi_{ij})\right) \,\middle|\, \mathbf{z}_{ij}, y_{ij} \right) \\
&\quad - \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{z}_{ij} \mathbf{z}'_{ij} \mathrm{E}^2_{\xi} \left((y_{ij} - \pi_{ij}) \big| \, \mathbf{z}_{ij}, y_{ij} \right) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{z}_{ij} \mathbf{z}'_{ij}\, \mathrm{E}_{\xi} \left(\big(y_{ij} - \pi_{ij}(\xi_{ij})\big)^2 - \pi_{ij}(\xi_{ij}) \left(1 - \pi_{ij}(\xi_{ij})\right) \,\middle|\, \mathbf{z}_{ij}, y_{ij} \right) \\
&\quad - \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{z}_{ij} \mathbf{z}'_{ij} \left(\frac{\mathrm{E}_{\xi}\big(\mathrm{Var}(y_{ij} \,|\, \mathbf{z}_{ij}, \xi_{ij})\big)}{\pi_{ij}} \right)^2
\end{aligned} \tag{4.12}$$

and

$$-\mathrm{E}\left(\frac{\partial^2 \ell(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \right) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(\frac{\mathbf{z}_{ij} \mathbf{z}'_{ij}}{\mu_{ij}(1 - \mu_{ij})} \right) \left(\mathrm{E}_{\xi}\big(\mathrm{Var}(y_{ij} \,|\, \mathbf{z}_{ij}, \xi_{ij})\big) \right)^2 \tag{4.13}$$

because

$$\frac{\partial^2 \mathrm{p}_{ij}(y_{ij} \,|\, \mathbf{z}_{ij}, \xi)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \mathbf{z}_{ij} \mathbf{z}'_{ij} \left(\big(y_{ij} - \pi_{ij}(\xi)\big)^2 - \pi_{ij}(\xi) \big(1 - \pi_{ij}(\xi)\big) \right) \mathrm{p}_{ij}(y_{ij} \,|\, \mathbf{z}_{ij}, \xi) \tag{4.14}$$

whose expectation is null, as well as

$$\mathrm{E}\left(\frac{1}{\pi_{ij}^2} \right) = \frac{1}{\mu_{ij}(1 - \mu_{ij})} \tag{4.15}$$

4.2.1 Computations

To evaluate the conditional expectations involved in the previous expressions we can use the following relationship. For $r = -1, 0, 1, 2, \ldots$, define

$$P_{ij}^{(r)} = \int_{-\infty}^{\infty} \pi_{ij}^{r+1}(\xi)\, \mathrm{f}_{\boldsymbol{\vartheta}}(\xi)\; d\xi \tag{4.16}$$

and

$$Q_{ij}^{(r)} = \int_{-\infty}^{\infty} \lambda_{ij}^{r+1}(\xi)\, \mathrm{f}_{\boldsymbol{\vartheta}}(\xi)\; d\xi \tag{4.17}$$

Then, we have that, for $r = 0, 1, 2, \ldots$,

$$\pi_{ij}\, \mathrm{E}_{\boldsymbol{\vartheta}} \left(\pi_{ij}^{r}(\xi_{ij}) \,\middle|\, \mathbf{z}_{ij}, y_{ij} \right) = (1 - y_{ij})\, P_{ij}^{(r-1)} + (-1)^{1-y_{ij}}\, P_{ij}^{(r)} \tag{4.18}$$

$$\pi_{ij}\, \mathrm{E}_{\boldsymbol{\vartheta}} \left(\lambda_{ij}^{r}(\xi_{ij}) \,\middle|\, \mathbf{z}_{ij}, y_{ij} \right) = y_{ij}\, Q_{ij}^{(r-1)} + (-1)^{y_{ij}}\, Q_{ij}^{(r)} \tag{4.19}$$

and

$$\begin{aligned} \mu_{ij} &= P_{ij}^{(0)} \\ &= 1 - Q_{ij}^{(0)} \end{aligned} \tag{4.20}$$

as well as

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\vartheta}}\big(\mathrm{Var}(y_{ij} \,|\, \mathbf{z}_{ij}, \xi_{ij})\big) &= P_{ij}^{(0)} - P_{ij}^{(1)} \\ &= Q_{ij}^{(0)} - Q_{ij}^{(1)} \end{aligned} \tag{4.21}$$

These expressions can be easily evaluated by quadrature methods. In this thesis, 128point Gauss Hermite quadrature was used.

4.2.2 Normal Case

If we assume that $\mathrm{f}_{\vartheta}(\cdot)$ has a normal distribution with null mean and variance $\vartheta > 0$. Then

$$\frac{\partial \ell(\boldsymbol{\theta}, \vartheta)}{\partial \vartheta} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(-\frac{1}{2\vartheta} + \frac{1}{2\vartheta^2} \mathrm{E}_{\xi_{ij}} \left(\xi_{ij}^2 \,|\, \mathbf{z}_{ij}, y_{ij} \right) \right) \tag{4.22}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(-\frac{1}{2\vartheta} \left(\frac{y_{ij} - \mu_{ij}}{1 - \mu_{ij}} \right) + \frac{1}{2\vartheta^2} \left(\frac{y_{ij} - \mu_{ij}}{\mu_{ij}(1 - \mu_{ij})} \right) \mathrm{E}_{\xi_{ij}} \left(\xi_{ij}^2 \, \pi_{ij}(\xi_{ij}) \right) \right) \tag{4.23}$$

and the remaining components of the Fisher information can be written as

$$-\mathrm{E}\left(\frac{\partial^2 \ell(\boldsymbol{\theta}, \vartheta)}{\partial \vartheta^2} \right) = \frac{1}{4\vartheta^4} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{\mu_{ij}^2 \vartheta^2 - 2\mu_{ij}\vartheta \mathrm{E}_{\xi_{ij}}\big(\xi_{ij}^2 \, \pi_{ij}(\xi_{ij})\big) + \mathrm{E}_{\xi_{ij}}^2\big(\xi_{ij}^2 \, \pi_{ij}(\xi_{ij})\big)}{\mu_{ij}(1 - \mu_{ij})} \tag{4.24}$$

as well as

$$-\mathrm{E}\left(\frac{\partial^2 \ell(\boldsymbol{\theta}, \vartheta)}{\partial \vartheta \partial \boldsymbol{\theta}'}\right) = \frac{1}{2\vartheta^2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{x}'_{ij}\, \mathrm{E}_{\xi_{ij}}\big(\mathrm{Var}(y_{ij} \,|\, \mathbf{z}_{ij}, \xi_{ij})\big) \left(\frac{\mathrm{E}_{\xi_{ij}}\big(\xi_{ij}^2\, \pi_{ij}(\xi_{ij})\big) - \vartheta \mu_{ij}}{\mu_{ij}(1-\mu_{ij})}\right) \tag{4.25}$$

4.3 Mixed Model with Linkage Component

In this case, it is common to assume that the vector of random effects $\boldsymbol{\xi}_i$ has null mean and variance $\vartheta \boldsymbol{\Sigma}_i$, where $\boldsymbol{\Sigma}_i$ is a positive definite matrix assumed to be known up to a multiplicative constant in ith family. Define

$$\Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i) = \prod_{j=1}^{n_i} \zeta_{ij}^{y_{ij}}(\xi_{ij})\, \lambda(\xi_{ij}). \tag{4.26}$$

Then,

$$\Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}) = \int_{\mathbb{R}^{n_i}} \Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i)\, \mathrm{f}(\boldsymbol{\xi}_i)\, d\boldsymbol{\xi}_i \tag{4.27}$$

By the same argument used before,

$$\frac{\partial \Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i)}{\partial \boldsymbol{\theta}} = \Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i) \sum_{j=1}^{n_i} \mathbf{z}_{ij} \left(y_{ij} - \pi_{ij}(\xi_{ij})\right) \tag{4.28}$$

and

$$\frac{\partial^2 \Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i) \sum_{j'=1}^{n_i} \sum_{j=1}^{n_i} \mathbf{z}_{ij'} \mathbf{z}'_{ij} \left(y_{ij'} - \pi_{ij'}(\xi_{ij'})\right) \left(y_{ij} - \pi_{ij}(\xi_{ij})\right)$$
$$- \Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i) \sum_{j=1}^{n_i} \mathbf{z}_{ij} \mathbf{z}'_{ij} \, \pi_{ij}(\xi_{ij}) \big(1 - \pi_{ij}(\xi_{ij})\big) \qquad (4.29)$$

so that

$$\frac{\partial \ell_i(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} = \mathrm{E}_{\boldsymbol{\xi}_i}\Big(\sum_{j=1}^{n_i} \mathbf{z}_{ij} \left(y_{ij} - \pi_{ij}(\xi_{ij})\right) \Big| \{\mathbf{z}_{ij}\}, \mathbf{y}_i \Big)$$
$$= \sum_{j=1}^{n_i} \mathbf{z}_{ij} y_{ij} - \mathrm{E}_{\boldsymbol{\xi}_i}\big(\sum_{j=1}^{n_i} \mathbf{z}_{ij} \pi_{ij}(\xi_{ij}) | \{\mathbf{z}_{ij}\}, \mathbf{y}_i \big) \qquad (4.30)$$

and

$$\frac{\partial^2 \ell_i(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -\left(\frac{\partial \ell_i(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} - \sum_{j=1}^{n_i} \mathbf{z}_{ij} \, y_{ij} \right) \left(\frac{\partial \ell_i(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta}} - \sum_{j=1}^{n_i} \mathbf{z}_{ij} \, y_{ij} \right)'$$
$$+ \mathrm{E}_{\boldsymbol{\xi}_i} \Bigg(\sum_{j'=1}^{n_i} \sum_{j=1}^{n_i} \mathbf{z}_{ij'} \mathbf{z}'_{ij} \, \pi_{ij'}(\xi_{ij'}) \, \pi_{ij}(\xi_{ij}) \qquad (4.31)$$
$$- \sum_{j=1}^{n_i} \mathbf{z}_{ij} \mathbf{z}'_{ij} \, \pi_{ij}(\xi_{ij}) \big(1 - \pi_{ij}(\xi_{ij})\big) \Big| \{\mathbf{z}_{ij}\}, \mathbf{y}_i \Bigg)$$

It is easy to show that

$$-\mathrm{E}\left(\frac{\partial^2 \ell_i(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right) = \mathrm{E}\Bigg(\mathrm{E}_{\boldsymbol{\xi}_i} \left(\sum_{j=1}^{n_i} \mathbf{z}_{ij} \left(y_{ij} - \pi_{ij}(\xi_{ij})\right) \Big| \{\mathbf{z}_{ij}\}, \mathbf{y}_i \right)$$
$$\mathrm{E}_{\boldsymbol{\xi}_i} \left(\sum_{j=1}^{n_i} \mathbf{z}'_{ij} \left(y_{ij} - \pi_{ij}(\xi_{ij})\right) \Big| \{\mathbf{z}_{ij}\}, \mathbf{y}_i \right) \Bigg) \qquad (4.32)$$

Now, under the assumption that $\boldsymbol{\xi}_i \sim N(\mathbf{0}, \vartheta\boldsymbol{\Sigma}_i)$ with $\vartheta > 0$,

$$\frac{\partial \ell_i(\boldsymbol{\theta}, \vartheta)}{\partial \vartheta} = -\frac{n_i}{2\vartheta} + \frac{1}{2\vartheta^2} \mathrm{E}_{\boldsymbol{\xi}_i} \left(\boldsymbol{\xi}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\xi}_i \,|\, \{\mathbf{z}_i\}, \mathbf{y}_i \right) \tag{4.33}$$

and

$$\begin{aligned} \frac{\partial^2 \ell_i(\boldsymbol{\theta}, \vartheta)}{\partial \vartheta^2} = &- \left(\frac{\partial \ell_i(\boldsymbol{\theta}, \vartheta)}{\partial \vartheta} \right)^2 - \left(\frac{n_i + 2}{\vartheta} \right) \left(\frac{\partial \ell_i(\boldsymbol{\theta}, \vartheta)}{\partial \vartheta} + \frac{n_i}{4\vartheta} \right) \\ &+ \frac{1}{4\vartheta^4} \mathrm{E}_{\boldsymbol{\xi}_i} \left((\boldsymbol{\xi}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\xi}_i)^2 \,|\, \{\mathbf{z}_{ij}\}, \mathbf{y}_i \right) \end{aligned} \tag{4.34}$$

by using the rules of conditional expectation

$$\begin{aligned} -\mathrm{E} \left(\frac{\partial^2 \ell_i(\boldsymbol{\theta}, \vartheta)}{\partial \vartheta^2} \right) &= \frac{1}{4\vartheta^4} \mathrm{E} \left(\mathrm{E}_{\boldsymbol{\xi}_i} \left(\boldsymbol{\xi}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}, \mathbf{y}_i \right) - n_i \vartheta \right)^2 \\ &= \frac{1}{4\vartheta^4} \mathrm{Var} \left(\mathrm{E}_{\boldsymbol{\xi}_i} \left(\boldsymbol{\xi}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}, \mathbf{y}_i \right) \right) \end{aligned} \tag{4.35}$$

we have

$$\begin{aligned} \frac{\partial^2 \ell_i(\boldsymbol{\theta}, \vartheta)}{\partial \boldsymbol{\theta} \partial \vartheta} = &- \left(\frac{\partial \ell_i(\boldsymbol{\theta}, \vartheta)}{\partial \boldsymbol{\theta}} - \sum_{j=1}^{n_i} \mathbf{z}_{ij} y_{ij} \right) \left(\frac{\partial \ell_i(\boldsymbol{\theta}, \vartheta)}{\partial \vartheta} - \frac{n_i}{2\vartheta} \right) \\ &- \frac{1}{2\vartheta} \mathrm{E} \left(\sum_{j=1}^{n_i} \mathbf{z}_{ij} \pi_{ij}(\xi_{ij}) \, \boldsymbol{\xi}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\xi} \,\Big|\, \{\mathbf{z}_{ij}\}, \mathbf{y}_i \right) \end{aligned} \tag{4.36}$$

and

$$-\mathrm{E}\left(\frac{\partial^2 \ell_i(\boldsymbol{\theta},\vartheta)}{\partial\vartheta\partial\boldsymbol{\theta}'}\right) = \frac{1}{2\vartheta^2}\mathrm{E}\Bigg(\left(\mathrm{E}_{\boldsymbol{\xi}_i}\left(\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i \mid \{\mathbf{z_{ij}}\},\mathbf{y}_i\right) - n_i\vartheta\right) \mathrm{E}_{\boldsymbol{\xi}_i}\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}'\left(y_{ij}-\pi_{ij}(\xi_{ij})\right)\Big|\,\{\mathbf{z}_{ij}\},\mathbf{y}_i\right)\Bigg) \tag{4.37}$$

When $\boldsymbol{\Sigma}_i$ can be written as

$$\boldsymbol{\Sigma}_i = \mathrm{I} + \varsigma\mathbf{A}_i$$

with unknown $\varsigma > 0$ and a nonnegative definite matrices A_i. Up to a multiplicative constant, we can think of A_i as the identity by descent matrix at any given locus in ith family, then ς can be thought as the signal-to-noise ratio at the given locus. Under this parameterizations, the score vector must be augmented with

$$\frac{\partial \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial\varsigma} = \frac{1}{2\vartheta}\,\mathrm{E}_{\boldsymbol{\xi}_i}\Big(\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i - \vartheta\,\mathrm{tr}\big(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\big)\;\Big|\;\{\mathbf{z}_{ij}\},\mathbf{y}_i\Big)\Big) \tag{4.38}$$

Similarly, the observed and Fisher information matrix must be augmented in a dimension while all the other elements of the score vector and information matrix remain the same. The additional elements of the observed and Fisher information matrix can be written as

$$\frac{\partial^2 \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial\varsigma^2} \approx -\left(\frac{\partial \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial\varsigma} + \frac{1}{2}\,\mathrm{tr}\big(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\big)\right)^2 + \frac{1}{2}\,\mathrm{tr}\big(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\big)$$
$$-\frac{1}{\vartheta}\,\mathrm{E}_{\boldsymbol{\xi}_i}\Big(\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i\Big|\;\{\mathbf{z}_{ij}\},\mathbf{y}_i\Big) + \frac{1}{4\vartheta^2}\,\mathrm{E}_{\boldsymbol{\xi}_i}\Big(\Big(\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i\Big)^2\Big|\;\{\mathbf{z}_{ij}\},\mathbf{y}_i\Big)$$

$$-\mathrm{E}\left(\frac{\partial^2 \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial \varsigma^2}\right) = \frac{1}{4\vartheta^2}\mathrm{Var}\left(\mathrm{E}_{\boldsymbol{\xi}_i}\left(\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}A\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i \,|\, \{\mathbf{z}_i\}, \mathbf{y}_i\right)\right) \tag{4.39}$$

and

$$\begin{aligned}\frac{\partial^2 \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial \varsigma \partial \boldsymbol{\theta}'} \approx &- \left(\frac{\partial \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial \boldsymbol{\theta}} - \sum_{j=1}^{n_i}\mathbf{z}_{ij}\mathbf{y}_i\right)\left(\frac{\partial \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial \varsigma} + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\right)\right) \\ &- \frac{1}{2\vartheta}\mathrm{E}_{\boldsymbol{\xi}_i}\Big(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\pi_{ij}(\xi_{ij})\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i\Big|\ \{\mathbf{z}_{ij}\}, \mathbf{y}_i\Big)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial \vartheta \partial \varsigma} \approx &- \left(\frac{\partial \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial \vartheta} + \frac{n_i}{2\vartheta}\right)\left(\frac{\partial \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial \varsigma} + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\right)\right) \\ &+ \frac{1}{4\vartheta^3}\mathrm{E}_{\boldsymbol{\xi}_i}\Big(\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i\Big|\ \{\mathbf{z}_{ij}\}, \mathbf{y}_i\Big)\end{aligned}$$

$$\begin{aligned}-\mathrm{E}\left(\frac{\partial^2 \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial \varsigma \partial \boldsymbol{\theta}'}\right) =&\frac{1}{2\vartheta}\mathrm{E}\Bigg(\mathrm{E}_{\boldsymbol{\xi}_i}\Big(\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i - \vartheta\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\right)\Big|\ \{\mathbf{z}_{ij}\}, \mathbf{y}_i\Big) \\ &\mathrm{E}_{\boldsymbol{\xi}_i}\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}'\left(y_{ij} - \pi_{ij}(\xi_{ij})\right)\Big|\ \{\mathbf{z}_{ij}\}, \mathbf{y}_i\right)\Bigg)\end{aligned} \tag{4.40}$$

$$\begin{aligned}-\mathrm{E}\left(\frac{\partial^2 \ell_i(\boldsymbol{\theta},\vartheta,\varsigma)}{\partial \vartheta \partial \varsigma}\right) =&\frac{1}{4\vartheta^3}\mathrm{E}\Bigg(\left(\mathrm{E}_{\boldsymbol{\xi}_i}\left(\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i - n_i\vartheta \,|\, \{\mathbf{z}_{ij}\}, \mathbf{y}_i\right)\right) \\ &\mathrm{E}_{\boldsymbol{\xi}_i}\Big(\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i - \vartheta\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\right)\Big|\ \{\mathbf{z}_{ij}\}, \mathbf{y}_i\Big)\Bigg)\end{aligned} \tag{4.41}$$

In the general mixed model, the evaluation of the score vector can be carried out

by direct Monte Carlo or by a combination of Monte Carlo and quadrature methods, while the Fisher information matrix can be evaluated by the Monte Carlo technique of importance sampling. For this work, the distribution described in the previous section was used an importance sampling distribution, i.e., assuming $\boldsymbol{\Sigma}_i = \mathrm{I}$, so that the sampling Monte Carlo sampling distribution was a set of independent $y_{ij} \sim \mathrm{Ber}(\mu_{ij})$, $j = 1, 2, \ldots, n_i$, with each μ_{ij} computed by 128-point Gauss-Hermite quadrature.

In some situation, particularly when the number of observations grows, the Monte Carlo scheme could be quite inefficient. In such cases, the Laplace approximation may be a good alternative.

The Laplace approximation for (4.27) goes as follows: define $F_0(\boldsymbol{\xi}_i)$ as

$$F_0(\boldsymbol{\xi}_i) = -\frac{1}{2\vartheta}\,\boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i + \log \Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i) \tag{4.42}$$

then

$$\begin{aligned} \Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}) &= \sqrt{\frac{1}{(2\pi\,\vartheta)^{n_i}\,|\boldsymbol{\Sigma}_i|}} \int_{\mathbb{R}^{n_i}} \mathrm{e}^{F_0(\boldsymbol{\xi}_i)}\, d\boldsymbol{\xi}_i \\ &\approx \sqrt{\frac{1}{\vartheta^{n_i}\,|\boldsymbol{\Sigma}_i|\,|\mathbf{R}_o|}}\, \mathrm{e}^{F_0(\breve{\boldsymbol{\xi}}_{io})} \end{aligned} \tag{4.43}$$

where $\breve{\mathbf{z}}_o$ is the solution to

$$\frac{\partial F_0(\boldsymbol{\xi}_i)}{\partial \boldsymbol{\xi}_i} = \mathbf{0}$$

and $\mathbf{R}_o$ is defined as

$$\mathbf{R}_o = -\frac{\partial^2 F_0(\boldsymbol{\xi}_i)}{\partial \boldsymbol{\xi}_i \, \partial \boldsymbol{\xi}_i'} \Big|_{\boldsymbol{\xi}_i = \breve{\boldsymbol{\xi}}_{io}}$$

i.e., $\breve{\boldsymbol{\xi}}_{io}$ satisfy

$$\breve{\boldsymbol{\xi}}_{io} = \vartheta \, \boldsymbol{\Sigma}_i \left(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{io}) \right) \tag{4.44}$$

and

$$\mathbf{R}_o = \frac{1}{\vartheta} \, \boldsymbol{\Sigma}_i^{-1} + \mathrm{Var}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{io}). \tag{4.45}$$

Similarly, to compute the sth entry in the score function (4.30), say ℓ_{is}', s = 1, 2, ... S+2, define $F_s(\boldsymbol{\xi}_i)$ as

$$F_s(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + \log g_s\big(\boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}\big) \tag{4.46}$$

where

$$g_s\big(\boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}\big) = \sum_{j=1}^{n_i} z_{ij,s} \, \pi_{ij}(\xi_{ij}).$$

Then

$$\ell_{is}' \approx S_{is} - \sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}_s|}} \, \mathrm{e}^{F_s(\breve{\boldsymbol{\xi}}_{is}) - F_0(\breve{\boldsymbol{\xi}}_{io})} \tag{4.47}$$

where

$$S_{is} = \sum_{j=1}^{n_i} z_{ij,s} \, y_{ij}$$

and $\breve{\boldsymbol{\xi}}_{is}$ and $\mathbf{R}_s$ can be computed by using

$$\breve{\boldsymbol{\xi}}_{is} = \vartheta \, \boldsymbol{\Sigma}_i \left(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{is}) + \mathbf{q}_s(\breve{\boldsymbol{\xi}}_{is}) \right) \tag{4.48}$$

and

$$\mathbf{R}_s = \frac{1}{\vartheta}\,\boldsymbol{\Sigma}_i^{-1} + \mathrm{Var}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{is}) + \mathbf{q}_s(\breve{\boldsymbol{\xi}}_{ij})\;\mathbf{q}_s'(\breve{\boldsymbol{\xi}}_{is}) - \mathrm{diag}\big(\big(\mathbf{1} - 2\,\mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{is})\big) \circ \mathbf{q}_s(\breve{\boldsymbol{\xi}}_s)\big) \tag{4.49}$$

with

$$\mathbf{q}_s(\boldsymbol{\xi}_i) = \frac{1}{g_s\big(\boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}\big)}\,\mathrm{Var}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i) \begin{pmatrix} x_{i1,s} \\ x_{i2,s} \\ \vdots \\ x_{in_i,s} \end{pmatrix} \tag{4.50}$$

and '$\mathbf{a} \circ \mathbf{b}$' denotes Hadamard product.

To find the Laplace approximation for (4.33), say $\ell'_{i\vartheta}$, define $F_\vartheta(\boldsymbol{\xi}_i)$ as

$$F_\vartheta(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + \log \boldsymbol{\xi}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\xi}_i \tag{4.51}$$

and compute

$$\ell'_{i\vartheta} \approx -\frac{n_i}{2\vartheta} + \frac{1}{2\vartheta^2}\sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}_\vartheta|}}\,\mathrm{e}^{F_\vartheta(\breve{\boldsymbol{\xi}}_{i\vartheta}) - F_0(\breve{\boldsymbol{\xi}}_{i_o})} \tag{4.52}$$

with $\breve{\boldsymbol{\xi}}_{i\vartheta} \neq \mathbf{0}$ and satisfying

$$\breve{\boldsymbol{\xi}}_{i\vartheta} = \vartheta\,\boldsymbol{\Sigma}_i\,\big(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\vartheta})\big) + \frac{2\vartheta}{\breve{\boldsymbol{\xi}}_{i\vartheta}'\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\vartheta}}\,\breve{\boldsymbol{\xi}}_{i\vartheta} \tag{4.53}$$

or, equivalently

$$\breve{\boldsymbol{\xi}}_{i\vartheta}'\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\vartheta} = 2\vartheta + \vartheta\,\breve{\boldsymbol{\xi}}_{i\vartheta}'\big(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\vartheta})\big)$$

and

$$\mathbf{R}_\vartheta = \frac{1}{\vartheta}\,\boldsymbol{\Sigma}_i^{-1} + \mathrm{Var}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\vartheta}) - \frac{2}{\breve{\boldsymbol{\xi}}_{i\vartheta}'\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\vartheta}}\,\boldsymbol{\Sigma}_i^{-1} + \left(\frac{2}{\breve{\boldsymbol{\xi}}_{i\vartheta}'\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\vartheta}}\right)^2 \boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\vartheta}\breve{\boldsymbol{\xi}}_{i\vartheta}'\boldsymbol{\Sigma}_i^{-1} \tag{4.54}$$

Finally, the Laplace approximation to (4.38), $\ell'_{i\varsigma}$, can be found as follows. Define

$$F_\varsigma(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + \log \boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i \tag{4.55}$$

and compute

$$\ell'_{i\varsigma} \approx -\frac{1}{2}\,\mathrm{tr}(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i) + \frac{1}{2\vartheta}\sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}_\varsigma|}}\,\mathrm{e}^{F_\varsigma(\breve{\boldsymbol{\xi}}_{i\varsigma}) - F_0(\breve{\boldsymbol{\xi}}_{io})} \tag{4.56}$$

with $\breve{\boldsymbol{\xi}}_{i\varsigma} \neq \mathbf{0}$ and satisfying

$$\breve{\boldsymbol{\xi}}_{i\varsigma} = \vartheta\,\boldsymbol{\Sigma}_i\left(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\varsigma})\right) + \frac{2\vartheta}{\breve{\boldsymbol{\xi}}_{i\varsigma}'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma}}\,\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma} \tag{4.57}$$

and

$$\begin{aligned}\mathbf{R}_\varsigma = \;& \frac{1}{\vartheta}\,\boldsymbol{\Sigma}_i^{-1} + \mathrm{Var}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\varsigma}) - \frac{2}{\breve{\boldsymbol{\xi}}_{i\varsigma}'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma}}\,\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1} \\ & + \left(\frac{2}{\breve{\boldsymbol{\xi}}_{i\varsigma}'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma}}\right)^2 \boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma}\breve{\boldsymbol{\xi}}_{i\varsigma}'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\end{aligned} \tag{4.58}$$

Now we would like to find the ss'th entry of the observed information matrix, or equivalently the ss'th entry of (4.31), say $\ell''_{iss'}$. To do so, define two functions $F^\circ_{ss'}(\boldsymbol{\xi}_i)$and

$F^{\bullet}_{ss'}(\boldsymbol{\xi}_i)$ as

$$F^{\circ}_{ss'}(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + \log g^{\circ}_{ss'}\big(\boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}\big) \quad \text{and} \quad F^{\bullet}_{ss'}(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + \log g^{\bullet}_{ss'}\big(\boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}\big) \tag{4.59}$$

where

$$\begin{aligned} g^{\circ}_{ss'}\big(\boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}\big) &= \sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} z_{ij,s}\, z_{ij',s'}\, \pi_{ij}(\xi_{ij}) \pi_{ij'}(\xi_{ij'}) \\ g^{\bullet}_{ss'}\big(\boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}\big) &= \sum_{j=1}^{n_i} z_{ij,s}\, z_{ij,s'}\, \pi_{ij}(\xi_{ij})(1 - \pi_{ij}(\xi_{ij})) \end{aligned}$$

Then

$$\ell''_{iss'} \approx -(\ell'_{is} - S_{is})(\ell'_{is'} - S_{is'}) + \sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}^{\circ}_{ss'}|}}\, \mathrm{e}^{F^{\circ}_{ss'}(\breve{\boldsymbol{\xi}}^{\,\circ}_{iss'}) - F_0(\breve{\boldsymbol{\xi}}_{io})} - \sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}^{\bullet}_{ss'}|}}\, \mathrm{e}^{F^{\bullet}_{ss'}(\breve{\boldsymbol{\xi}}^{\,\bullet}_{iss'}) - F_0(\breve{\boldsymbol{\xi}}_{io})} \tag{4.60}$$

with $\breve{\boldsymbol{\xi}}^{\,\circ}_{iss'}$ satisfying

$$\breve{\boldsymbol{\xi}}^{\,\circ}_{iss'} = \vartheta\, \boldsymbol{\Sigma}_i \left(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}^{\,\circ}_{iss'}) + \mathbf{q}_{ss'_\circ}(\breve{\boldsymbol{\xi}}^{\,\circ}_{iss'})\right) \tag{4.61}$$

and

$$\begin{aligned} \mathbf{R}^{\circ}_{ss'} \;=\; & \frac{1}{\vartheta}\, \boldsymbol{\Sigma}_i^{-1} + \mathrm{Var}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}^{\,\circ}_{iss'}) + \mathbf{q}_{ss'_\circ}(\breve{\boldsymbol{\xi}}^{\,\circ}_{iss'}) \mathbf{q}'_{ss'_\circ}(\breve{\boldsymbol{\xi}}^{\,\circ}_{iss'}) \\ & - \mathrm{diag}\big(\big(\mathbf{1} - 2\,\mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}^{\,\circ}_{iss'})\big) \circ \mathbf{q}_{ss'_\circ}(\breve{\boldsymbol{\xi}}^{\,\circ}_{iss'})\big) - \mathbf{B}^{\circ}_{ss'}(\breve{\boldsymbol{\xi}}^{\,\circ}_{iss'}) / g^{\circ}_{ss'}\big(\boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}\big) \end{aligned} \tag{4.62}$$

where

$$\mathbf{q}_{ss'_\circ}(\boldsymbol{\xi}_i) = \frac{1}{g^\circ_{ss'}\big(\boldsymbol{\xi}_i \,|\, \{\mathbf{z}_{ij}\}\big)} \operatorname{Var}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i)\, \mathbf{h}^\circ_{ss'}(\boldsymbol{\xi}_i)$$

and the elements of $\mathbf{h}^\circ_{ss'} \in \mathbb{R}^{n_i}$ and $\mathbf{B}^\circ_{ss'} \in \mathbb{R}^{n_i \times n_i}$ are given by

$$h^\circ_{ss',j'}(\boldsymbol{\xi}_i) = z_{ij',s} \sum_j z_{ij,s'}\, \pi_{ij}(\xi_{ij}) + z_{ij',s'} \sum_{j=1} z_{ij,s}\, \pi_{ij}(\xi_{ij})$$

and

$$B^\circ_{ss',jj'}(\boldsymbol{\xi}_i) = \begin{cases} 2\, z_{ij,s}\, z_{ij,s'}\, \left(\operatorname{Var}(y_{ij} \,|\, \mathbf{z}_{ij}, \xi_{ij})\right)^2 & \text{if } j = j' \\ (z_{ij,s}\, z_{ij',s'} + z_{ij,s'}\, z_{ij',s}) \operatorname{Var}(y_{ij} \,|\, \mathbf{z}_{ij}, \xi_{ij}) \operatorname{Var}(y_{ij'} \,|\, \mathbf{z}_{ij'}, \xi_{ij'}) & \text{otherwise.} \end{cases}$$

The equations for $\breve{\boldsymbol{\xi}}^{\bullet}_{i\,ss'}$ are the similar with '$\circ$' replaced by '$\bullet$' and the elements of $\mathbf{h}^\bullet_{ss'} \in \mathbb{R}^{n_i}$ and $\mathbf{B}^\bullet_{ss'}$ defined as

$$h^\bullet_{ss',j}(\boldsymbol{\xi}_i) = z_{ij,s}\, z_{ij,s'}(1 - 2\,\pi_{ij}(\xi_{ij}))$$

and

$$B^\bullet_{ss',jj'}(\boldsymbol{\xi}_i) = \begin{cases} 2\, z_{ij,s}\, z_{ij,s'}\, \left(\operatorname{Var}(y_{ij} \,|\, \mathbf{z}_{ij}, \xi_{ij})\right)^2 & \text{if } j = j' \\ 0 & \text{otherwise.} \end{cases}$$

To find the observed information Laplace approximation for (4.34), $\ell''_{\vartheta\vartheta}$, define

$$F_{\vartheta\vartheta}(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + 2\log \boldsymbol{\xi}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\xi}_i \tag{4.63}$$

and compute

$$\ell''_{i\vartheta\vartheta} \approx -{\ell'_{i\vartheta}}^2 - \left(\frac{n_i + 2}{\vartheta}\right)\left(\ell'_{i\vartheta} + \frac{n_i}{4\vartheta}\right) + \frac{1}{4\vartheta^4}\sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}_{\vartheta\vartheta}|}}\, \mathrm{e}^{F_{\vartheta\vartheta}(\breve{\boldsymbol{\xi}}_{i\vartheta\vartheta}) - F_0(\breve{\boldsymbol{\xi}}_{io})} \tag{4.64}$$

where $\breve{\boldsymbol{\xi}}_{i\vartheta\vartheta}$ can be found recursively by using the relationship

$$\breve{\boldsymbol{\xi}}_{i\vartheta\vartheta} = \vartheta\, \boldsymbol{\Sigma}_i \left(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \mid \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\vartheta\vartheta})\right) + \frac{4\vartheta}{\breve{\boldsymbol{\xi}}_{i\vartheta\vartheta}' \boldsymbol{\Sigma}_i^{-1} \breve{\boldsymbol{\xi}}_{i\vartheta\vartheta}} \breve{\boldsymbol{\xi}}_{i\vartheta\vartheta} \tag{4.65}$$

and

$$\begin{aligned} \mathbf{R}_{\vartheta\vartheta} =& \frac{1}{\vartheta}\boldsymbol{\Sigma}_i^{-1} + \mathrm{Var}(\mathbf{y}_i \mid \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\vartheta\vartheta}) - \frac{4}{\breve{\boldsymbol{\xi}}_{i\vartheta\vartheta}' \boldsymbol{\Sigma}_i^{-1} \breve{\boldsymbol{\xi}}_{i\vartheta\vartheta}} \boldsymbol{\Sigma}_i^{-1} \\ &+ 2\left(\frac{2}{\breve{\boldsymbol{\xi}}_{i\vartheta\vartheta}' \boldsymbol{\Sigma}_i^{-1} \breve{\boldsymbol{\xi}}_{i\vartheta\vartheta}}\right)^2 \boldsymbol{\Sigma}_i^{-1} \breve{\boldsymbol{\xi}}_{i\vartheta\vartheta} \breve{\boldsymbol{\xi}}_{i\vartheta\vartheta}' \boldsymbol{\Sigma}_i^{-1} \end{aligned} \tag{4.66}$$

Similarly, to find the sth entry of (4.36), say $\ell''_{is\vartheta}$, define

$$F_{s\vartheta}(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + \log \boldsymbol{\xi}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\xi}_i + \log \sum_{j=1}^{n_i} z_{ij,s} \pi_{ij}(\xi_{ij}) \tag{4.67}$$

Then

$$\ell''_{is\vartheta} \approx -\left(\ell'_{is} - S_{is}\right)\left(\ell'_{i\vartheta} + \frac{n_i}{2\vartheta}\right) - \frac{1}{2\vartheta^2}\sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}_{s\vartheta}|}}\,\mathrm{e}^{F_{s\vartheta}(\breve{\boldsymbol{\xi}}_{is\vartheta}) - F_0(\breve{\boldsymbol{\xi}}_{io})} \tag{4.68}$$

where $\breve{\boldsymbol{\xi}}_{is\vartheta}$ satisfies

$$\breve{\boldsymbol{\xi}}_{is\vartheta} = \vartheta\,\boldsymbol{\Sigma}_i\left(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \mid \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{is\vartheta}) + \mathbf{q}_s(\breve{\boldsymbol{\xi}}_{is\vartheta})\right) + \frac{2\vartheta}{\breve{\boldsymbol{\xi}}'_{is\vartheta}\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{is\vartheta}}\,\breve{\boldsymbol{\xi}}_{is\vartheta} \tag{4.69}$$

with $\mathbf{q}_s(\breve{\boldsymbol{\xi}}_{is\vartheta})$ defined by (4.50) and

$$\begin{aligned}\mathbf{R}_{s\vartheta} =& \frac{1}{\vartheta}\,\boldsymbol{\Sigma}_i^{-1} + \mathrm{Var}(\mathbf{y}_i \mid \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{is\vartheta}) - \mathrm{diag}\big(\big(\mathbf{1} - 2\,\mathrm{E}(\mathbf{y}_i \mid \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{is\vartheta})\big) \circ \mathbf{q}_s(\breve{\boldsymbol{\xi}}_{is\vartheta})\big) \\ &+ \mathbf{q}_s(\breve{\boldsymbol{\xi}}_{is\vartheta})\mathbf{q}'_s(\breve{\boldsymbol{\xi}}_{is\vartheta}) - \frac{2}{\breve{\boldsymbol{\xi}}'_{is\vartheta}\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{is\vartheta}}\,\boldsymbol{\Sigma}_i^{-1} + \left(\frac{2}{\breve{\boldsymbol{\xi}}'_{is\vartheta}\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{is\vartheta}}\right)^2 \boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{is\vartheta}\breve{\boldsymbol{\xi}}'_{is\vartheta}\boldsymbol{\Sigma}_i^{-1}\end{aligned} \tag{4.70}$$

The Laplace approximation for the remaining components of the observed information matrix can be found in the same way:

$$\ell''_{is\varsigma} \approx -\left(\ell'_{is} - S_{is}\right)\left(\ell'_{i\varsigma} + \frac{1}{2}\,\mathrm{tr}\big(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\big)\right) - \frac{1}{2\vartheta}\sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}_{s\varsigma}|}}\,\mathrm{e}^{F_{s\varsigma}(\breve{\boldsymbol{\xi}}_{is\varsigma}) - F_0(\breve{\boldsymbol{\xi}}_{io})} \tag{4.71}$$

with

$$F_{s\varsigma}(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + \log \boldsymbol{\xi}'_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i + \log\sum_{j=1}^{n_i} z_{ij,s}\pi_{ij}(\xi_{ij}) \tag{4.72}$$

$$\ell''_{i\vartheta\varsigma} \approx -\left(\ell'_{i\vartheta} + \frac{n_i}{2\vartheta}\right)\left(\ell'_{i\varsigma} + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\right)\right) + \frac{1}{4\vartheta^3}\sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}_{\vartheta\varsigma}|}}\,\mathrm{e}^{F_{\vartheta\varsigma}(\breve{\boldsymbol{\xi}}_{i\vartheta\varsigma}) - F_0(\breve{\boldsymbol{\xi}}_{io})} \quad (4.73)$$

with

$$F_{\vartheta\varsigma}(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + \log \boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i + \log \boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i \quad (4.74)$$

and

$$\begin{aligned}\ell''_{i\varsigma\varsigma} \approx -\left(\ell'_{i\varsigma} + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\right)\right)^2 + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\right) - \frac{1}{\vartheta}\sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}_{\varsigma\varsigma_\circ}|}}\,\mathrm{e}^{F_{\varsigma\varsigma_\circ}(\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ}) - F_0(\breve{\boldsymbol{\xi}}_{io})} \\ + \frac{1}{4\vartheta^2}\sqrt{\frac{|\mathbf{R}_o|}{|\mathbf{R}_{\varsigma\varsigma_\bullet}|}}\,\mathrm{e}^{F_{\varsigma\varsigma_\bullet}(\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet}) - F_0(\breve{\boldsymbol{\xi}}_{io})}\end{aligned} \quad (4.75)$$

with

$$F_{\varsigma\varsigma_\circ}(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + \log \boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i \quad (4.76)$$

and

$$F_{\varsigma\varsigma_\bullet}(\boldsymbol{\xi}_i) = F_0(\boldsymbol{\xi}_i) + 2\,\log \boldsymbol{\xi}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\xi}_i \quad (4.77)$$

where $\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ}$ and $\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet}$ satisfy

$$\begin{aligned}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ} &= \vartheta\,\boldsymbol{\Sigma}_i\left(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ})\right) + \frac{2\vartheta}{\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ}'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ}}\,\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ} \\ \breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet} &= \vartheta\,\boldsymbol{\Sigma}_i\left(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet})\right) + \frac{4\vartheta}{\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet}'\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet}}\,\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet}\end{aligned}$$

and

$$\mathbf{R}_{\varsigma\varsigma_\circ} = \frac{1}{\vartheta}\,\boldsymbol{\Sigma}_i^{-1} + \mathrm{Var}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ}) - \frac{2}{\breve{\boldsymbol{\xi}}'_{i\varsigma\varsigma_\circ}\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ}}\,\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}$$
$$+ \left(\frac{2}{\breve{\boldsymbol{\xi}}'_{i\varsigma\varsigma_\circ}\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ}}\right)^2 \boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\circ}\breve{\boldsymbol{\xi}}'_{i\varsigma\varsigma_\circ}\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}$$
$$\mathbf{R}_{\varsigma\varsigma_\bullet} = \frac{1}{\vartheta}\,\boldsymbol{\Sigma}_i^{-1} + \mathrm{Var}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet}) - 2\,\frac{2}{\breve{\boldsymbol{\xi}}'_{i\varsigma\varsigma_\bullet}\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet}}\,\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}$$
$$+ 2\,\left(\frac{2}{\breve{\boldsymbol{\xi}}'_{i\varsigma\varsigma_\bullet}\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet}}\right)^2 \boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}\breve{\boldsymbol{\xi}}_{i\varsigma\varsigma_\bullet}\breve{\boldsymbol{\xi}}'_{i\varsigma\varsigma_\bullet}\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i\boldsymbol{\Sigma}_i^{-1}$$

Beginning with a reasonable initial guess about the parameters, the system of equations above describes an iterative algorithm that proceeds until the relative change in the estimated parameters is sufficiently small, such as 10^{-4}, and the $(k+1)$th iteration of $\boldsymbol{\theta}$ and ϑ will be:

$$\begin{pmatrix} \hat{\boldsymbol{\theta}}^{(k+1)} \\ \hat{\vartheta}^{(k+1)} \\ \hat{\varsigma}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\theta}}^{(k)} \\ \hat{\vartheta}^{(k)} \\ \hat{\varsigma}^{(k)} \end{pmatrix} + \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}^{(k)}, \hat{\vartheta}^{(k)}, \hat{\varsigma}^{(k)}) \begin{pmatrix} \sum_i \ell'_{i\hat{\boldsymbol{\theta}}^{(k)}} \\ \sum_i \ell'_{i\hat{\vartheta}^{(k)}} \\ \sum_i \ell'_{i\hat{\varsigma}^{(k)}} \end{pmatrix} \tag{4.78}$$

or

$$\begin{pmatrix} \hat{\boldsymbol{\theta}}^{(k+1)} \\ \hat{\vartheta}^{(k+1)} \\ \hat{\varsigma}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\theta}}^{(k)} \\ \hat{\vartheta}^{(k)} \\ \hat{\varsigma}^{(k)} \end{pmatrix} + \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}^{(k)}, \hat{\vartheta}^{(k)}, \hat{\varsigma}^{(k)}; \mathbf{y}) \begin{pmatrix} \sum_i \ell_{i\hat{\boldsymbol{\theta}}^{(k)}} \\ \sum_i \ell_{i\hat{\vartheta}^{(k)}} \\ \sum_i \ell_{i\hat{\varsigma}^{(k)}} \end{pmatrix} \tag{4.79}$$

with

$$\mathbf{I}(\hat{\boldsymbol{\theta}}, \hat{\vartheta}, \hat{\varsigma}) = \begin{bmatrix} -\mathrm{E}\sum_i \ell''_{i\boldsymbol{\theta}\boldsymbol{\theta}'} & -\mathrm{E}\sum_i \ell''_{i\boldsymbol{\theta}\vartheta} & -\mathrm{E}\sum_i \ell''_{i\boldsymbol{\theta}\varsigma} \\ -\mathrm{E}\sum_i \ell''_{i\vartheta\boldsymbol{\theta}'} & -\mathrm{E}\sum_i \ell''_{i\vartheta\vartheta} & -\mathrm{E}\sum_i \ell''_{i\vartheta\varsigma} \\ -\mathrm{E}\sum_i \ell''_{i\varsigma\boldsymbol{\theta}'} & -\mathrm{E}\sum_i \ell''_{i\varsigma\vartheta} & -\mathrm{E}\sum_i \ell''_{i\varsigma\varsigma} \end{bmatrix}$$

and

$$\mathbf{I}(\hat{\boldsymbol{\theta}}, \hat{\vartheta}, \hat{\varsigma}; \mathbf{y}) = \begin{bmatrix} -\sum_i \ell''_{i\boldsymbol{\theta}\boldsymbol{\theta}'} & -\sum_i \ell''_{i\boldsymbol{\theta}\vartheta} & -\sum_i \ell''_{i\boldsymbol{\theta}\varsigma} \\ -\sum_i \ell''_{i\vartheta\boldsymbol{\theta}'} & -\sum_i \ell''_{i\vartheta\vartheta} & -\sum_i \ell''_{i\vartheta\varsigma} \\ -\sum_i \ell''_{i\varsigma\boldsymbol{\theta}'} & -\sum_i \ell''_{i\varsigma\vartheta} & -\sum_i \ell''_{i\varsigma\varsigma} \end{bmatrix}$$

4.4 Test of Joint Linkage and Association

For joint linkage and association test, we write the hypothesis test as:

$$H_o : \alpha = 0 \text{ and } \varsigma = 0$$

$$H_A : \alpha \neq 0 \text{ and/or } \varsigma > 0.$$

The hypotheses of interest involve parameter α and the signal-to-noise ratio ς. The parameter α quantifies association between $\mathbf{y}$ and $\mathbf{g}$, ς quantifies linkage between marker locus and disease locus. If $\alpha = 0$, the traits and the marker gene are not associated. Otherwise, the traits are associated with the marker gene. If $\varsigma > 0$, the marker locus

and disease locus are linked together. If the estimation gives a negative estimate of ς due to random sampling, but we know that signal-to-noise ratio cannot be negative, we use zero instead of a negative number of ς, the marker locus does not linked with disease locus, the only association effect α be tested.

Under H_o, the random vector variable y_{ij} has a binomial distribution with independent random variable ξ_{ij}, and the model is similar to (4.1), where $\xi_{ij} \sim N(0, \vartheta)$ with $\vartheta > 0$. The parameter estimators $(\hat{\mu}, \hat{\boldsymbol{\gamma}}, \hat{\vartheta})$ can be obtained through the procedure described in previous section.

Under H_A, the random vector $\mathbf{y}_i$ has a binomial distribution with dependent random variable $\boldsymbol{\xi}_i$, and the model is similar to (4.27), where $\boldsymbol{\xi}_i \sim N(\mathbf{0}, \vartheta\boldsymbol{\Sigma}_i)$ with $\vartheta > 0$. The parameter estimators $(\hat{\mu}, \hat{\alpha}, \hat{\boldsymbol{\gamma}}, \hat{\vartheta}, \hat{\varsigma})$ can be obtained through the procedure described in previous section. For our intersection null hypothesis $\alpha = 0$ and $\varsigma = 0$, the alternative hypothesis has three cases:

$$
\begin{aligned}
H_{A1} &: \alpha \neq 0,\ \varsigma = 0 \\
H_{A2} &: \alpha = 0,\ \varsigma > 0 \\
H_{A3} &: \alpha \neq 0,\ \varsigma > 0
\end{aligned}
$$

Because of the three cases alternative hypothesis, the test statistic can be used for making inferences about signals arising from the linkage, the association, and both. Parallel to the study in Chapter 3, for hypothesis test $H_o : \varsigma = 0, \alpha = 0$ vs. $\varsigma >$

0 or/and $\alpha \neq 0$, the likelihood ratio test is as follows:

$$-2\sum_{i=1}^{N}\{\log Pr(\mathbf{y}_i|\{\mathbf{z}_i\})|_{H_o} - \log Pr(\mathbf{y}_i|\{\mathbf{z}_i\})|_{H_1}\}$$

Under the Wald statistical test, the estimate $(\hat{\alpha}, \hat{\varsigma})$ of the parameter(s) of interest (α, ς) is compared with the proposed value (α_0, ς_0), with the assumption that the difference between the two for each parameter will be approximately normally distributed. Typically under $H_o : \alpha_0 = 0, \varsigma_0 = 0$, the Wald test statistic is,

$$(\hat{\alpha}, \hat{\varsigma})I(\hat{\alpha}, \hat{\varsigma}; \mathbf{y}) \begin{pmatrix} \hat{\alpha} \\ \\ \hat{\varsigma} \end{pmatrix}$$

where

$$I(\hat{\alpha}, \hat{\varsigma}; \mathbf{y}) = \begin{bmatrix} -\sum_i \ell''_{i\alpha\alpha} & -\sum_i \ell''_{i\alpha\varsigma} \\ -\sum_i \ell''_{i\varsigma\alpha} & -\sum_i \ell''_{i\varsigma\varsigma} \end{bmatrix}$$

at $(\hat{\mu}, \hat{\alpha}, \hat{\boldsymbol{\gamma}}, \hat{\vartheta}, \hat{\varsigma})$ under H_A.

The score test is a statistical test of a simple null hypothesis that a parameter of interest (α, ς) is equal to some particular value (α_0, ς_0). It is the most powerful test when the true value of (α, ς) is close to (α_0, ς_0). The main advantage of the score test is that it does not require an estimate of the information under the alternative hypothesis or unconstrained maximum likelihood. This makes testing feasible when the unconstrained maximum likelihood estimate is a boundary point in the parameter

space. Typically under $H_o : \alpha = 0, \varsigma = 0$, the score test is,

$$(\ell'_{i\alpha}, \ell'_{i\varsigma}) I^{-1}(\tilde{\alpha}, \tilde{\varsigma}; \mathbf{y}) \begin{pmatrix} \ell'_{i\alpha} \\ \ell'_{i\varsigma} \end{pmatrix}$$

at $(\tilde{\mu}, 0, \tilde{\boldsymbol{\gamma}}, \tilde{\vartheta}, 0)$ under H_o.

In particular, these three statistics do not follow a standard chi-square distribution because the true parameter values will reach the boundary parameter $\{0\}$, the asymptotic distribution of the three joint tests under H_o is :

$$(\frac{5}{8} - \frac{\rho}{2\pi})\chi_0^2 + (\frac{3}{8} + \frac{\rho}{4\pi})\chi_1^2 + \frac{\rho}{4\pi}\chi_2^2 \tag{4.80}$$

where

$$\rho = \arccos \frac{(-\sum_i \ell''_{i\alpha\varsigma})}{\sqrt{(-\sum_i \ell''_{i\alpha\alpha})(-\sum_i \ell''_{i\varsigma\varsigma})}} \tag{4.81}$$

at $(\hat{\mu}, \hat{\alpha}, \hat{\boldsymbol{\gamma}}, \hat{\vartheta}, \hat{\varsigma})$ for Wald test, and at $(\tilde{\mu}, 0, \tilde{\boldsymbol{\gamma}}, \tilde{\vartheta}, 0)$ for LR and score tests.

4.5 Test of Association

For association test, we write the hypothesis test as: $H_o : \alpha = 0$ vs. $H_1 : \alpha \neq 0$. When $\alpha = 0$ and ς is free, the vector of random effects $\boldsymbol{\xi}_i$ has null mean and variance $\vartheta^\bullet \boldsymbol{\Sigma}_i$, where $\boldsymbol{\xi}_i \sim N(\mathbf{0}, \vartheta^\bullet \boldsymbol{\Sigma}_i)$ with $\vartheta^\bullet > 0$ and $\boldsymbol{\Sigma}_i = \mathrm{I} + \varsigma \mathbf{A}_i$. Define $\boldsymbol{\theta}_0^\bullet = \{\mu_0^\bullet, \boldsymbol{\gamma}_0^\bullet\}', \mathbf{z}_{ij} =$

$\{1, \mathbf{x}_{ij}\}$, the estimators $\hat{\boldsymbol{\theta}}_0^\bullet, \hat{\vartheta}_0^\bullet$ and $\hat{\varsigma}_0^\bullet$ can be obtained through the procedure described in the previous section by setting $\alpha^\bullet = \alpha^{\bullet(k)} = 0$ for all k, and for fixed α, the estimators $\hat{\boldsymbol{\theta}}^\bullet, \hat{\vartheta}^\bullet$ and $\hat{\varsigma}^\bullet$ can be obtained by setting $\alpha^\bullet = \alpha^{\bullet(k)}$ for all k. Testing association, the likelihood ratio test:

$$-2\sum_{i=1}^{N}(\log Pr^\bullet(\mathbf{y}_i|\{\mathbf{z}_i\})|_{H_o} - \log Pr^\bullet(\mathbf{y}_i|\{\mathbf{z}_i\})|_{H_A}),$$

Wald test

$$\alpha^{\bullet 2}\,(-\sum_i \ell''_{i\alpha^\bullet\alpha^\bullet})|_{H_A},$$

and score test

$$(\sum_i \ell'_{i\alpha^\bullet})^2/(-\sum_i \ell''_{i\alpha^\bullet\alpha^\bullet})|_{H_o}$$

are approximately χ^2 distributed with the 1 degree of freedom.

4.6 Test of Linkage

For linkage test, we write the hypothesis test as: $H_o : \varsigma = 0$ vs. $H_1 : \varsigma > 0$. When α is free and $\varsigma = 0$, the random vector variable y_{ij} has a binomial distribution with parameter $\pi^\star_{ij0}$

$$\begin{aligned}\pi^\star_{ij0} &= \Pr\left(y_{ij} = 1 \mid \boldsymbol{\xi}_i, \xi_{ij}\right) \\ &= \frac{\mathrm{e}^{\boldsymbol{\xi}_i\boldsymbol{\theta}^\star_0+\xi_{ij}}}{1+\mathrm{e}^{\boldsymbol{\xi}_i\boldsymbol{\theta}^\star_0+\xi_{ij}}} \quad \text{for } i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, n_i.\end{aligned} \tag{4.82}$$

where $\boldsymbol{\theta}_0^\star = \{\mu_0^\star, \alpha_0^\star, \boldsymbol{\gamma}_0^\star\}'$, $\mathbf{z}_{ij} = \{1, g_{ij}, \mathbf{x}_{ij}\}$, and the independent random variable $\xi_{ij} \sim \mathrm{N}(0, \vartheta^\star)$. The estimators under the reduced model can be obtained through the procedure described in the previous section, and for fixed ς, the estimators $\hat{\boldsymbol{\theta}}^\star, \hat{\vartheta}^\star$ can be obtained by setting $\varsigma^\star = \varsigma^{\star(k)}$ for all k. The asymptotic distribution of the likelihood ratio test becomes:

$$-2\sum_{i=1}^{N}\{\log Pr^\star(\mathbf{y}_i|\{\mathbf{z}_i\})|_{H_o} - \log Pr^\star(\mathbf{y}_i|\{\mathbf{z}_i\})|_{H_A}\}$$

Wald test is:

$$\varsigma^{\star 2}\,(-\sum_i \ell''_{i\varsigma^\star\varsigma^\star})|_{H_A},$$

and score test is:

$$(\sum_i \ell'_{i\varsigma^\star})^2/(-\sum_i \ell''_{i\varsigma^\star\varsigma^\star})|_{H_o},$$

which are the mixture of χ_0^2 and χ_1^2 distribution with mixing probabilities 1/2 and 1/2 respectively.

Chapter 5

Binary Phenotype with Multivariate Random Effects - Distribution Unknown

For binary data, however, the multivariate normal distribution assumption may not be true or unknown population distribution sometimes, the parameters can not be estimated under the alternative hypothesis, LR and Wald tests can not be used, score test can be applied. Zelterman (1988) described the score function based on a set of mutually independent random variables for a general mixture sampling distribution. Jacqmin-Gadda and Commenges (1995) proposed a score test for testing homogeneity among clustered data, adjusting for the effects of covariates. Silvapulle and Silvapulle (1995) introduced a score type statistic for testing one-sided hypotheses for indepen-

dent and identically distributed observations. Lin (1997) developed a score test based on an integrated quasi-likelihood function for the marginal distribution of the response vector. Hall and Praestgaard (2001) introduced the restricted score tests that improves upon the earlier work of Lin (1997) in terms of efficiency.

Most of the literature on score test is concerned with independent and identically distributed observations. In this chapter, we discuss central mixture alternative hypotheses based on a set of dependent unknown distribution random variables. we explore a score test that is derived from a Taylor series expansion of the likelihood function for testing one-sided and two-sided hypotheses where the true parameter values may be on the boundary of parameter space. The main advantages of score tests are that they require estimation of models only under the null hypothesis that no random variables in model. If the exact population distribution is unknown or the exact likelihood is unknown, the parameters can not be estimated to the likelihood ratio and other equivalent forms under the alternative hypothesis. Therefore, score tests are convenient to apply because the full model does not need to be estimated, and we do not need to know the exact likelihood because score tests are based on estimating equations rather than likelihoods.

5.1 Model Specification

Suppose $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})$ is a vector of phenotypes with the binary variables taking values 0 and 1. 0 means that the jth individual in ith family is unaffected, while 1

means affected. Genotypes $\mathbf{g}_i = \{g_{i1}, g_{i2}, ..., g_{in_i}\}$ and covatiates $\mathbf{x}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{in_i}\}$ for each members in the ith family, and the particular model is:

$$\begin{aligned} \text{logit } P(\mathbf{y}_i = 1) &= \mu\mathbf{1} + \mathbf{x}_i\boldsymbol{\gamma} + \alpha\mathbf{g}_i + \sqrt{\beta}\boldsymbol{\xi}_i \\ &= \boldsymbol{\lambda}_i + \alpha\mathbf{g}_i + \sqrt{\beta}\boldsymbol{\xi}_i \end{aligned} \tag{5.1}$$

where μ is overall mean, α quantifies association between $\mathbf{y}$ and $\mathbf{g}$, β quantifies linkage between marker locus and disease locus, $\boldsymbol{\gamma}$ is nuisance parameter, random vector $\boldsymbol{\xi}_i$ has multivariate distribution function G with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\xi_i}$ that is known positive definite IBD matrix at tested locus in family i, and $\boldsymbol{\lambda}_i = \mu\mathbf{1} + \mathbf{x}_i\boldsymbol{\gamma}$. With this model, the hypothesis testing looks like:

$$H_o : \alpha = 0 \text{ and } \beta = 0$$

$$H_A : \alpha \neq 0 \text{ and/or } \beta > 0.$$

The hypotheses of interest involve parameter α and the random effect β. The parameter α quantifies association between $\mathbf{y}$ and $\mathbf{g}$, β quantifies linkage between marker locus and disease locus. If $\alpha = 0$, the traits and the marker gene are not associated. Otherwise, the traits are associated with the marker gene. If $\beta > 0$, the marker locus and disease locus are linked together. If the estimation gives a negative estimate of β due to random sampling, but we know that random effect cannot be negative, we use zero instead of a negative number of β, the marker locus does not linked with disease locus, the only association effect α be tested.

Again H_A includes three cases: $H_{A1} : \alpha \neq 0, \beta = 0$, $H_{A2} : \alpha = 0, \beta > 0$ and $H_{A3} : \alpha \neq 0, \beta > 0$. Under the null hypothesis ($H_0 : \alpha = 0, \beta = 0$), the reduced model will be:

$$\text{logit } P(\mathbf{y}_i = 1) = \mu_0 + \mathbf{x}_i\boldsymbol{\gamma}_0 = \boldsymbol{\lambda}_{i0} \tag{5.2}$$

$\mathbf{y}_i$ has independent identical distribution, such that the model can be expressed as a Bernoulli distribution:

$$y_{ij} \sim Ber(p_{ij}),$$

with the logistic link function:

$$p_{ij} = \frac{e^{\mu_0 + \mathbf{x}_{ij}\boldsymbol{\gamma}_0}}{1 + e^{\mu_0 + \mathbf{x}_{ij}\boldsymbol{\gamma}_0}}.$$

5.2 Test of Joint Linkage and Association

Under alternative hypothesis $\alpha \neq 0$ or/and $\beta > 0$, in ith family, $\boldsymbol{\lambda}_i = \{\lambda_{ij}\}$, and $\boldsymbol{\xi}_i = \{\xi_{ij}\}, j = 1, ..., n_i$ have multivariate distribution function G with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\xi_i}$, the likelihood function L_i is given by:

$$L_i = \int \cdots \int L(\boldsymbol{\lambda}_i, \alpha, \beta | \mathbf{y}_i) dG(\boldsymbol{\xi}_i). \tag{5.3}$$

Let $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_i\}$ and $l(\boldsymbol{\lambda}, \alpha, \beta) = \sum_{i=1}^{N} \ln L_i$ denotes the log-likelihood function of $(\boldsymbol{\lambda}, \alpha, \beta)$, the efficient score for $\alpha(\alpha = 0), \beta(\beta = 0)$ is denoted by:

$$\begin{aligned} U(\alpha, \beta) &= \left(\frac{\partial l}{\partial \alpha}\Big|_{\substack{\alpha=0 \\ \beta=0}} , \frac{\partial l}{\partial \beta}\Big|_{\substack{\alpha=0 \\ \beta=0}} \right) \\ &= \left(\sum_{i=1}^{N} \left[\frac{1}{L_i} \times \frac{\partial L_i}{\partial \alpha} \right] \Big|_{\substack{\alpha=0 \\ \beta=0}} , \sum_{i=1}^{N} \left[\frac{1}{L_i} \times \frac{\partial L_i}{\partial \beta} \right] \Big|_{\substack{\alpha=0 \\ \beta=0}} \right) \end{aligned} \tag{5.4}$$

The Taylor series expansion for $L(\boldsymbol{\lambda}_i, \alpha, \beta|\mathbf{y}_i)$ around $\boldsymbol{\lambda}_{i0}$ is

$$\begin{aligned} &L(\boldsymbol{\lambda}_i, \alpha, \beta|\mathbf{y}_i) = L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i) + L^{'}(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)(\alpha \mathbf{g}_i + \sqrt{\beta}\boldsymbol{\xi}_i) \\ &+\frac{1}{2}(\alpha \mathbf{g}_i + \sqrt{\beta}\boldsymbol{\xi}_i)^T L^{''}(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)(\alpha \mathbf{g}_i + \sqrt{\beta}\boldsymbol{\xi}_i) + \mathbf{r}_i \end{aligned} \tag{5.5}$$

where $L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)$ is the likelihood function under H_o,

$$L^{'}(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i) = \left(\frac{\partial L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{i10}}, \frac{\partial L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{i20}}, ..., \frac{\partial L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{in_i0}} \right) \Big|_{\boldsymbol{\lambda}_{i0}}$$

and

$$L^{''}(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i) = \left(\begin{array}{cccc} \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{i10}^2} & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{i1} \partial \lambda_{i20}} & \cdots & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{i10} \partial \lambda_{in_i0}} \\ \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{i20} \partial \lambda_{i10}} & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{i20}^2} & \cdots & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{i20} \partial \lambda_{in_i0}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{in_i0} \partial \lambda_{i10}} & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{in_i0} \partial \lambda_{i20}} & \cdots & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)}{\partial \lambda_{in_i0}^2} \end{array} \right) \Bigg|_{\boldsymbol{\lambda}_{i0}} \tag{5.6}$$

are the first two partial derivatives of $L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)$ with respect to $\boldsymbol{\lambda}_{i0}$, and $\mathbf{r}_i$ is the remainder term. Since

$$\begin{aligned}\int\cdots\int L^{'}(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)\boldsymbol{\xi}_i dG(\boldsymbol{\xi}_i) &= 0\\ \int\cdots\int \boldsymbol{\xi}_i^T L^{''}(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)\boldsymbol{\xi}_i dG(\boldsymbol{\xi}_i) &= \text{tr}[L^{''}(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)\Sigma_{\boldsymbol{\xi}_i}]\end{aligned}$$

L_i can be represented as

$$\begin{aligned}L_i &= L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i) + \alpha L'(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)\mathbf{g}_i + \frac{1}{2}\alpha^2\mathbf{g}_i^T L''(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)\mathbf{g}_i\\ &\quad +\frac{1}{2}\beta\text{tr}[L^{''}(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)\Sigma_{\boldsymbol{\xi}_i}] + \mathbf{R}_i \qquad (5.7)\end{aligned}$$

where $\mathbf{R}_i = \int\cdots\int \mathbf{r}_i dG(\boldsymbol{\xi}_i)$. From the result given in Zelterman and Chen (1988), $\partial^3 L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)/\partial\boldsymbol{\lambda}_{i0}^3$ is bounded, both $\partial\mathbf{R}_i/\partial\alpha$ and $\partial\mathbf{R}_i/\partial\beta$ approach to 0 for all individuals as α and $\beta \downarrow 0$. The efficient score $U(\alpha,\beta)$ then becomes:

$$\begin{aligned}U(\alpha,\beta) &= \left(\sum_{i=1}^{N} L^{'}(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)\mathbf{g}_i/L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i), \frac{1}{2}\sum_i \text{tr}[L^{''}(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)\Sigma_{\boldsymbol{\xi}_i}]/L(\boldsymbol{\lambda}_{i0}|\mathbf{y}_i)\right)\\ &= \left(\sum_{i=1}^{N}\sum_{j=1}^{n_i} g_{ij}\frac{f'_{ij}}{f_{ij}}\bigg|_{\boldsymbol{\lambda}_{i0}}, \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(\Sigma_{\boldsymbol{\xi}_i})_{jj}\frac{f''_{ij}}{f_{ij}}\bigg|_{\boldsymbol{\lambda}_{i0}}\right.\\ &\quad \left. + \sum_{i=1}^{N}\sum_{j<j'}(\Sigma_{\boldsymbol{\xi}_i})_{jj'}\frac{f'_{ij}f'_{ij'}}{f_{ij}f_{ij'}}\bigg|_{\boldsymbol{\lambda}_{i0}}\right) \qquad (5.8)\end{aligned}$$

where f is the probability density function of y, and f', f'' is the first$-$, second$-$order partial derivatives with respect to λ_{ij0} under H_0, the information matrix is:

$$\begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix}$$

where

$$\begin{aligned}
I_{11} &= E[(\partial l(\boldsymbol{\lambda},\alpha,\beta)/\partial\boldsymbol{\lambda})^2|H_0] = \sum_{i=1}^{N}\sum_{j=1}^{n_i} E[(f'_{ij}/f_{ij})^2|H_0], \\
I_{12} &= I_{21} = E[(\partial l(\boldsymbol{\lambda},\alpha,\beta)/\partial\boldsymbol{\lambda})(\partial l(\boldsymbol{\lambda},\alpha,\beta)/\partial\alpha)|H_0] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{n_i} E[(f'_{ij}/f_{ij})^2 g_{ij}|H_0], \\
I_{22} &= E[(\partial l(\boldsymbol{\lambda},\alpha,\beta)/\partial\alpha)^2|H_0] = \sum_{i=1}^{N}\sum_{j=1}^{n_i} E[(f'_{ij}/f_{ij})^2 g_{ij}^2|H_0], \\
I_{13} &= I_{31} = E[(\partial l(\boldsymbol{\lambda},\alpha,\beta)/\partial\boldsymbol{\lambda})(\partial l(\boldsymbol{\lambda},\alpha,\beta)/\partial\beta)|H_0] \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i} E\left.\frac{f''_{ij}f'_{ij}}{f_{ij}^2}(\Sigma_{\xi_i})_{jj}\right|_{H_0} \\
I_{23} &= I_{32} = E[(\partial l(\boldsymbol{\lambda},\alpha,\beta)/\partial\alpha)(\partial l(\boldsymbol{\lambda},\alpha,\beta)/\partial\beta)|H_0] \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i} E g_{ij}\left.\frac{f''_{ij}f'_{ij}}{f_{ij}^2}(\Sigma_{\xi_i})_{jj}\right|_{H_0} \\
I_{33} &= E[(\partial l(\boldsymbol{\lambda},\alpha,\beta)/\partial\beta)^2|H_0] = \frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{n_i} E\left.\left(\frac{f''_{ij}}{f_{ij}}(\Sigma_{\xi_i})_{jj}\right)^2\right|_{H_0} \\
&\quad + \sum_{i=1}^{N}\sum_{j<j'} E\left.\left((\Sigma_{\boldsymbol{\xi}_i})_{jj'}\frac{f'_{ij}f'_{ij'}}{f_{ij}f_{ij'}}\right)^2\right|_{H_0}
\end{aligned}$$

The probability density function of y_{ij} under H_0 is:

$$f_{ij} = p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}},$$

so that, $U(\alpha,\beta)$ becomes:

$$\left(\sum_{i=1}^{N}\sum_{j=1}^{n_i}(y_{ij}-p_{ij})g_{ij}, \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(\Sigma_{\boldsymbol{\xi}_i}\right)_{jj}\left[(y_{ij}-p_{ij})^2 - p_{ij}(1-p_{ij})\right] + \sum_{i=1}^{N}\sum_{j<j'}\left(\Sigma_{\boldsymbol{\xi}_i}\right)_{jj'}(y_{ij}-p_{ij})(y_{ij'}-p_{ij'})\right) \tag{5.9}$$

and the elements of Fisher information matrix are:

$$\begin{aligned}
I_{11} &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} p_{ij}(1-p_{ij}) \\
I_{12} &= I_{21} = \sum_{i=1}^{N}\sum_{j=1}^{n_i} p_{ij}(1-p_{ij})g_{ij} \\
I_{22} &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} p_{ij}(1-p_{ij})g_{ij}^2 \\
I_{13} &= I_{31} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i} p_{ij}(1-p_{ij})(1-2p_{ij})(\Sigma_{\xi_i})_{jj} \\
I_{23} &= I_{32} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i} p_{ij}(1-p_{ij})(1-2p_{ij})(\Sigma_{\xi_i})_{jj}g_{ij} \\
I_{33} &= \frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{n_i} p_{ij}(1-p_{ij})(2p_{ij}-1)^2(\Sigma_{\xi_i})_{jj}^2 \\
&\quad + \sum_{i=1}^{N}\sum_{j<j'} p_{ij}(1-p_{ij})p_{ij'}(1-p_{ij'})(\Sigma_{\xi_i})_{jj'}^2
\end{aligned}$$

Since p_{ij} is unknown, we replace it by its MLE $\hat{p_{ij}}$ under H_0,

$$\hat{p}_{ij} = e^{\hat{\mu}_0 + \mathbf{x}_{ij}\hat{\boldsymbol{\gamma}}_0}/(1 + e^{\hat{\mu}_0 + \mathbf{x}_{ij}\hat{\boldsymbol{\gamma}}_0})$$

where the values $(\hat{\mu_0}, \hat{\boldsymbol{\gamma}}_0)$ are the MLE of $(\mu_0, \boldsymbol{\gamma}_0)$ under H_o, Bartoo and Puni (1967) demonstrated that the score test statistic is:

$$U(\alpha, \beta) \left[\begin{pmatrix} I_{22} & I_{23} \\ I_{32} & I_{33} \end{pmatrix} - \begin{pmatrix} I_{21} \\ I_{31} \end{pmatrix} I_{11}^{-1} \begin{pmatrix} I_{12} & I_{13} \end{pmatrix} \right]^{-1} U^T(\alpha, \beta)$$

In particular, this score statistic does not follow a standard chi-square distribution because the true parameter values are on the boundary of the parameter space. The same as in chapter 3, for hypothesis test $H_0 : \beta = 0, \alpha = 0$ vs. $H_a : \beta > 0$ or/and $\alpha \neq 0$, the parameter β or α will reach the boundary parameter $\{0\}$ at H_A. As before, the asymptotic distribution of the score test is as following:

$$(\frac{5}{8} - \frac{\rho}{2\pi})\chi_0^2 + (\frac{3}{8} + \frac{\rho}{4\pi})\chi_1^2 + \frac{\rho}{4\pi}\chi_2^2 \tag{5.10}$$

with

$$\rho = \arccos \frac{I_{23}}{\sqrt{I_{22} I_{33}}} \tag{5.11}$$

5.3 Test of Association

For testing association alone, setting free $\beta^\bullet$, the model can be expressed as:

$$\begin{aligned} \text{logit } P(\mathbf{y}_{ij}=1) &= \mu^\bullet + \alpha^\bullet g_{ij} + \mathbf{x}_{ij}\boldsymbol{\gamma}^\bullet + \xi_{ij}^\bullet \\ &= \boldsymbol{\lambda}_{ij}^\bullet + \alpha^\bullet g_{ij} \end{aligned} \tag{5.12}$$

Under $\alpha = 0$, the model can be reduced as,

$$\text{logit } P(\mathbf{y}_{ij}=1) = \mu_0^\bullet + \mathbf{x}_{ij}\boldsymbol{\gamma}_0^\bullet + \xi_{ij0}^\bullet = \boldsymbol{\lambda}_{ij0}^\bullet \tag{5.13}$$

assuming $\{y_{ij}\}$ are independent and $\xi_{ij0}^\bullet \sim N(0, \beta_0^{\bullet 2})$. The parameters of fixed effects and random effect jointly in these models can be obtained through the procedure described in the previous section 5.2.

Let $\boldsymbol{\lambda}^\bullet = \{\boldsymbol{\lambda}_i^\bullet\}, \boldsymbol{\lambda}_i^\bullet = \{\lambda_{ij}^\bullet\}$, and $l(\boldsymbol{\lambda}^\bullet, \alpha^\bullet) = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \ln f_{ij}$ denotes the log-likelihood function of $(\boldsymbol{\lambda}^\bullet, \alpha^\bullet)$, the efficient score for $\alpha^\bullet(\alpha^\bullet = 0)$ is:

$$U(\alpha^\bullet) = \left.\frac{\partial l}{\partial \alpha^\bullet}\right|_{\alpha^\bullet=0}. \tag{5.14}$$

The Taylor series expansion for f_{ij} around $\lambda_{ij0}^\bullet$ is:

$$f_{ij} = f(y_{ij}|\lambda_{ij0}^\bullet) + \alpha^\bullet\, g_{ij} f_{ij}^{'}|\lambda_{ij0}^\bullet + \frac{1}{2}\alpha^{\bullet 2} g_{ij}^2 f_{ij}^{''}|\lambda_{ij0}^\bullet + r_{ij}^\bullet$$

where $f(y_{ij}|\lambda^{\bullet}_{ij0})$ is probability density function of y_{ij} under H_0, similarly

$$\begin{aligned} f^{'}_{ij}|\lambda^{\bullet}_{ij0} &= \left.\frac{\partial f(y_{ij}|\lambda^{\bullet}_{ij0})}{\partial \lambda^{\bullet}_{ij0}}\right|_{\lambda^{\bullet}_{ij0}} \\ f^{''}_{ij}|\lambda^{\bullet}_{ij0} &= \left.\frac{\partial^2 f(y_{ij}|\lambda^{\bullet}_{ij0})}{\partial \lambda^{\bullet 2}_{ij0}}\right|_{\lambda^{\bullet}_{ij0}} \end{aligned}$$

$r^{\bullet}_{ij}$ is the remainder term which equals zero as $\alpha^{\bullet} \downarrow 0$.

If $\mu^{\bullet}_0, \boldsymbol{\gamma}^{\bullet}_0, \beta^{\bullet}_0$ are known, $U^2(\alpha^{\bullet})/I_{22}$ will be a χ^2_1 distribution asymptotically when H_0 is true. If $\boldsymbol{\lambda}^{\bullet}$ is replaced by its estimation $\hat{\boldsymbol{\lambda}}^{\bullet}$ under H_0, then Bartoo and Puni (1967) demonstrated the test statistic:

$$U^2(\alpha^{\bullet})/(I_{22} - I^2_{12}/I_{11}) \tag{5.15}$$

which is approximately distributed as χ^2_1 distribution under null hypothesis, where I_{ij} is component in the information matrix with respect to $(\boldsymbol{\lambda}^{\bullet}, \alpha^{\bullet})$.

Under the model (5.13) and score function (5.14),

$$U(\alpha^{\bullet}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}(y_{ij} - p_{ij})g_{ij} \tag{5.16}$$

and the elements of information matrix are:

$$\begin{aligned} I_{11} &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} p_{ij}(1-p_{ij}) \\ I_{12} &= I_{21} = \sum_{i=1}^{N}\sum_{j=1}^{n_i} p_{ij}(1-p_{ij})g_{ij} \\ I_{22} &= \sum_{i=1}^{N}\sum_{j=1}^{n_i} p_{ij}(1-p_{ij})g_{ij}^2 \end{aligned}$$

Replacing the unknown p_{ij} by its MLE $\hat{p_{ij}}$ under H_0,

$$\hat{p}_{ij} = e^{\hat{\mu}_0^{\bullet}+\mathbf{x}_{ij}\hat{\gamma}_0^{\bullet}+\hat{\xi}_{ij0}^{\bullet}}/(1+e^{\hat{\mu}_0^{\bullet}+\mathbf{x}_{ij}\hat{\gamma}_0^{\bullet}+\hat{\xi}_{ij0}^{\bullet}}),$$

the values $(\hat{\mu}_0^{\bullet}, \hat{\boldsymbol{\gamma}}_0^{\bullet}, \hat{\xi}_{ij0}^{\bullet})$ are obtained by model (5.13).

5.4 Test of Linkage

For linkage test, setting free $\alpha^{\star}$, the model is

$$\begin{aligned} \text{logit } P(\mathbf{y}_i = 1) &= \mu^{\star}\mathbf{1} + \mathbf{x}_i\boldsymbol{\gamma}^{\star} + \alpha^{\star}\mathbf{g}_i + \sqrt{\beta^{\star}}\boldsymbol{\xi}_i \\ &= \boldsymbol{\lambda}_i^{\star} + \sqrt{\beta^{\star}}\boldsymbol{\xi}_i \end{aligned} \tag{5.17}$$

and the likelihood function L_i is given by:

$$L_i = \int \cdots \int L(\boldsymbol{\lambda}_i^{\star}, \beta^{\star}|\mathbf{y}_i) dG(\boldsymbol{\xi}_i), \tag{5.18}$$

where $\beta^\star \geq 0$ behaves as a scale parameter, $\boldsymbol{\lambda}_i^\star = \mu^\star \mathbf{1} + \mathbf{x}_i\boldsymbol{\gamma}^\star + \alpha^\star \mathbf{g}_i$, and $\boldsymbol{\xi}_i = \{\xi_{ij}\}, j = 1, ..., n_i$ have multivariate distribution function G with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\xi_i}$.

Let $\boldsymbol{\lambda}^\star = \{\boldsymbol{\lambda}_i^\star\}$, $l(\boldsymbol{\lambda}^\star, \beta^\star) = \sum_{i=1}^{N} \ln L_i$ denotes the log-likelihood function of $(\boldsymbol{\lambda}^\star, \beta^\star)$, the efficient score for $\beta^\star (\beta^\star = 0)$ is denoted by:

$$U(\beta^\star) = \left.\frac{\partial l}{\partial \beta^\star}\right|_{\beta^\star=0} = \sum_{i=1}^{N} \left[\frac{1}{L_i} \times \frac{\partial L_i}{\partial \beta^\star}\right]\Bigg|_{\beta^\star=0}. \tag{5.19}$$

The Taylor series expansion for likelihood function $L(\boldsymbol{\lambda}_i^\star, \beta^\star|\mathbf{y}_i)$ around $\boldsymbol{\lambda}_{i0}^\star$ is:

$$\begin{aligned} L(\boldsymbol{\lambda}_i^\star, \beta^\star|\mathbf{y}_i) &= L(\boldsymbol{\lambda}_{i0}^\star|\mathbf{y}_i) + \sqrt{\beta^\star} L^{'}(\boldsymbol{\lambda}_{i0}^\star|\mathbf{y}_i)\boldsymbol{\xi}_i \\ &\quad + \frac{1}{2}\beta^\star \boldsymbol{\xi}_i^T L^{''}(\boldsymbol{\lambda}_{i0}^\star|\mathbf{y}_i)\boldsymbol{\xi}_i + \mathbf{r}_i^\star \end{aligned} \tag{5.20}$$

where $\boldsymbol{\lambda}_{i0}^\star = \mu_0^\star \mathbf{1} + \mathbf{x}_i\boldsymbol{\gamma}_0^\star + \alpha_0^\star \mathbf{g}_i$, and $L(\boldsymbol{\lambda}_{i0}^\star|\mathbf{y}_i)$ is the likelihood function under H_0 in the ith family.

$$L^{'}(\boldsymbol{\lambda}_{i0}^\star|\mathbf{y}_i) = \left(\frac{\partial L(\boldsymbol{\lambda}_{i0}^\star|\mathbf{y}_i)}{\partial \lambda_{i10}^\star}, \frac{\partial L(\boldsymbol{\lambda}_{i0}^\star|\mathbf{y}_i)}{\partial \lambda_{i20}^\star}, ..., \frac{\partial L(\boldsymbol{\lambda}_{i0}^\star|\mathbf{y}_i)}{\partial \lambda_{in_i0}^\star}\right)\Bigg|_{\boldsymbol{\lambda}_{i0}^\star}$$

and

$$L^{''}(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i) = \left(\begin{array}{cccc} \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i)}{\partial \lambda_{i10}^{\star 2}} & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i)}{\partial \lambda_{i10}^{\star} \partial \lambda_{i20}^{\star}} & \cdots & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i)}{\partial \lambda_{i10}^{\star} \partial \lambda_{in_i0}^{\star}} \\ \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i)}{\partial \lambda_{i20}^{\star} \partial \lambda_{i1}^{\star}} & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i)}{\partial \lambda_{i20}^{\star 2}} & \cdots & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i)}{\partial \lambda_{i20}^{\star} \partial \lambda_{in_i0}^{\star}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i)}{\partial \lambda_{in_i0}^{\star} \partial \lambda_{i10}^{\star}} & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i)}{\partial \lambda_{in_i0}^{\star} \partial \lambda_{i20}^{\star}} & \cdots & \frac{\partial^2 L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i)}{\partial \lambda_{in_i0}^{\star 2}} \end{array} \right) \Bigg|_{\boldsymbol{\lambda}_{i0}^{\star}}$$

are the first two partial derivatives of $L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i)$ with respect to $\boldsymbol{\lambda}_{i0}^{\star} = \{\lambda_{ij0}^{\star}\}$, $j = 1, 2, \cdots, n_i$, and $\mathbf{r}_i^{\star}$ is the remainder term. Since

$$\int \cdots \int L^{'}(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i) \boldsymbol{\xi}_i dG(\boldsymbol{\xi}_i) = 0$$

and

$$\int \cdots \int \boldsymbol{\xi}_i^T L^{''}(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i) \boldsymbol{\xi}_i dG(\boldsymbol{\xi}_i) = \text{tr}[L^{''}(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i) \Sigma_{\boldsymbol{\xi}_i}]$$

Therefore, L_i from (5.17) can be represented as

$$L_i = L(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i) + \frac{1}{2}\beta^{\star} \text{tr}[L^{''}(\boldsymbol{\lambda}_{i0}^{\star}|\mathbf{y}_i) \Sigma_{\boldsymbol{\xi}_i}] + \mathbf{R}_i \tag{5.21}$$

Following the result in Zelterman and Chen (1988), $\mathbf{R}_i = \int \cdots \int \mathbf{r}_i dG(\boldsymbol{\xi}_i) = 0$ as

$\beta^\star \downarrow 0$. The efficient score $U(\beta^\star)$ of (5.19) then becomes:

$$U(\beta^\star) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(\Sigma_{\boldsymbol{\xi}_i})_{jj}\frac{f''_{ij}}{f_{ij}}\bigg|_{\boldsymbol{\lambda}^\star_{i0}} + \sum_{i=1}^{N}\sum_{j<j'}(\Sigma_{\boldsymbol{\xi}_i})_{jj'}\frac{f'_{ij}f'_{ij'}}{f_{ij}f_{ij'}}\bigg|_{\boldsymbol{\lambda}^\star_{i0}} \tag{5.22}$$

where f_{ij} is the probability density function of y_{ij}, and f'_{ij}, f''_{ij} are the first two partial derivatives with respect to $\lambda^\star_{ij0}$ under H_0, $(\Sigma_{\xi_i})_{jj}$ is the jth diagonal value of covariance matrix Σ_{ξ_i} and $(\Sigma_{\xi_i})_{jj'}$ is the jth row and the j'th columns non-diagonal value of covariance matrix Σ_{ξ_i} in the ith family.

Under H_0, the information matrix with respect to $(\boldsymbol{\lambda}^\star, \alpha^\star)$is given by:

$$\begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$$

where

$$\begin{aligned}
I_{11} &= E[(\partial l(\boldsymbol{\lambda}^\star,\beta^\star)/\partial\boldsymbol{\lambda}^\star)^2|H_0] = \sum_{i=1}^{N}\sum_{j=1}^{n_i}E[(f'_{ij}/f_{ij})^2|H_0],\\
I_{12} &= I_{21} = E[(\partial l(\boldsymbol{\lambda}^\star,\beta^\star)/\partial\boldsymbol{\lambda}^\star)(\partial l(\boldsymbol{\lambda}^\star,\beta^\star)/\partial\beta^\star)|H_0]\\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}E\,\frac{f''_{ij}f'_{ij}}{f^2_{ij}}(\Sigma_{\xi_i})_{jj}\bigg|_{H_0}\\
I_{22} &= E[(\partial l(\boldsymbol{\lambda}^\star,\beta^\star)/\partial\beta^\star)^2|H_0]\\
&= \frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{n_i}E\left(\frac{f''_{ij}}{f_{ij}}(\Sigma_{\xi_i})_{jj}\right)^2\bigg|_{H_0} + \sum_{i=1}^{N}\sum_{j<j'}E\left((\Sigma_{\boldsymbol{\xi}_i})_{jj'}\frac{f'_{ij}f'_{ij'}}{f_{ij}f_{ij'}}\right)^2\bigg|_{H_0}
\end{aligned}$$

If $\boldsymbol{\lambda}^\star$ is known, $U^2(\beta^\star)/I_{22}$ will be a χ^2_1 distribution asymptotically when H_0 is true. If

$\boldsymbol{\lambda}^\star$ is replaced by its MLE $\hat{\boldsymbol{\lambda}}^\star$ under H_0, the test statistic

$$U^2(\beta^\star)/(I_{22} - I_{12}^2/I_{11}) \tag{5.23}$$

has a 50:50 mixture of χ_0^2 and χ_1^2 distribution under null hypothesis.

If the probability density function of y_{ij} under H_0 is:

$$f_{ij} = p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}},$$

and p_{ij} is replaced by its MLE $\hat{p_{ij}}$ under H_0, where $\hat{p}_{ij} = e^{\hat{\mu}_0^\star + \mathbf{x}_{ij}\hat{\boldsymbol{\gamma}}_0^\star + \hat{\alpha}_0^\star g_{ij}}/(1+e^{\hat{\mu}_0^\star + \mathbf{x}_{ij}\hat{\boldsymbol{\gamma}}_0^\star + \hat{\alpha}_0^\star g_{ij}})$, the values $(\hat{\mu_0}^\star, \hat{\boldsymbol{\gamma}}_0^\star, \hat{\alpha}_0^\star)$ are the MLE of $(\mu_0^\star, \boldsymbol{\gamma}_0^\star, \alpha_0^\star)$ under $\beta^\star = 0$. The efficient score $U(\beta^\star)$ becomes:

$$\begin{aligned} U(\beta^\star) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(\Sigma_{\boldsymbol{\xi}_i})_{jj}\left[(y_{ij}-\hat{p}_{ij})^2 - \hat{p}_{ij}(1-\hat{p}_{ij})\right] \\ &\quad + \sum_{i=1}^{N}\sum_{j<j'}(\Sigma_{\boldsymbol{\xi}_i})_{jj'}(y_{ij}-\hat{p}_{ij})(y_{ij'}-\hat{p}_{ij'}) \end{aligned} \tag{5.24}$$

and

$$\begin{aligned} I_{11} &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\hat{p}_{ij}(1-\hat{p}_{ij}) \\ I_{12} &= I_{21} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\hat{p}_{ij}(1-\hat{p}_{ij})(1-2\hat{p}_{ij})(\Sigma_{\xi_i})_{jj} \\ I_{22} &= \frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\hat{p}_{ij}(1-\hat{p}_{ij})(2\hat{p}_{ij}-1)^2(\Sigma_{\xi_i})_{jj}^2 \\ &\quad + \sum_{i=1}^{N}\sum_{j<j'}\hat{p}_{ij}(1-\hat{p}_{ij})\hat{p}_{ij'}(1-\hat{p}_{ij'})(\Sigma_{\xi_i})_{jj'}^2 \end{aligned}$$

Chapter 6

Simulation Study

6.1 Introduction

We have conducted extensive simulation studies to assess the performance of the proposed joint linkage and association test, linkage test, and association test for quantitative traits and binary traits. Phenotypes are simulated on 73 individuals using 550K SNP genotype data from five pedigrees. A random genotype quality, SNP rs3859167 on chromosome 16, is selected to be the single marker locus explaining our simulation studies. Simulation is done in three steps: First, 73 individual genotypes are selected for SNP rs3859167 on chromosome 16 of five pedigrees; Second, 73 random traits are obtained, which may be influenced by marker genotype and by environment (age); Third, proposed methods are used to calculate the statistic value. Each simulation of 73 observations is generated as follows: Each observation consists of genotypes, a covariates (age), and an outcome y that is the phenotype of the individual.

6.2 Simulation Results for Quantitative Traits

First, to explore the selected sample properties of the proposed likelihood ratio test for quantitative traits, we run a series of simulations using multivariate normal data with several fixed effects and a random effect. Data are generated from the multinormal mixed model (3.1) with the values of the fixed parameters at $\mu = 0, \gamma = 0.02$. We perform a simulation study based on 10,000 replicates of data sets and set $\alpha = 0, \beta^2 = 0$, and $\sigma^2 = 1$ to examine the level of significance of the proposed joint test. Table (6.1) presents the levels of significance of the tests for the multinormal mixed model. We note from the table that the level of significance of the joint test is generally much closer to the nominal 0.05, 0.01, and 0.001.

Nominal level	.05	.01	.001
Type I error	.0455	.0091	.0009

Table 6.1: Empirical level of significance for a multinormal mixed model of joint test.

The power of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is actually false. In general, the power is a function of the possible distributions which is determined by parameters under the alternative hypothesis. To investigate the power of the proposed likelihood ratio tests, we choose $\alpha = -0.8, -0.4, 0, 0.4, 0.8, \beta^2 = 0.2, 0.4, 0.6, 0.8$ at $\sigma^2 = 1$, and use 1,000 simulation replications for each simulation configuration, to find the P-value of the joint linkage and association test, linkage test, and association test. Table (6.2) presents the empirical powers of the likelihood ratio tests for the joint test, linkage test, and association test. It is clear from the table that the proposed joint test performs slightly worse than

association test, sometimes performs significantly better than both tests. When the value of $|\alpha|$ increases, likelihood ratio method tends to provide increased powers of the joint test and association test significantly, remained fairly constant powers of linkage test. When the value of β^2 increases, the likelihood ratio method tends to provide results obvious increased powers of linkage test and joint test except larger $|\alpha|$ which slightly decrease power, and moderately decreases the power of association test except smaller $|\alpha|$ which slightly increase the power.

α	β^2	$Test$		
		Joint	Association	Linkage
-.8	.2	.984	.946	.016
	.4	.981	.914	.106
	.6	.954	.922	.203
	.8	.956	.898	.280
-.4	.2	.704	.530	.015
	.4	.652	.543	.089
	.6	.676	.535	.216
	.8	.684	.486	.283
0	.2	.008	0	.022
	.4	.086	0	.114
	.6	.161	0	.205
	.8	.246	0	.298
.4	.2	.692	.566	.010
	.4	.679	.540	.110
	.6	.690	.506	.196
	.8	.679	.505	.288
.8	.2	.987	.936	.024
	.4	.973	.934	.096
	.6	.961	.917	.220
	.8	.958	.896	.283

Table 6.2: Empirical power for a multinormal mixed model with $\sigma^2 = 1$

6.3 Simulation Results of Binary Traits with Multivariate Normal Random Effects

We perform a simulation study for binary traits with multivariate normal random effects, where the value of the logistic regression parameter μ is fixed at - 1.5, and γ is fixed at 0.02. The value $\alpha = 0, \varsigma = 0$ are used to examine the level of significance of the proposed joint test. For each simulation configuration, we illustrate the simulation study based on 1,000 replicates of data sets. From Table (6.3) we can see that the 0.01 level of significance of the joint test is slightly larger than the nominal 0.01 level of significance.

Nominal level	.01		
test	LRT	Wald	score
Type I error	0.021	0.046	0.017

Table 6.3: Empirical level of significance for a binary mixed model of joint test with multivariate normal random effects

To investigate the powers of proposed joint, linkage, and association tests for binary mixed models with multivariate normal assumption of random variables, we choose $\alpha = 0.35, 0.4, 0.45,\ \varsigma = 0.1, 0.15, 0.2$, and use 1,000 simulation replications for each simulation configuration to find the P-value of the joint, linkage, and association studies with LRT, Wald and score tests. Table (6.4) presents the empirical powers of the joint, linkage, and association studies when $\mu = 2$ and $\gamma = 0.02$. It is clear from the table (6.4) that the proposed joint test is generally more powerful than the linkage test or the association alone for most of cases, the Wald test is much more powerful than LRT and score tests, and LRT test is slightly powerful than score test. When the value of α

increases, three tests tend to provide increased power for the joint and association test, but the power of the linkage test tends no change. When the value of ς increases, three tests tend to provide increased powers of the joint test and Wald test tends to provide increased powers of the linkage test. The Wald test shows no powers of association while α is small , and no powers of linkage score test.

α	ς	Joint			Association			Linkage		
		LRT	Wald	score	LRT	Wald	score	LRT	Wald	score
.35	.1	.364	.918	.363	.297	0	.253	.038	.506	0
	.15	.363	.992	.358	.306	0	.286	.058	.949	0
	.2	.413	1	.387	.317	0	.314	.076	1	0
.4	.1	.415	.966	.368	.396	.013	.327	.037	.553	0
	.15	.424	1	.373	.408	.019	.360	.078	.923	0
	.2	.482	1	.458	.466	.020	.380	.1	1	0
.45	.1	.433	.977	.377	.400	.983	.363	.032	.494	0
	.15	.452	1	.378	.424	.984	.369	.082	.936	0
	.2	.522	1	.514	.498	.992	.433	.1	1	0

Table 6.4: Empirical power of tests for binary mixed models with multivariate normal random effects

6.4 Simulation Results of Score Tests for Binary Traits with Multivariate Random Effects - Distribution Unknown

A simulation study based on the score tests for binary traits with multivariate random effects that distribution unknown, where the value of the logistic regression parameters μ is fixed at -3 and γ is fixed at 0.02. The value $\alpha = 0,\ \beta = 0$ are used to examine the level of significance of the proposed joint score test. For each simulation configuration, a simulation study based on 10,000 replicates of data sets was performed. Table (6.5)

presents the levels of significance of the joint score test based on the mixture of chi-square distributions. Here we observe that the proposed joint score test provides level of significance that is generally larger than the nominal level at 0.05, at 0.01, and at 0.001.

Nominal level	.05	.01	.001
Type I error	.0564	.0200	.0079

Table 6.5: Empirical level of significance for binary mixed model of joint score test with multivariate random effects - distribution unknown

To investigate the power of proposed joint, linkage, and association score tests, we consider $\alpha = 0.2, 0.4, 0.6, 0.8,\ \beta = 0.2, 0.4, 0.6, 0.8$, and use 1,000 simulation replications for each simulation configuration to find the P-value of the score tests. Table (6.6) presents the empirical powers of the joint, linkage, and association score tests when $\mu = -3$ and $\mu = -5$. It is clear from table (6.6) that the proposed joint score test is generally more powerful than the linkage score test or the association score test alone. When the value of α increases, the power of the joint score test and association score test increase, and the power for the linkage test tends to slightly decrease. When the value of β increases, the power of the joint test and the linkage test increase, but the power of association seems no change.

6.5 Overall Simulation Results

In the generalized mixed model, the proposed joint test is a simple alternative to compute approximate P-values based on a mixture of chi-square distributions. The overall results from the simulation study demonstrate that type I errors of proposed joint LRTs

α	β	μ = -3			μ = -5		
		Joint	Association	Linkage	Joint	Association	Linkage
.2	.2	.394	.142	.129	.350	.117	.147
	.4	.463	.140	.192	.416	.109	.180
	.6	.562	.150	.235	.498	.106	.278
	.8	.604	.118	.298	.615	.095	.378
.4	.2	.632	.382	.115	.461	.216	.109
	.4	.709	.362	.146	.545	.213	.181
	.6	.758	.353	.221	.628	.251	.271
	.8	.790	.372	.268	.722	.232	.371
.6	.2	.884	.746	.112	.623	.385	.124
	.4	.883	.692	.142	.701	.398	.190
	.6	.930	.695	.201	.749	.392	.275
	.8	.938	.660	.260	.836	.434	.380
.8	.2	.978	.920	.106	.773	.586	.117
	.4	.969	.901	.127	.840	.603	.196
	.6	.980	.898	.164	.882	.636	.273
	.8	.984	.887	.213	.928	.619	.368

Table 6.6: Empirical power of score tests for binary mixed models

for quantitative traits have generally correct level at nominal 0.05, 0.01, and 0.001 significance respectively. However, type I errors of the proposed LRT, Wald and score tests for the binary traits with multivariate normal assumption of random variables and score test for the binary traits with dependent unknown distribution random variables are generally larger than nominal levels. The combined test is a robust alternative; it does not substantially under-perform relative to either linkage or association test, and sometimes significantly out-perform both tests.

The joint tests can keep the power advantage even when the data has either no linkage or no association evidence. For the cases where the association parameter increases, linkage tests may provide slightly decreased power or remain fairly constant, while if the linkage parameter increases, association tests may lead moderately change.

However, for either cases, the proposed joint tests result significantly increased power.

In a simulation study, we compare the Wald joint test to the LRT and score joint tests for the binary data with multivariate normal assumption of random variables. It is confirmed that the Wald joint test is too liberal whereas the LRT and score joint tests are too conservative for 'large' α and ς, and the powers of the score test and the LRT are non-significantly different. Our simulation studies also have confirmed that Wald joint test has dramatically inflated type I error. For small linkage and association parameters, the robust Wald test does not perform well.

The LR, Wald, and score tests require different models to be estimated. More specifically, Wald test only requires the unrestricted model, score test needs the restricted model only, whereas LRT requires both the restricted and unrestricted models to be estimated. Several authors have identified problems with the use of the Wald statistic. Menard (1995) warns that for large coefficients, standard error is inflated, lowering the Wald statistic (chi-square) value. Agresti (1996) states that the likelihood-ratio test is more reliable for small sample sizes than the Wald test. Comparing the proposed joint LRT and joint score test for binary traits, the power of the joint score test is just slightly smaller than the power of the LRT with the same parameter values. In addition, the joint score test only requires estimation of the fixed effects regression coefficients under the null hypothesis such that the computation of joint score test is much faster than joint LRT (joint score test in chapter 5 with 1,000 simulation replications took 12 minutes, joint LRT in chapter 4 with 1,000 simulation replications took 8 hours of R 3.0.1 program). For large data sets, such as genome-wide SNPs, the joint

score statistic is recommended. If the exact population distribution is unknown or the exact likelihood is unknown, the joint score statistic has to be used. As in our data set with 600470 SNP's markers and unknown distribution of Familial Pulmonary Fibrosis, the joint score test should be used.

Chapter 7

Testing the Model with Familial Pulmonary Fibrosis

7.1 Introduction

Many common diseases in humans are caused by complex interactions among multiple genes with environment. A genetic predisposition may make a person vulnerable to developing a disease, while an environmental exposure may actually cause a disease to manifest. For example, pulmonary fibrosis (PF) is a complicated illness that the most frequent cases are related to sarcoidosis, fibrosis associated with certain occupational diseases, and older age, male sex, and history of cigarette smoking are important risk factors for the development of disease. Knowing that age, cigarette smoking are the risk factors suggest environmental factors may accentuate genetic risk and that gene-environment interactions may be important in PF disease pathogenesis (Canadian lung association: http://www.lung.ca/). Although we know that both environmental factors

such as age and lifestyle factors add tremendously to the uncertainty of developing a disease, it is difficult to measure and evaluate their overall effect on a disease process. Here, we analyze mainly a person's genetic predisposition and an environmental factor (age).

Identifying common ancestors can be very important. In one extended family members, association due to linkage disequilibrium should exist between ancestral disease susceptibility genes and closely linked markers. Investigation of genetic diseases has been facilitated by large families, close family ties, and modest out migration. A joint association and linkage tests that use marker and phenotype data from a number of families should have greater power in detecting a disease susceptibility locus.

The island of Newfoundland is a sparsely populated region of Canada in which 50% of the population of 560,000 reside in small coastal communities. The colonization of the island occurred primarily by natural increase from northern European settlers of predominantly English and Irish extraction who arrived before 1835. Most founders originated from the West Country of England and from southeast Ireland. Mating segregation between Irish Catholics and English Protestants, low immigration, and geographical isolation of communities have resulted in genetic isolation of the population. The Newfoundland population can be considered to have relatively homogenous origins and consist of multiple genetically simplified isolates. Its geography, settlement, and socioeconomic development have produced a population group which is ideal for study of genetic diseases (Young et al. 1999; Parfrey et al. 2002). It is known that population genetic isolation results in an increase in the possibility

of allele frequencies being affected by founder effects and a high coefficient of kinship. Familial pulmonary fibrosis is characterized by the presence of two or more primary biological family members (parent, child, or sibling) with the diagnosis of Idiopathic Pulmonary Fibrosis (IPF) or any other form of Idiopathic Interstitial Pneumonia (IIP). IPF is a late-onset disease characterized by inflammation and scarring of the lung parenchyma. A 10-15% of IPF is attributed to genetic causes.

FPF is considered a complex disease. This means a combination of genetic predisposition and environmental triggers contribute to an individual developing pulmonary fibrosis. FPF appears to transmit through families in an autosomal dominant fashion with reduced penetrance. An autosomal dominant inheritance pattern implies that if an affected mutation carrier has offspring, an average of 50% of the offspring of the mutation-carrier will carry the disease-causing variant. In a disease that has reduced penetrance, such as FPF, there are individuals who may carry the disease-causing genetic variant but will not present with the disease phenotype during their lifetime. Furthermore, phenocopies can also be present in families with FPF. Phenocopies are defined as the same phenotype being displayed by different individuals due to differing genetic and/or environmental causes. Disease heterogeneity in FPF within individual families has been found recently. Genetic heterogeneity is a phenomenon in which a single phenotype or genetic disorder may be caused by any one of a multiple number of alleles or non-allele alterations, like insertions or deletions. Genetic heterogeneity can be classified as either “allelic” or “locus”. Allelic heterogeneity means that different mutations within a single gene locus (forming multiple alleles of that gene) cause

the same phenotypic expression. Locus heterogeneity means that variation in possibly unrelated gene loci can cause the disorder.

FPF is a rare disease with a strong genetic component that has a high prevalence in the Newfoundland population compared to other populations, the possibility of one or few founder mutation cannot be ruled out as being a cause. Alternatively, without the founder effect hypothesis, one must conclude that it is the heterogeneous nature of the disease the cause of the high prevalence of FPF in Newfoundland. Although the complete pathogenesis of FPF is still not completely understood, the five genes (TERT, TERC, ABCA3, SFTPC and SFTPA2) known to carry variants causing familial pulmonary fibrosis (FPF) have been screened in our NL cohort, and Dr. Michael Woods' laboratory showed the liability class as table (7.1)

Table 7.1: Liability class

age	normal homozygous	disease heterozygous	disease homozygous
< 40	0.000006	0.10	0.10
40-49	0.000012	0.13	0.13
50-59	0.000034	0.33	0.33
60-69	0.000082	0.50	0.50
70-79	0.000164	0.73	0.73
> 80	0.0002	1.00	1.00

To illustrate our method, Dr. Michael Woods' laboratory, at Memorial University collected blood or tissue samples of families with clinically confirmed FPF and extracted genomic DNA from these samples. A total of five Familial Pulmonary Fibrosis (FPF) families which have 73 individual genotypes with 600470 SNP's markers are available for study. There are two distinct statistical goals: (1) Do the SNP data provide evidence that the genetic variation contributes to FPF? (2) What is the most

likely location of the disease variant(s)? For these goals, we have discrete outcomes (phenotypes) which have been clinically verified, and genotypes and covariate (age).

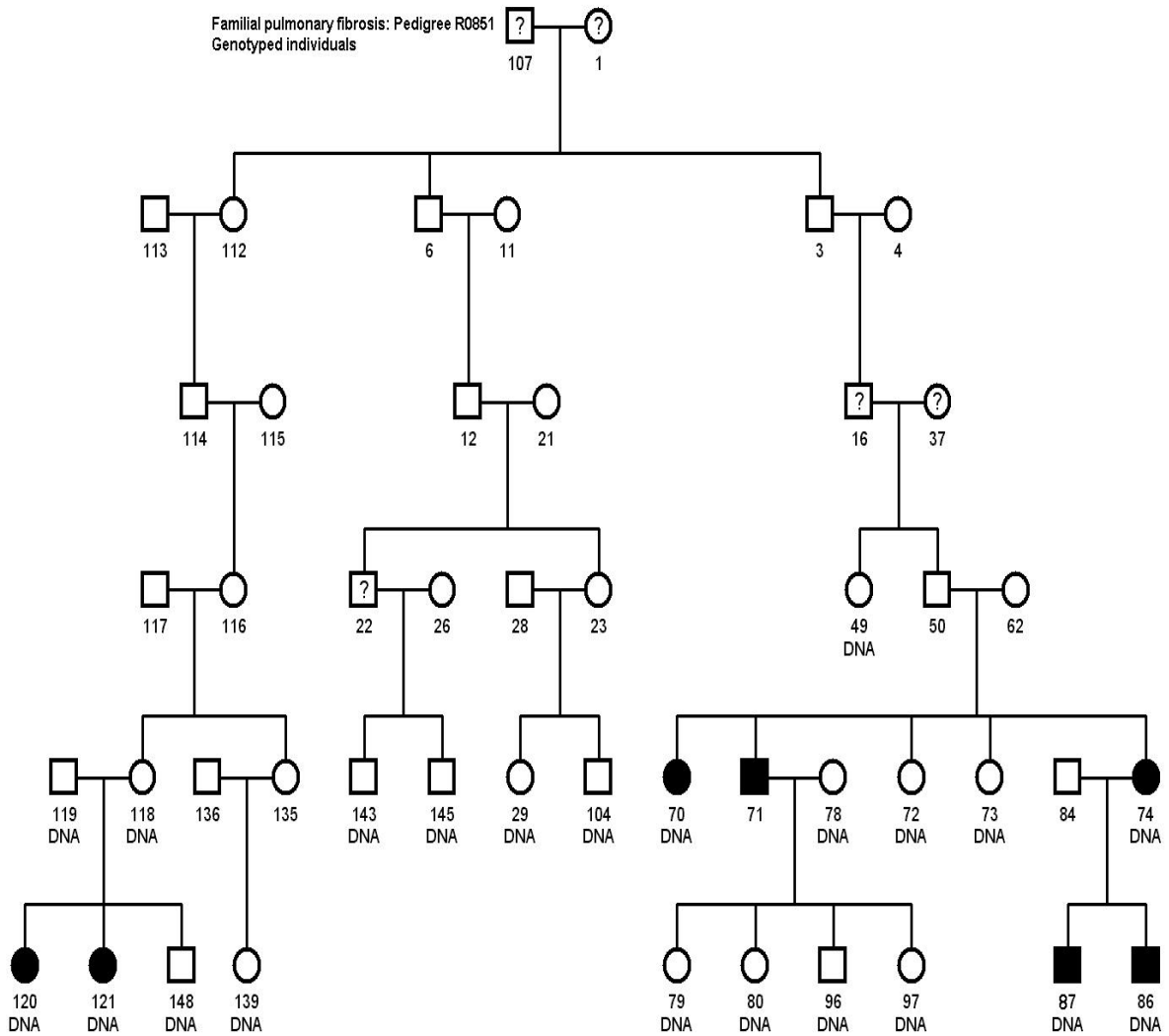


Figure 7.1: Pedigree structure, phenotype for the FPF pedigree R0851

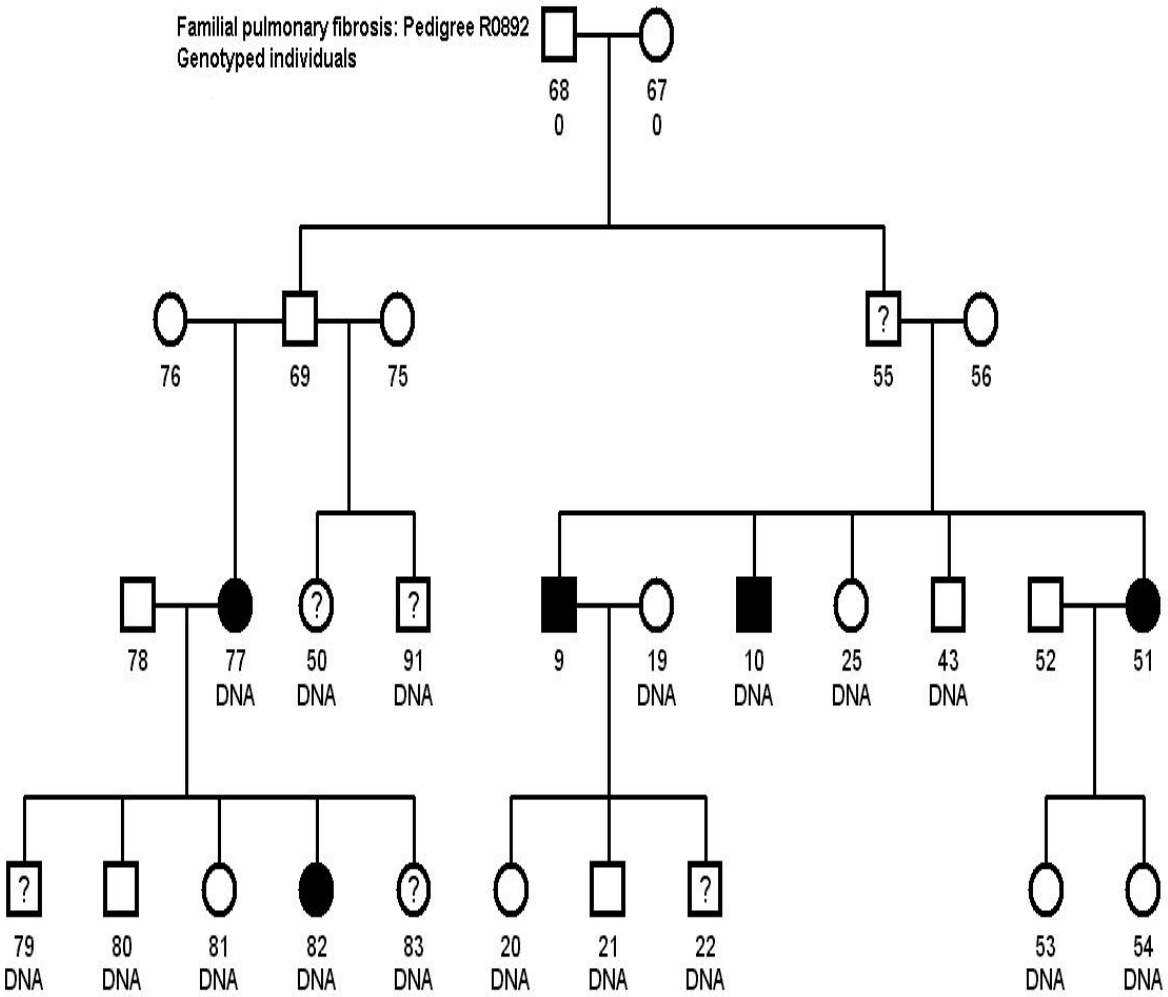


Figure 7.2: Pedigree structure, phenotype for the FPF pedigree R0892

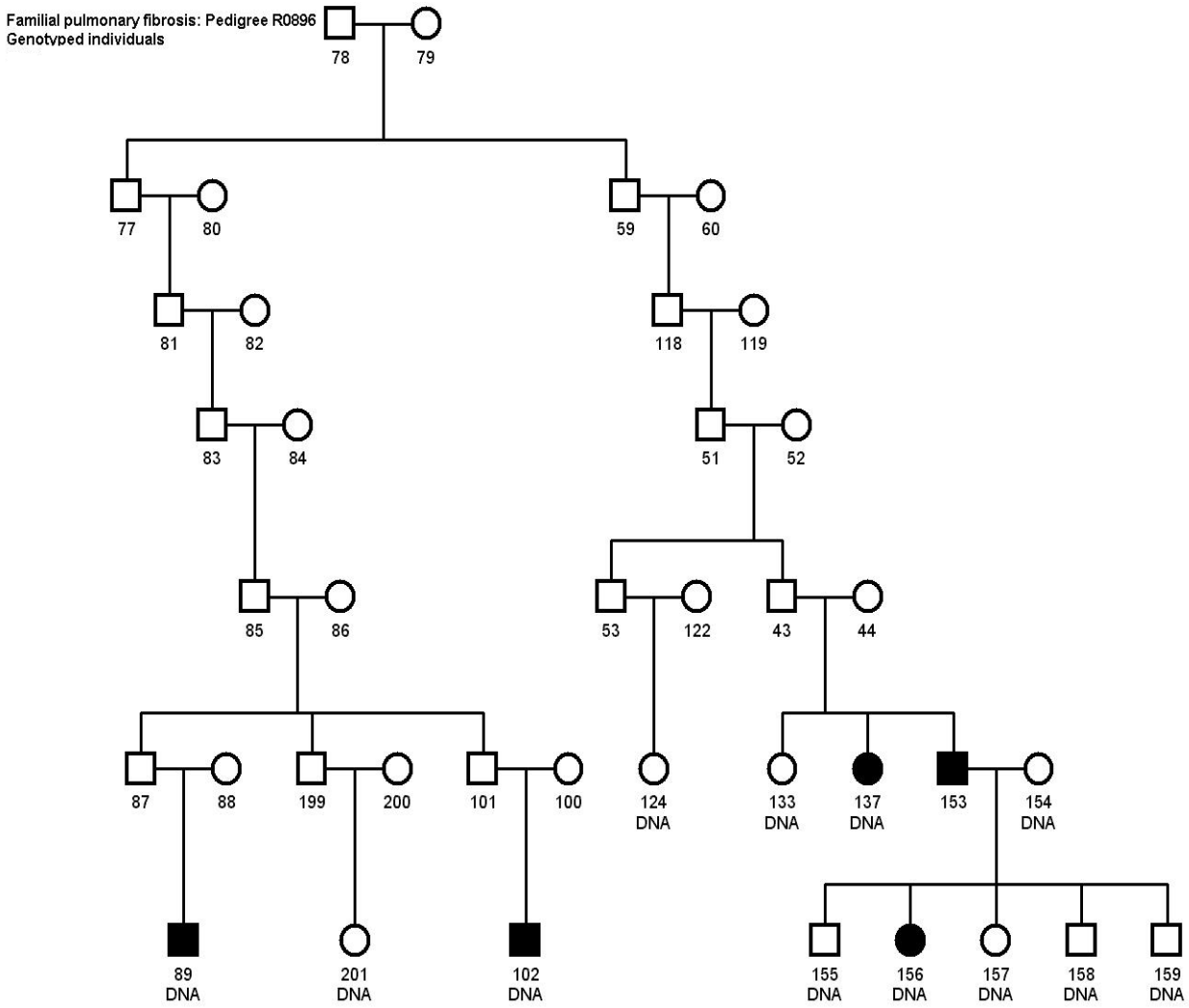


Figure 7.3: Pedigree structure, phenotype for the FPF pedigree R0896

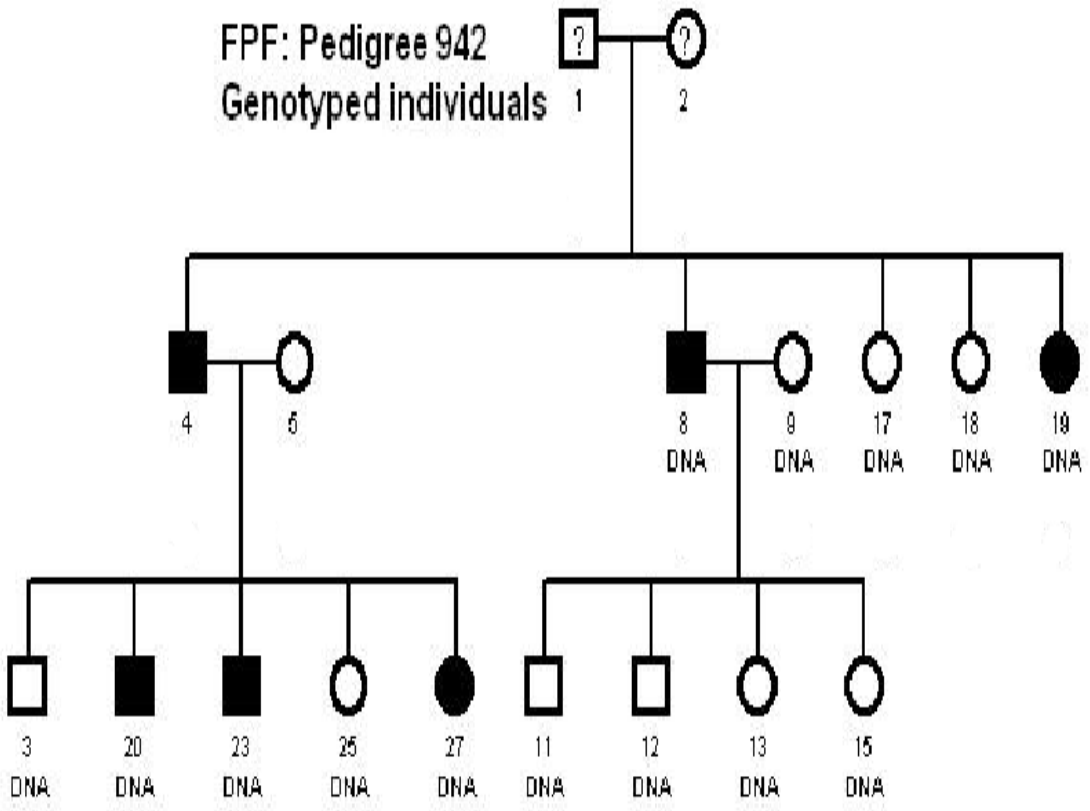


Figure 7.4: Pedigree structure, phenotype for the FPF pedigree R0942

7.2 Linkage Analysis Results

In a pedigree, it is not possible to identify recombinants unambiguously and counts them. Morton (1955) demonstrated that the LOD score method represents the most

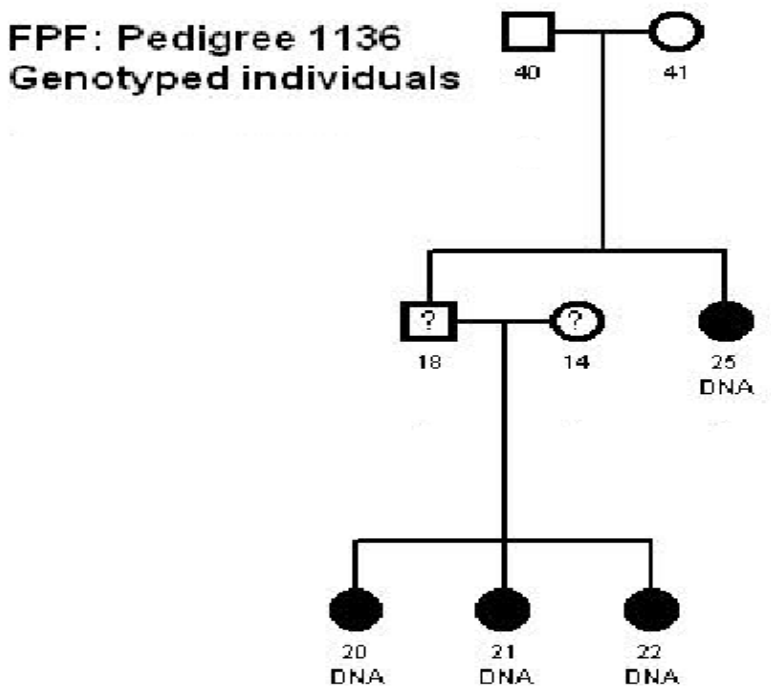


Figure 7.5: Pedigree structure, phenotype for the FPF pedigree R1136

efficient statistic for linkage analyses between Mendelian characters. In a set of families, the overall probability of linkage is the product of the probabilities in each family, therefore LOD scores can be added up across families. The LOD score with value 3 which corresponds to $1000:1$ odds $[\log_{10}(1000) = 3.0]$ is the threshold for accepting linkage with 5% chance of error. This can be quantified in a Bayesian calculation such that $1000:1$ odds corresponds precisely to the conventional $p = 0.05$ threshold of significance.

Two-point and multi-point linkage analysis are conducted for five pedigrees (R0851, R0942, R0892, R0896, and R1136) where 73 individuals are genotyped from Illumina 610 Quad Array. To speed up the linkage analyses, pedigrees were trimmed to remove non-genotyped founders (parents of spouses), non-genotyped individuals, and little impact individuals on linkage (Figure 7.1 to 7.5, legend for these Figures: Square

- male; Cycle - female; Black - affected; White: - unaffected; ? mark - uncertainty phenotype; DNA - genotyped individual).

Two-point analyses have the advantage of being (relatively) easy to do and computationally fast. We carry out the Two-point linkage analyses using Merlin on Linux and the LOD scores were compiled by extracting results from the Merlin output files. We force 363 SNPs with the highest two-point LOD (LOD $\geq$ 2) scores for tag SNPs. As seen in Figures 7.6 to 7.9 and table 7.2, there are three chromosomes (2, 6, and 18) with interesting peaks with LOD scores higher, or near to, 3. Chromosome 6 gives the highest LOD score peak of 3.22 at marker rs942631 and 3.15 at marker rs3130922 (see figure 7.8). Chromosome 18 gives the best result with a small number of negative values near the peak region (see figure 7.9) where the peak is 2.40 for three markers: rs12607533, rs11082034 and rs4528652. The region can be generally defined to be located between the first result with LOD $<$ - 2 at rs3786276 and rs307082 in an interval of 3 cM genetic distance which corresponds to physical distance from 320914936 to 35798935 base pairs of the peak region.

However the main disadvantage of two-point analyses is that the confidence intervals for the estimates of the two-point recombination frequencies are often very wide.

Chr	Marker	Lod	cM	Chr	Marker	Lod	cM	Chr	Marker	Lod	cM
2	rs13410055	2.07	129.665	6	rs581046	2.39	30.053	6	rs9466615	2.29	46.345
2	rs4849902	2.08	130.460	6	rs9369762	2.38	30.450	6	rs12523660	2.30	46.348
2	rs6706968	2.73	130.493	6	rs6907651	2.38	30.450	6	rs13216722	2.30	46.355
2	rs277554	2.21	132.048	6	rs7775901	2.38	30.475	6	rs4464783	2.29	46.365
2	rs277547	2.17	132.068	6	rs4715077	2.38	30.488	6	rs4566882	2.26	46.370
2	rs17007730	2.86	134.975	6	rs9369815	2.29	30.527	6	rs4310044	2.26	46.373
2	rs6713772	2.32	134.983	6	rs1937774	2.38	30.529	6	rs4571550	2.26	46.381
2	rs6541829	2.28	134.984	6	rs11751831	2.37	30.539	6	rs12527913	2.29	46.391
6	rs12526976	2.12	2.413	6	rs11755240	2.38	30.541	6	rs17302729	2.08	47.930
6	rs9328087	2.04	6.268	6	rs7742342	2.33	31.029	6	rs3765502	2.74	48.012
6	rs2505658	2.04	6.434	6	rs7764728	2.06	33.046	6	rs9461082	2.14	49.153
6	rs9378384	2.14	10.257	6	rs957387	2.03	34.632	6	rs587009	2.38	49.205
6	rs7741360	2.36	19.490	6	rs6919364	2.12	37.385	6	rs401671	2.27	49.248
6	rs6927500	2.23	19.531	6	rs10949381	2.11	37.412	6	rs2078527	2.46	49.659
6	rs11964049	2.23	19.537	6	rs9383214	2.03	37.864	6	rs2744267	2.39	49.660
6	rs12198986	2.25	19.558	6	rs11756169	2.37	38.999	6	rs4713108	2.01	50.473
6	rs7762096	2.16	22.383	6	rs6923060	2.44	40.385	6	rs6903282	2.01	50.476
6	rs2294729	2.78	22.745	6	rs9368069	2.20	41.199	6	rs2205831	2.03	50.481
6	rs12203770	2.03	23.038	6	rs6915939	2.05	41.207	6	rs3130922	**3.15**	51.279
6	rs855377	2.96	23.348	6	rs7743281	2.05	41.209	6	rs2395043	2.69	51.279
6	rs1206963	2.56	23.366	6	rs16882179	2.21	41.210	6	rs2395045	2.01	51.283
6	rs707782	2.56	23.387	6	rs2876555	2.08	41.412	6	rs3131631	2.01	51.283
6	rs1925768	2.56	23.515	6	rs9358258	2.08	41.413	6	rs2269475	2.05	51.296
6	rs9477228	2.56	23.522	6	rs4712423	2.16	41.418	6	rs2280800	2.05	51.305
6	rs1322826	2.29	23.882	6	rs9356703	2.43	41.422	6	rs2242653	2.05	51.309
6	rs6906943	2.11	24.251	6	rs9358259	2.44	41.424	6	rs3763305	2.04	51.873
6	rs9358307	2.10	24.255	6	rs9358260	2.47	41.425	6	rs11753634	2.10	54.742
6	rs796102	2.04	24.782	6	rs9368104	2.42	41.426	6	rs11758426	2.08	54.751
6	rs645297	2.04	24.787	6	rs9460399	2.36	41.441	6	rs2814982	2.08	54.770
6	rs942631	**3.22**	25.126	6	rs1209816	2.65	41.491	6	rs2814985	2.03	54.771
6	rs9357002	2.28	25.675	6	rs10946363	2.30	42.199	18	rs1469945	2.23	54.050
6	rs2179179	2.32	26.833	6	rs9466024	2.34	43.612	18	rs2919999	2.23	54.067
6	rs7760294	2.53	26.890	6	rs1322884	2.12	44.010	18	rs717948	2.08	55.581
6	rs10484453	2.55	26.918	6	rs196048	2.09	45.269	18	rs3826608	2.12	57.306
6	rs10498677	2.55	27.112	6	rs7451606	2.03	45.562	18	rs505601	2.36	57.763
6	rs17533974	2.33	28.177	6	rs1935005	2.04	45.563	18	rs9948912	2.02	57.994
6	rs209779	2.26	28.179	6	rs1033440	2.63	46.050	18	rs4799982	2.23	59.275
6	rs9296224	2.03	28.295	6	rs2744143	2.39	46.051	18	rs4129469	2.23	59.303
6	rs913021	2.27	28.295	6	rs2655439	2.08	46.052	18	rs12607533	2.40	59.305
6	rs511574	2.13	30.031	6	rs4280956	2.20	46.269	18	rs11082034	2.40	59.318
6	rs522923	2.06	30.040	6	rs4345386	2.29	46.340	18	rs4528652	2.40	59.331
6	rs560810	2.39	30.046	6	rs13208193	2.29	46.340	18	rs12970162	2.37	59.335

Table 7.2: Markers with LOD scores higher than 2 obtained by two-point linkage analysis of chromosome 2, 6, and 18

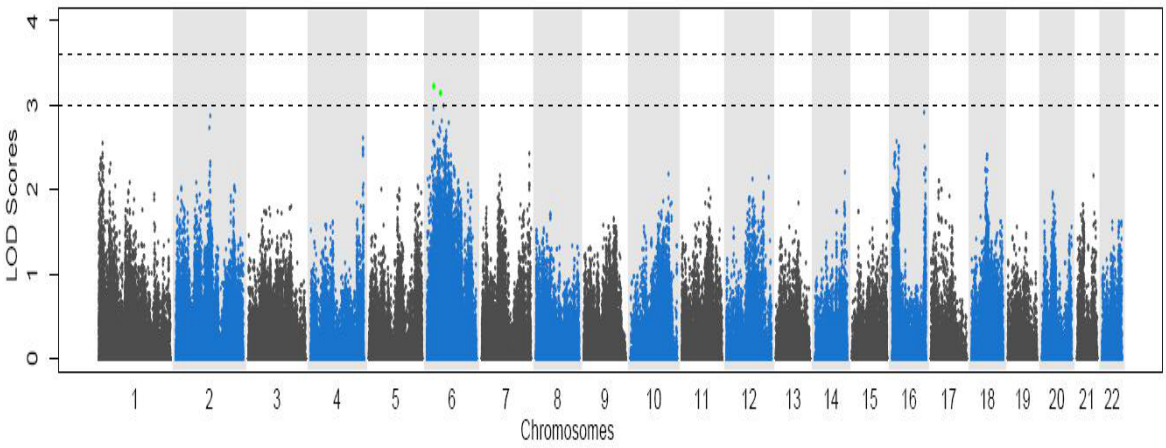


Figure 7.6: Two-point linkage LOD score on 550k SNPs for 5 FPF pedigrees on 22 autosomal chromosomes

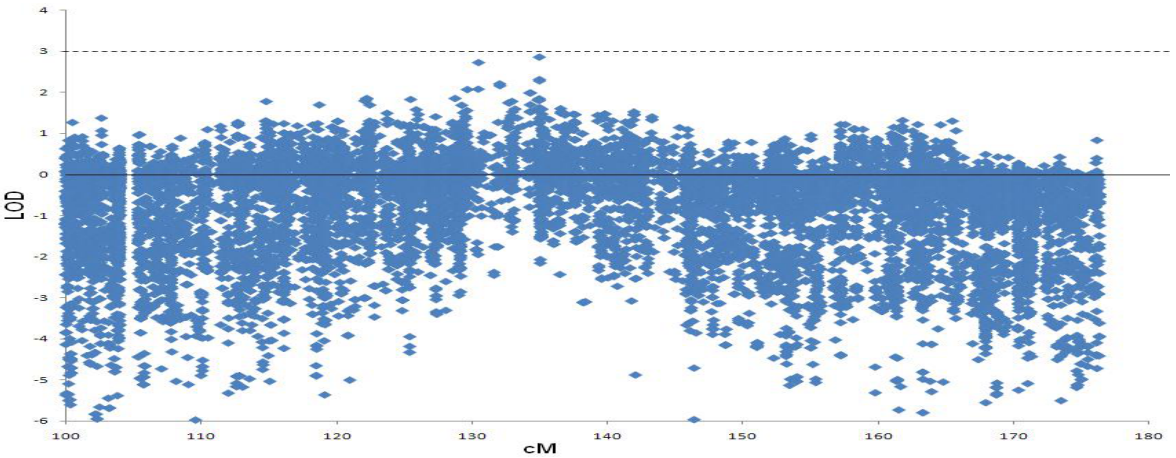


Figure 7.7: Two-point linkage LOD score on 550k SNPs for 5 FPF pedigrees on chromosome 2

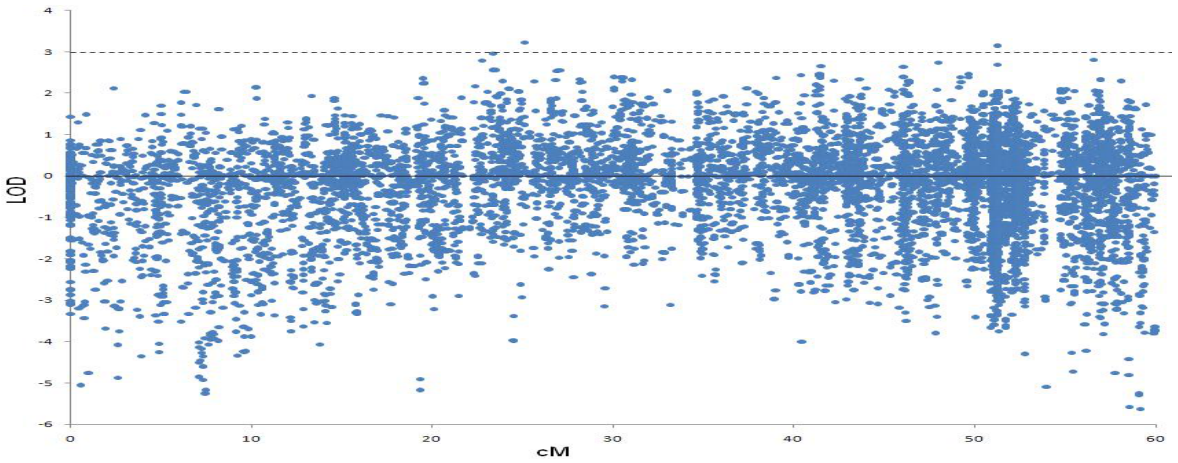


Figure 7.8: Two-point linkage LOD score on 550k SNPs for 5 FPF pedigrees on chromosome 6

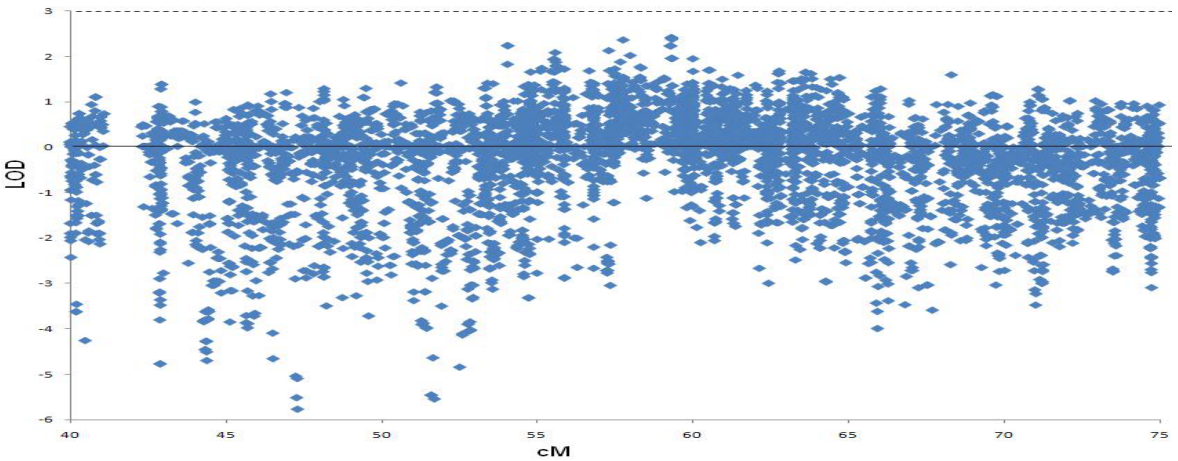


Figure 7.9: Two-point linkage LOD score on 550k SNPs for 5 FPF pedigrees on chromosome 18

To overcome this disadvantage, we can use multi-point analysis. Multi-point analysis is the most efficient method of detecting linkage, estimating recombination, and determining gene order compared to two-point analysis from family data. Experimental geneticists have used multi-point analysis for linkage study for long time.

The multi-point linkage analyses results are showed here that obtained by Merlin software. There are three chromosomes (chromosome 6, 16, and 18) with peaks near 2 or above 2 for three-point parametric linkage analysis (see figures 7.10 to 7.13, and tables 7.3). Chromosome 18 is the only one with LOD scores above 2.

Chr	Marker	Lod	cM	Chr	Marker	Lod	cM	Chr	Marker	Lod	cM
6	rs9367137	1.81	64.369	16	rs8055935	1.97	18.117	18	rs12956578	2.88	57.533
6	rs9394894	1.94	64.468	16	rs17669255	1.97	18.170	18	rs12607135	2.89	57.534
6	rs12195244	1.94	64.510	16	rs1478713	1.97	18.183	18	rs1050265	2.96	57.561
6	rs9349232	1.94	64.680	16	rs1551960	1.97	18.200	18	rs1390430	2.98	57.575
6	rs365387	1.94	64.682	16	rs1038106	1.97	18.244	18	rs12957930	**3.01**	57.596
6	rs260253	1.94	64.701	16	rs1038103	1.97	18.252	18	rs12956471	**3.01**	57.601
6	rs12201447	1.94	64.705	16	rs17669791	1.97	18.285	18	rs16967980	**3.02**	57.615
6	rs7756342	1.94	64.792	16	rs4411516	1.97	18.368	18	rs9652996	**3.03**	57.637
6	rs9367148	1.94	64.811	16	rs2346602	1.97	18.386	18	rs2469881	**3.06**	57.736
6	rs7753593	1.91	65.036	16	rs2178720	1.97	18.500	18	rs519309	**3.07**	57.770
6	rs5014584	1.91	65.037	16	rs12709192	1.97	18.524	18	rs11665085	**3.07**	57.770
6	rs375435	1.89	65.139	16	rs2103403	1.97	18.539	18	rs2847593	**3.07**	57.777
6	rs621627	1.89	65.143	16	rs4536494	1.96	18.647	18	rs3747899	**3.08**	57.873
16	rs11862743	1.81	16.856	16	rs12444565	1.96	18.905	18	rs1786060	**3.08**	57.945
16	rs8055674	1.89	16.939	16	rs17671833	1.96	18.965	18	rs4799911	**3.09**	58.053
16	rs43142	1.97	17.017	16	rs17563428	1.86	19.129	18	rs4077472	**3.09**	58.089
16	rs1476968	1.97	17.066	18	rs17649254	2.07	57.313	18	rs611473	**3.10**	58.131
16	rs7195768	1.97	17.194	18	rs2096889	2.09	57.332	18	rs1539847	**3.11**	58.297
16	rs17140584	1.97	17.263	18	rs567058	2.12	57.363	18	rs11660785	**3.11**	58.303
16	rs17140687	1.96	17.425	18	rs12454634	2.12	57.366	18	rs7232868	**3.11**	58.322
16	rs9935419	1.96	17.441	18	rs3786279	2.12	57.367	18	rs2027754	**3.11**	58.440
16	rs9302824	1.96	17.453	18	rs1057251	2.12	57.369	18	rs13380988	**3.10**	58.466
16	rs8057575	1.96	17.480	18	rs12961465	2.13	57.382	18	rs16968965	**3.10**	58.478
16	rs8053669	1.96	17.480	18	rs680423	2.14	57.395	18	rs1786802	**3.10**	58.509
16	rs12445315	1.96	17.480	18	rs9957382	2.18	57.436	18	rs6507207	**3.08**	58.655
16	rs17722735	1.96	17.480	18	rs1790649	2.22	57.490	18	rs9953231	**3.08**	58.692
16	rs12598879	1.96	17.648	18	rs8087319	2.22	57.494	18	rs4799952	**3.06**	58.796
16	rs12599604	1.96	17.661	18	rs17746694	2.47	57.500	18	rs751947	**3.01**	59.019
16	rs11860754	1.96	17.759	18	rs1546564	2.47	57.500	18	rs930027	**3.01**	59.034
16	rs17141214	1.96	17.778	18	rs7238168	2.48	57.500	18	rs9954636	2.88	59.200
16	rs8046305	1.97	17.833	18	rs7238355	2.48	57.500	18	rs12607533	2.78	59.305
16	rs8057091	1.97	17.911	18	rs9652993	2.49	57.500	18	rs9948897	2.13	59.411
16	rs11643737	1.97	18.048	18	rs7236364	2.87	57.531	18	rs9807741	2.12	59.416
16	rs8053066	1.97	18.108	18	rs12955215	2.88	57.533				

Table 7.3: Markers with LOD scores higher than 2 obtained by multipoint linkage analysis

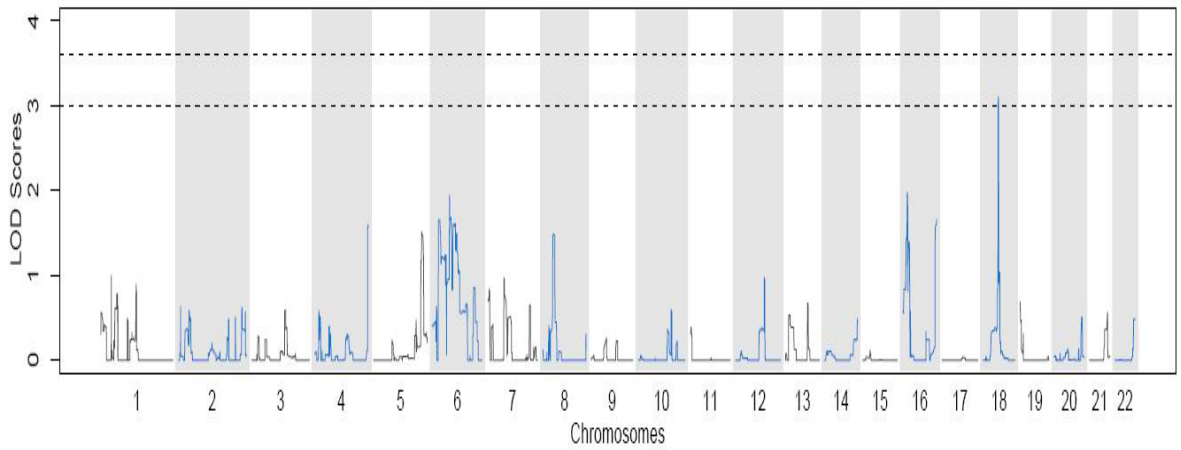


Figure 7.10: multipoint linkage LOD score on 550k SNPs for FPF pedigrees on 22 autosomal chromosomes

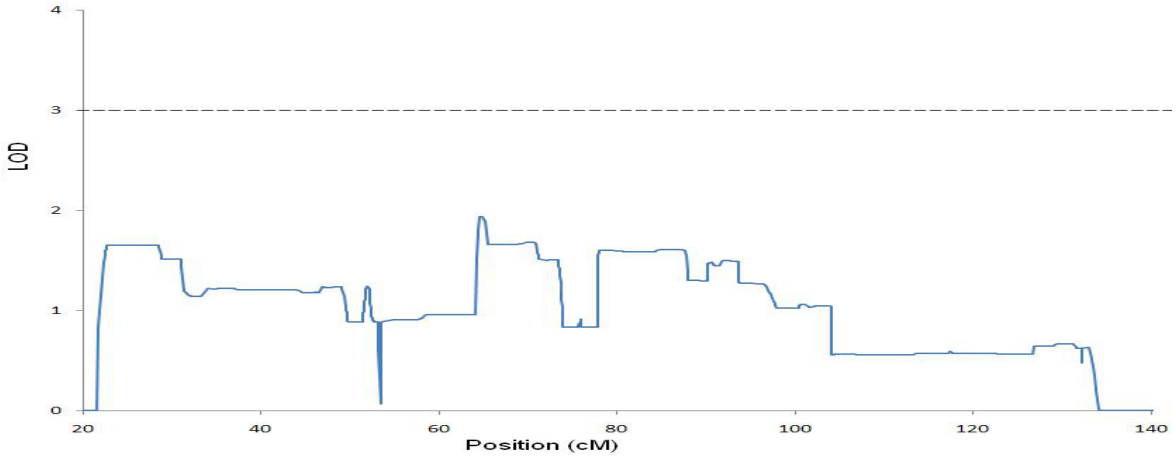


Figure 7.11: Multipoint linkage LOD score on 550k SNPs for 5 FPF pedigrees on chromosome 6

Chromosome 6 has 13 loci with LOD scores above 1.8. The near 2 LOD score markers are from rs9367137 (64.389 cM) to marker rs621627 (65.147 cM), and the

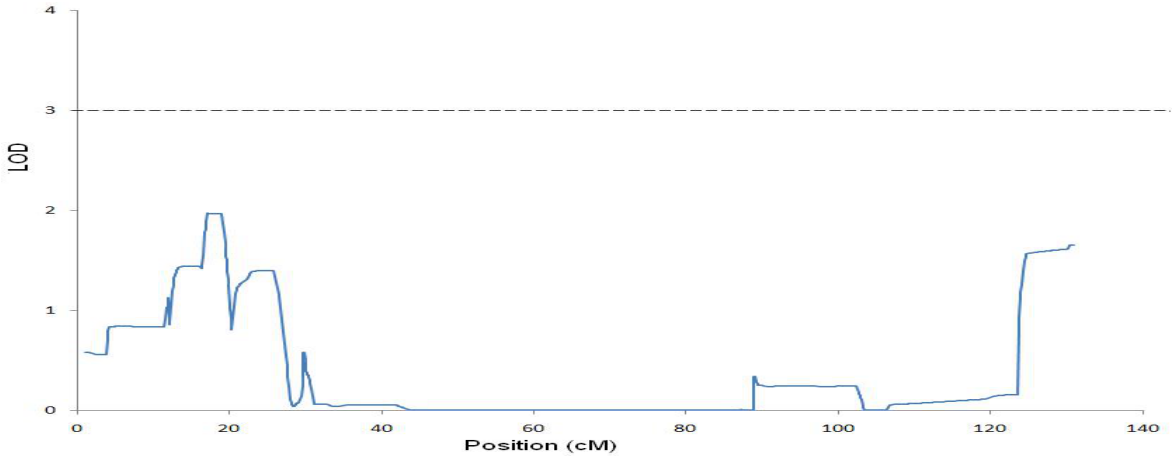


Figure 7.12: Multipoint linkage LOD score on 550k SNPs for 5 FPF pedigrees on chromosome 16

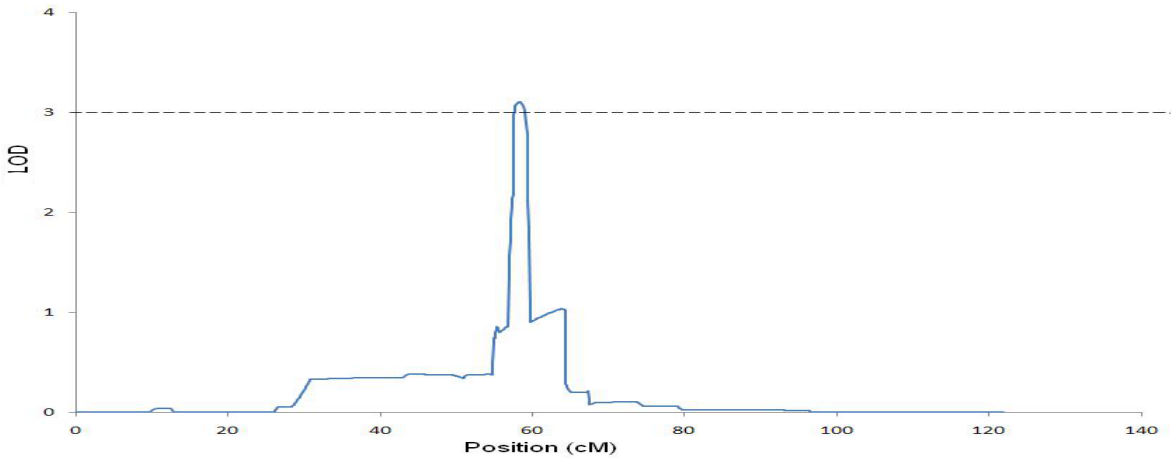


Figure 7.13: Multipoint linkage LOD score on 550k SNPs for 5 FPF pedigrees on chromosome 18

highest parametric LOD score for the peak is 1.94 from marker rs9394894 (64.468 cM) to marker rs9367148 (64.811 cM). Chromosome 16 shows that the markers that LOD scores above 1.8 are from rs11862743 (16.856 cM) to marker rs17563428 (19.129 cM), the highest parametric LOD score for the first cluster of peaks is 1.97 from marker rs43142 (17.017 cM) to marker rs17140584 (17.263 cM), and the second peak from marker rs8046305 (17.833 cM) to marker rs2103403 (18.539 cM). Chromosome 18 is the only one with LOD scores above 2 from marker rs17649254 (57.313 cM) to rs9807741 (59.416 cM). The highest parametric LOD score is 3.11 from marker rs1539847 (58.297 cM) to marker rs2027754 (58.440 cM). For this peak, every family

is contributing positively to the LOD score.

The major disadvantage of multipoint linkage analysis is that for n loci there may be up to $2^{n-1} - 1$ parameters to be estimated. So, if n is greater than 3, either the number of families must be reasonably large , or a priori knowledge concerning values of some of the multi-point recombination frequencies is needed.

7.3 Joint Score Linkage and Association Study Results

For testing the joint score linkage and association in a generalized mixed model with unknown distribution of random variables, we can find an empirical P-value of the test of the mixed chi-square distribution for single-locus one-by-one. The most popular and simplest is the so-called minP test that takes the minimum p value of the individual tests. In the dense-map case, occurrences of spuriously small P-value at nearby markers are no longer independent events. The results of P-value immediately translate into statements about how small a P-value will be expected to occur by chance, given the penalty size of the genome. Specifically, the penalty size M of 550k Illumina SNP data is 1867, and putative trait loci penalty function is given by:

$$P^* = MP$$

Equivalently, one can combine individual score test statistics and $-\log_{10} p$ by their maximum, and the penalty function is:

$$-\log_{10} p^* = -\log_{10} p - 3.27$$

We perform a joint linkage and association study based on the score test for binary

traits with multivariate unknown distribution random effects for FPF families. The results are presented in Table 7.4, Figure 7.14 to 7.18.

There are 16 chromosomes (except chromosome 8, 9, 13, 15, 19 and 20) with values $-\log_{10} p^*$ above 2 for joint linkage and association score test(see figures 7.14 and tables 7.4). Only chromosomes 6, 17, 18, and 22 have the value $-\log_{10} p^*$ around 3 (see figures 7.15 to 7.18). Marker rs4605929 (11.1421 cM) has highest score test statistic 24.56885 and $-\log_{10} p^*$ value 2.800226 in chromosome 6; Marker rs11078200 (39.7220 cM) has statistic 24.30622 and $-\log_{10} p^*$ value 2.735171 in chromosome 17; Marker rs1941686 (55.5969 cM) has statistic value 25.56352 and $-\log_{10} p^*$ value 3.026009 in chromosome 18; Marker rs114682 (36.8622 cM) has statistic 25.55801 and $-\log_{10} p^*$ value 3.019527 in chromosome 22. These four are the most significance markers that relate disease gene.

Chr	Marker	bp	cM	Test-Stat	$-\log_{10} p^*$
chr1	rs12739892	66077155	92.2933	24.13593	2.697791
chr1	rs2734690	86717023	110.7731	23.02958	2.465658
chr1	rs4847183	93957932	117.3230	22.33328	2.312328
chr2	rs6544340	40535657	65.1119	21.32333	2.084857
chr2	rs2121304	53116620	78.1286	22.04461	2.247345
chr2	rs13002109	63200206	84.7203	21.46243	2.114602
chr2	rs2121304	53116620	78.1286	22.04461	2.247345
chr2	rs13002109	63200206	84.7203	21.46243	2.114602
chr3	rs1394764	24380054	46.1834	21.60298	2.151027
chr3	rs9829159	36007577	61.1164	24.11680	2.694197
chr3	rs1147696	121602169	126.9800	22.32073	2.310858
chr3	rs1515577	121611630	126.9800	22.08907	2.259584
chr3	rs1881919	134537735	140.5947	22.28300	2.295886
chr3	rs36059	136215422	143.4800	23.68631	2.600389
chr3	rs4371486	144385643	149.8817	22.47789	2.336017
chr3	rs4839656	144397917	149.9000	22.47789	2.336017
chr3	rs4839629	144410136	149.9182	22.47789	2.336017
chr3	rs4839637	144422638	149.9369	21.66923	2.157343
chr4	rs6838690	162279202	155.0975	22.39278	2.322098
chr5	rs1363576	169332702	179.4815	21.29043	2.081115
chr5	rs7701794	170110717	182.8091	21.00161	2.023576
chr6	rs11759003	2690371	7.5082	22.16126	2.268680
chr6	rs4605929	4096517	11.1421	24.56885	2.800226
chr6	rs2064108	5609463	14.7939	23.32504	2.524652
chr6	rs4716001	9918688	23.3248	23.02176	2.458014
chr6	rs9368621	11344664	26.5768	22.98379	2.438531
chr6	rs7750679	12999287	30.2790	21.50840	2.126581
chr6	rs865226	20299817	42.2858	21.44344	2.114750
chr7	rs156675	131844848	137.5191	22.25209	2.289600
chr7	rs156974	131851408	137.5372	22.16564	2.270321
chr7	rs10260766	131868326	137.5840	22.27549	2.294533
chr7	rs2253200	137592748	145.0690	24.12334	2.709005
chr10	rs10997481	68434476	83.6192	22.83058	2.417149
chr10	rs10885336	114101192	129.3316	21.61788	2.149678
chr10	rs11198686	120696565	141.3733	22.88129	2.427675
chr10	rs11018214	129278677	159.5559	21.97567	2.214423
chr11	rs1528640	14027404	22.7726	21.32439	2.091597
chr12	rs7974181	13309964	30.9746	21.34480	2.086817
chr14	rs11158329	60752466	60.7764	22.23052	2.285244
chr16	rs6497441	9737349	24.9599	22.50869	2.351173
chr16	rs1549662	74880036	92.0534	22.57699	2.369739
chr16	rs7198446	86936870	128.3817	22.95407	2.448039
chr17	rs11078200	13696863	39.7220	24.30622	2.735171
chr18	rs1941686	29554557	55.5969	25.56352	3.026009
chr18	rs12456032	55683064	82.4427	21.09905	2.035142
chr21	rs2828183	23761888	22.8980	20.96074	2.002030
chr21	rs2226674	23840748	23.0123	22.10160	2.250789
chr21	rs8134891	28888157	30.4692	21.33917	2.095521
chr22	rs2073760	17886456	8.0471	23.31451	2.528482
chr22	rs2073762	18151568	8.6980	22.60177	2.375552
chr22	rs2073765	18180322	8.7686	22.91433	2.447694
chr22	rs114682	31398118	36.8622	25.55801	3.019527
chr22	rs1159220	31410753	36.8772	23.57574	2.582233
chr22	rs3788483	31414345	36.8815	21.54420	2.133546

Table 7.4: Markers with $-\log_{10} p^*$ larger than 2 obtained by joint linkage and association analysis

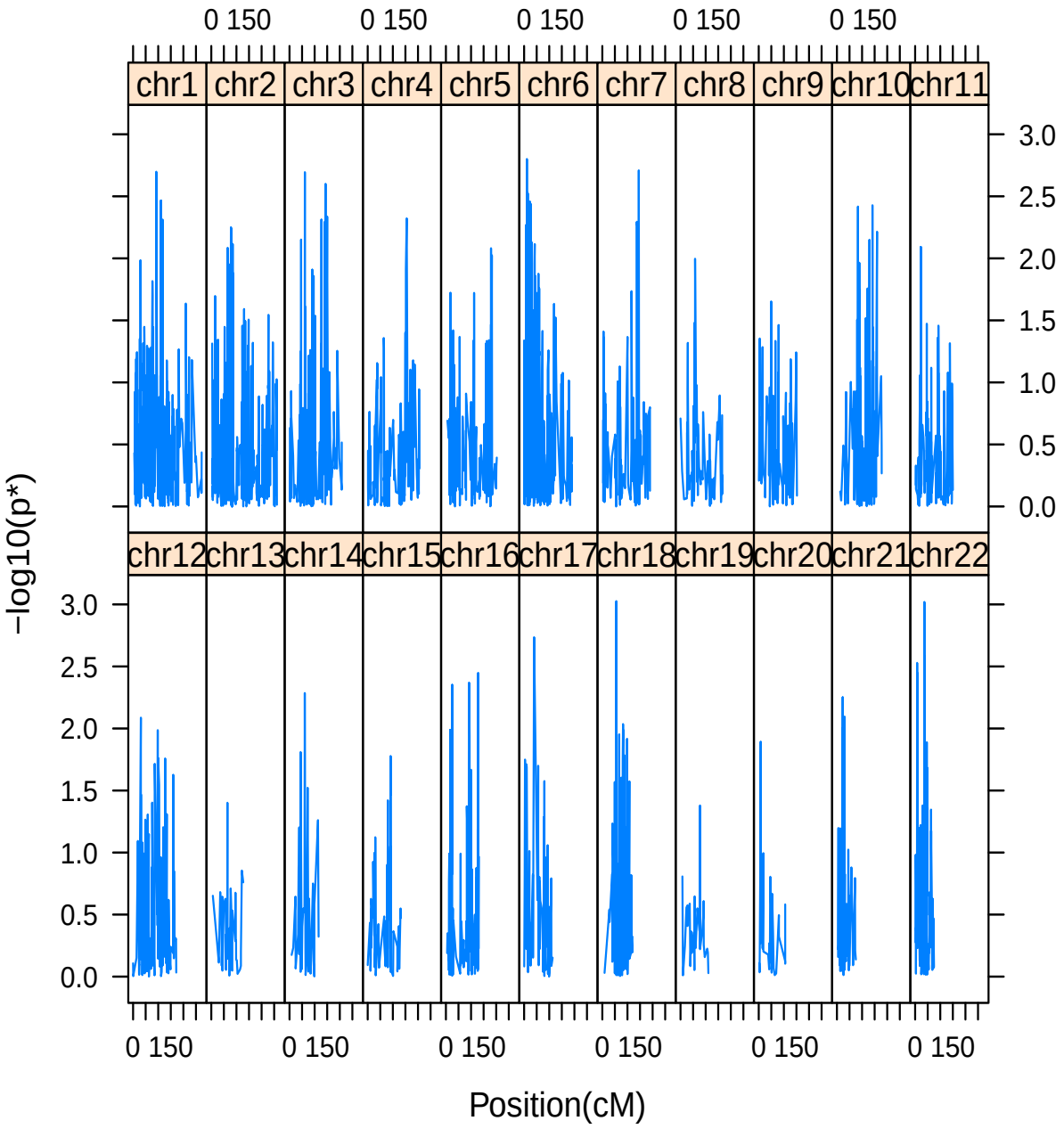


Figure 7.14: $-\log_{10} p^*$ with position of 22 autosomal chromosomes

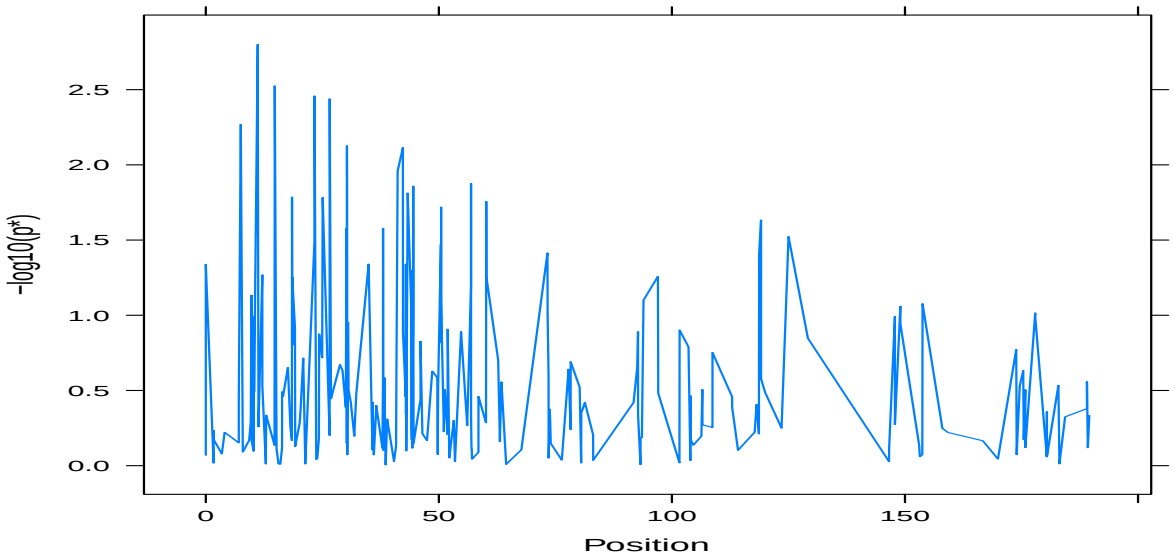


Figure 7.15: $-\log_{10} p^*$ with position of chromosome 6

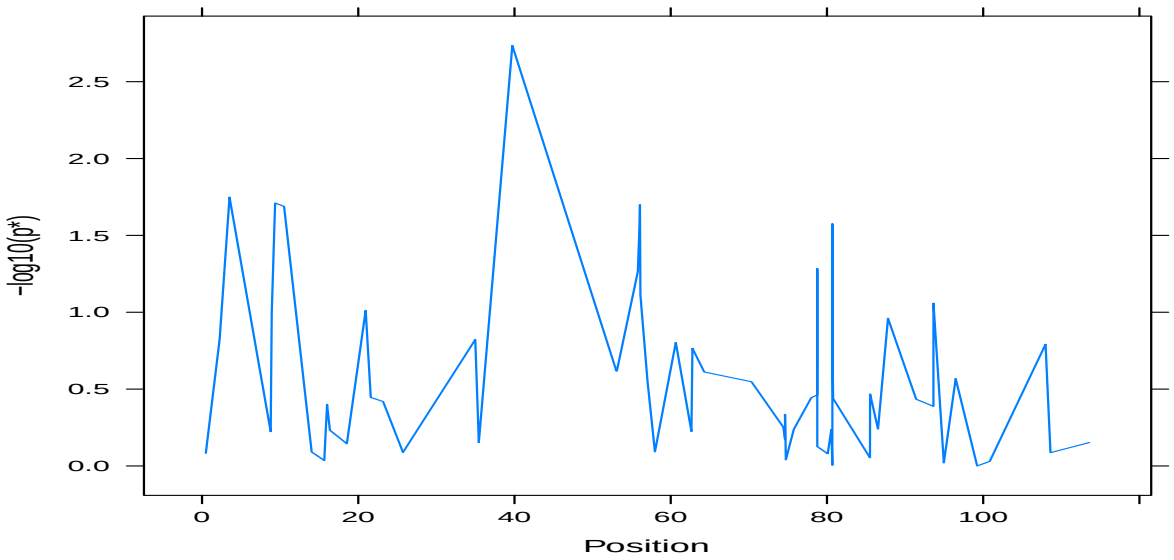


Figure 7.16: $-\log_{10} p^*$ with position of chromosome 17

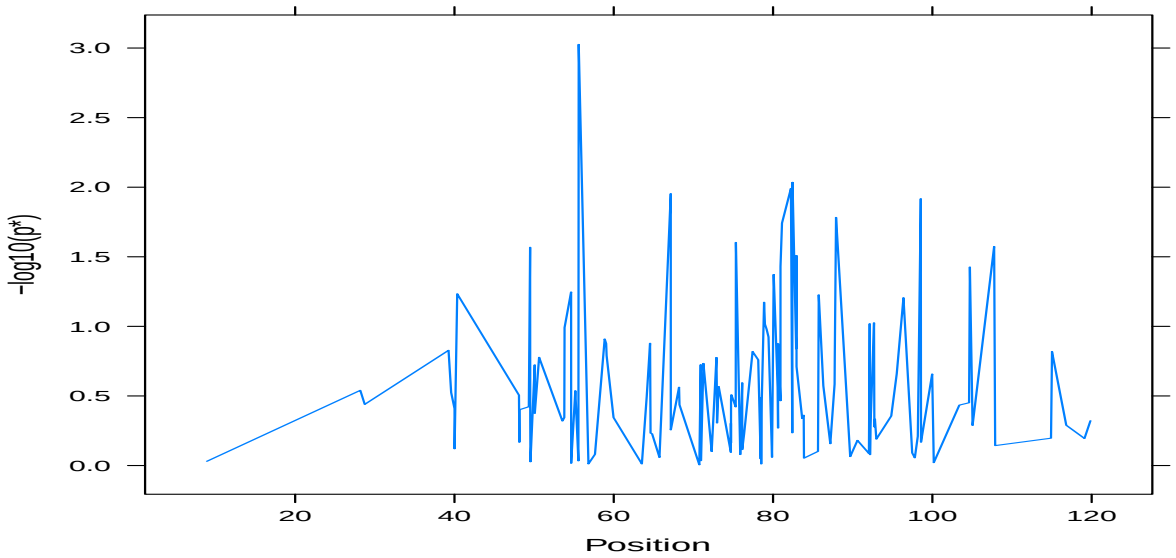


Figure 7.17: $-\log_{10} p^*$ with position of chromosome 18

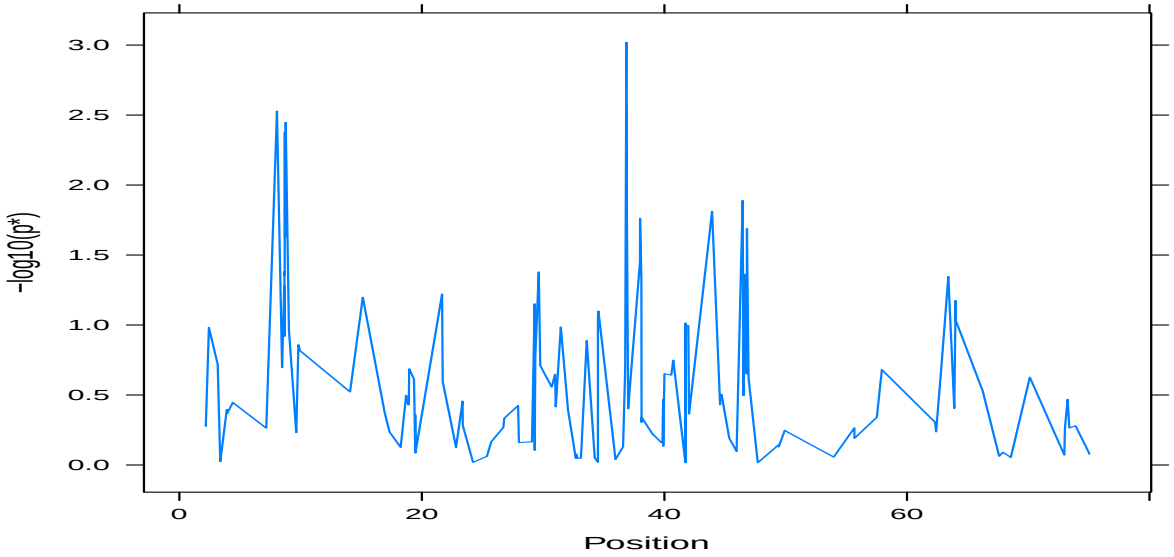


Figure 7.18: $-\log_{10} p^*$ with position of chromosome 22

Chapter 8
Conclusions and Future Works

A joint modeling of genetic linkage and association for combinations of pedigree structures and relationship of relative individuals within pedigrees has been reviewed for families with or without a remote common ancestor. This joint modeling uses a likelihood approach that allows the inclusion of other covariates into the model of quantitative and binary traits.

For quantitative traits, the approach tested has similarities with that of Zhao et al. (1998) and Sham et al. (2000). Zhao et al. (1998) estimated the recombination fraction (RF) for linkage analysis to determine the genetic distance between the putative disease locus and the marker locus. If the estimated RF is significantly less than .5, then a positive linkage can be declared. The presence of linkage disequilibrium implies that the disease allele at the putative disease locus is associated with an allele at the marker locus. Linkage disequilibrium is equivalent to the association between the putative disease allele and the marker allele; that is, they are no longer independent. Zhao et al. (1998) also used the odds ratio for linkage disequilibrium test. Furthermore, combining both linkage and linkage-disequilibrium analysis, this method pro-

duces a combined test statistic that tests the significance of both linkage and linkage disequilibrium, or against the significance of linkage where there is no association, or significance of association where there is no linkage. But this joint modeling does not allow other covariates into the model and it can not be used in a maximum likelihood framework. Sham et al. (2000) introduced a joint likelihood ratio test for linkage and family-based association under a variance-components model. This joint test contains several parameters for linkage and several parameters for association respectively, and assumes the true parameters do not reach the boundary. This joint modeling allows other covariates into the model and it uses the maximum likelihood framework. This method produces a combined test statistic that only detects the presence of both linkage and linkage disequilibrium.

We consider LR, Wald and score tests on testing the joint linkage and association components for binary traits with multivariate normal assumption random variables in our non-linear mixed model. It is confirmed that the Wald joint test is too liberal whereas the LRT and score joint tests are too conservative for large α and ς, and the powers of the score test and the LRT are non-significantly different.

Also, we have explored the score test - the alternative hypotheses based on a set of binary traits with multivariate unknown distribution random variables. The joint score test requires estimation of the model only under the null hypothesis. This approach has some similarities with that of Zelterman and Chen (1988), who derived test statistic through a score test that base on independent of the particular mixing distribution. Although our research shows none powers of linkage score test, but the joint score

test for linkage and association provides a fast and comparable power to LRTs for analysing large and complicated pedigrees.

In this thesis, we have derived a joint test for linkage and association. It can be used for combinations of pedigree structures and relationship of relative individuals within a pedigree. We also have defined separate tests for linkage and association by using either of the two components separately. Our framework facilitates efficiency comparisons between the joint, association, and linkage tests. When comparing linkage and association, no method is uniformly superior. The simulation study shows that the joint tests have a level of significance close to the nominal levels of quantitative traits, but the levels of significance of the joint tests are slightly larger than the nominal 0.01 level of significance of binary traits; in addition, the joint test is more powerful than linkage or association test alone when both sources of variation are present. Furthermore, the joint method may also test against specific alternatives - for example, against the significance of linkage where there is no association, against significance of association where there is no linkage, against significance of both linkage and association.

From our joint genome-wide family-based approach, we could verify several markers which already confirmed evidence of linkage and association for FPF in 5 pedigrees. Through linkage and association analyses, marker rs4605929 in chromosome 6, marker rs11078200 in chromosome 17, marker rs1941686 in chromosome 18, and marker rs114682 in chromosome 22 are the most significance markers related the disease gene.

Our joint modeling of genetic linkage and association can be present phenocopies

in families, modeling the contributions of environmental covariates with or without familial correlations. With this extension, our proposed methods are even more useful for assessing the performance of various pedigree analysis methods that incorporate environmental covariates or search for more than one disease gene at a time.

In human studies, the putative disease alleles are generally unobserved and may need to be inferred on the basis of the observed phenotypes and their marker genes. A joint linkage and association test based on inferred latent variables will be investigated in a future study. To improve the efficiency of joint linkage and association analysis, multipoint joint linkage and association analysis should be considered.

Reference

Abecasis GR, Cardon LR, Cookson WO (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66**: 279-292.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) MERLIN-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**: 97-101.

Agresti A. (1996) An introduction to categorical data analysis. John Wiley and Sons, Inc.

Allison DB (1997) Transmission disequilibrium tests for quantitative traits. *Am J Hum Genet* **60**: 676-690.

Allison DB, Fernandez JR, Heo M, Beasley TM (2000) Testing the robustness of the new Haseman-Elston quantitative-trait loci-mapping procedure. *Am. J. Hum. Genet.* **67**: 249-252.

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **62**: 1198-1211.

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* **54**: 535-543.

Amos CI, Zhu DK, Boerwinkle E (1996) Assessing genetic linkage and association with robust components of variance approaches. *Ann Hum Genet* **60**: 143-160.

Armanios MY, Chen JJ, Cogan JD, Alder JK, Ingersoll RG, Markin C, Lawson WE, Xie M, Vulto I, Phillips JA, Lansdorp PM, Greider CW, Loyd JE (2007) Telomerase mutations in families with idiopathic pulmonary fibrosis. *N Engl J Med.* **356**: 1317-26.

Arminger G, K$\ddot{u}$sters U (1988) Latent trait models with indicators of mixed measurement level. In latent trait and latent class models (eds Langeheine R and Rost J). New York: Plenum.

Arunachalam V, Owen A (1971) Polymorphisms with linked loci (London: Chapman and Hall)

Aulchenko YS, Struchalin1 MV, Duijn CM (2010) ProbABEL package for genome wide association analysis of imputed data. *BMC Bioinformatics* **11**: 134.

Bain L, Yang JD, Guo TW, Duan Y, Qin W, Sun Y, Feng GY and He L (2005) Association study of the A2M and LRP1 genes with Alzheimer disease in the Han Chinese. *Biol. Psychiat.* **58**: 731-737.

Baksh MF, Balding DJ, Yyes TJ, Whittaker JC (2007) Family-based association analysis with ordinal categorical phenotypes, covariates and interactions. *Genet Epidemiol* **31**: 1-8.

Barrett JC and Cardon LR (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**: 659-662.

Bartoo JB, Puri PS (1967) On optimal asymptotic tests of composite statistical hypotheses, *Annals of Mathematical Statistics* **38**: 1845-1852.

Berkhof J, Snijders T (2001) Variance component testing in multilevel models. *Journal of educational and behavioral statistics* **26**: 133-152.

Biernacka JM, Cordell HJ (2007) Exploring causality via identification of SNPs or haplotypes responsible for a linkage signal. *Genet Epidemiol.* **31**: 727-740.

Blangero J, Almasy L (1997) Multipoint oligogenic linkage analysis of quantitative traits. *Genet Epidemiol* **14**: 959-964.

Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am J Hum Genet* **62**: 950-961.

Boomsma DI, Beem AL, van den Berg M, Dolan CV, Koopmans JR, Vink JM, De G, Slagboom PE (2000) Netherlands twin family study of anxious depression (NEDSAD). *Twin Res.* **3**: 323-334.

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex diseases. *Nat Genet Suppl* **33**: 228-237.

Cardon LR, Abecasis GR (2000) Some properties of a variance components model for fine-mapping quantitative trait loci. *Behav Genet* **30**: 235-243.

Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci USA* **85**: 9119-9123.

Chant D (1974) On asymptotic tests of composite hypotheses in nonstandard conditions, *Biometrika* **61**: 291-298.

Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power *Hum Hered* **56**: 18-31.

Chernoff H (1954) On the distribution of the likelihood ratio. *Ann. Math. Statist.* **25**: 573-578.

Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* **65**: 1170-1177.

Cox DR, Hinkley DV (1974) *Theoretical Statistics*. Chapman and Hall, London.

Crowder MJ (1978) Beta-binomial ANOVA for proportions, *Applied Statistics* **27**: 34-37.

Curtis D (1997) Use of siblings as controls in case-control association studies. *Ann Hum Genet* **61**: 319-333.

Curtis D, Sham PC (1996) Population stratifications can cause false positive linkage results if founders are untyped. *Ann Hum Genet* **60**: 261263.

Curtis D, Knight J, Sham PC (2006) Program report: GENECOUNTING support programs. *Ann. Hum. Genet.* **70**: 277-279.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B* **39**: 1-38.

Deng HW (2001) Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. *Genetics* **159**: 13191323.

Diao G, Lin DY (2005) A powerful and robust method for mapping quantitative trait loci in general pedigrees. *Am J Hum Genet* **77**: 97-111.

Diao G, Lin DY (2006) Improving the power of association tests for quantitative traits in family studies. *Genet Epidemiol* **30**: 301-313.

Diao G, Lin DY (2010) Variance-components methods for linkage and association analysis of ordinal traits in general pedigrees *Genet Epidemiol.* **34**: 232-237.

Drum M, Mccullagh P (1993) REML estimation with exact covariance in the logistics mixed model. *Biometrics* **49**: 677-89.

Duerr RH, Taylor KD, Brant SR, et al. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**: 1461-1463.

Epstein .P, Lin XH, Boehnke M (2003) A tobit variance-component method for linkage analysis of censored trait data. *Am J Hum Genet* **72**: 611-620.

Evans DM, Medland SE (2003) A note on including phenotypic information from monozygotic twins in variance components QTL linkage analysis. *Ann. Hum. Genet.* **67**: 613-617.

Fahrmeir L, Tutz G (2001) *Multivariate Statistical Modelling based on Generalized Linear Models, Springer Series in Statistics*, 2nd ed., Springer,
New York/Berlin/Heidelberg.

Fan R, Xiong M (2003) Combined high resolution linkage and association mapping of quantitative trait loci. *Eur J Hum Genet* **11**: 125-137.

Farrer LA, Cupples LA, Haines JL, et al. (1997) Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* **278**: 1349-1356.

Feng Z, McCulloch CE (1992) Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics and Probability Letters* **13**: 325-332.

Falconer DC, Mackay TFC (1996) *Introduction to Quantitative Genetics*. Pearson Education, Harlow.

Fitzmaurice GM, Lipsitz SR, Ibrahim JG (2007) Anote on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics* **63**: 942-946.

Fog A (2008) Calculation methods for Wallenius' noncentral hypergeometric distribution. *Communications in Statistics, Simulation and Computation* **37 (2)**: 258273

Fulker DW, Cherny SS (1996) An improved multipoint sib-pair analysis of quantitative traits. *Behav. Genet.* **26**: 527-532.

Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* **64**: 259-267.

Goldgar DE (1990) Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet* **47**: 957-967.

Goring HH, Terwilliger JD (2000) Linkage analysis in the presence of error. II. Marker-locus genotyping errors modeled with hypercomplex recombination fraction. *Am. J. Hum. Genet.* **66**: 1107-1118.

Gosso MF, van Belzen M, De Geus EJC, Polderman JC, Heutink P, Boomsma DI, Posthuma D (2006) Association between the CHRM2 gene and intelligence in a sample of 304 Dutch families. *Genes Brain Behav.* **5**: 577-584.

Gusella JF, Wexler NS, Conneally PM, et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**: 234-238.

Haldane JBS (1919) The combination of linkage values and the calculation of distance between the loci of linked factors. *J Genet* **8**: 299309.

Hall CB, Praestgaard JT (2001) Order-restricted score tests for homogeneity in generalized linear and nonlinear mixed models. *Biometrika* **88**: 739-751.

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2**: 3-19.

Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72**: 320338.

Herbert A, Gerry NP, McQueen MB, et al. (2006) A common genetic variant is associated with adult and childhood obesity. *Science* **312**: 279-283.

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**: 226-231.

Horvath SM, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet* **63**: 1886-1897.

Horvath S, Xu X, Laird NM (2001) The family based association test method: strategies for studying general genotype-phenotype associations. *European J. Hum. Genet* **9**: 301-306.

H*ö*ssjer O (2003) Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. *Ann. Statist* **31**: 1075-1109.

H*ö*ssjer O (2005a) Conditional likelihood score functions for mixed models in linkage analysis. *Biostatistics* **6**: 313-332.

H*ö*ssjer O (2005b) Combined association and linkage analysis for general pedigree and genetic models. *Statistical Application in Genetics and Molecular Biology* **4**, Article 11.

Hopper JL, Mathews JD (1982) Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet* **46**: 373-383.

Jacqmin GH, Commenges D (1995) Tests of homogeneity for generalized linear models. *Journal of the American Statistical Association* **90**: 1237-1246.

Kerem B, Rommens JM, Buchanan JA, et al. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**: 1073-1080.

Klein RJ, Zeiss C, Chew EY, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308**: 385-389.

Knapp M (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *Am J Hum Genet* **64**: 861-870.

Koenig M, Hoffman EP, Bertelson CJ, et al. (1987) Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* **50**: 509-517.

Kong X, Murphy K, Raj T, He C, White PS, Matise TC (2004) A combined linkage-physical map of the human genome. *Am J Hum Genet* **75**: 1143-1148.

Kosambi DD (1944) The estimation of map distance from recombination values. *Annals of Eugenics* **12**: 172175.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified approach. *Am J Hum Genet* **58**: 1347-1363

Labbe A, Bureau A, Merette C (2009) Integration of genetic familial dependence structure in latent class models, *The International Journal of Biostatistics* **5**, Iss. 1, Article 6.

Lake SL, Blacker D, Laird NM (2000) Family-based tests of association in the presence of linkage. *Am. J. Hum. Genet* **67**: 1515-1525.

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. *Science* **265**: 2037-2048.

Lange K (2002) *Mathematical and Statistical Methods for Genetic Analysis.* Springer-Verlag, NY.

Lange C, DeMeo DL, Laird NM (2002) Power and design considerations for a general class of family-based association tests: quantitative traits. *Am J Hum Genet* **71**: 1330-1341.

Lange K, Westlake J, Spence MA (1976) Extensions to pedigree analysis. III. Variance components by the scoring method. *Ann Hum Genet* **39**: 485-491.

Laird N, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. *Nature Reviews genetics.* **7**: 385-394.

Laird NM, Ware JH (1982) Random effects models for longitudinal data. *Biometrics* **38**: 963-974.

Lee SY (2007) *Handbook of Latent Variable and Related Models*. North-Holland.

Lee SY, Poon WY, Bentler P (1992) Structural equation models with continuous and polytomous variables. *Psychometrika* **57**: 89-105.

Lee WC (2003) Searching for disease susceptibility loci by testing for Hardy-Weinberg disequilibrium in a gene bank of affected individuals. *Am J Epidemiol* **158**: 397400.

Legler JM, Ryan LM (1997) Latent variable models for multiple birth outcomes. *J. Am. Statist. Ass.* **92**: 13-20.

Lehmann EL (1983) *Theory of Point Estimation*. Wiley, New York.

Lemire (2005) A simple nonparametric multipoint procedure to test for linkage through mothers or fathers as well as imprinting effects in the presence of linkage. *BMC Genetics* **6**(Suppl 1): S159.

lewontin R, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. *Evolution* **14**: 458-472.

Li M, Boehnke M, Abecasis GR (2005) Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. *Am. J. Hum. Genet.* **76**: 934-949.

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models, *Biometrika* **73**: 13-22.

Lin X (1997) Variance component testing in generalized linear models with random effects. *Biometrika* **84**: 309-326.

Lin X, Breslow NE (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Am. Statist. Assoc.* **91**: 1007-1016.

Little RJ, Rubin DB (1987) *Statistical Analysis with Missing Data*. New York: Wiley

Loredo-Osti JC (2014) A cautionary note on ignoring polygenic background when mapping quantitative trait loci via recombinat congenic strains. *Front. Genet.* 5:68, doi:10.3389/fgene.2014.00068.

Martin ER, Bass MP, Hauser ER, Kaplan NL (2003) Accounting for linkage in family-based tests of association with missing parental genotypes. *Am J Hum Genet* **73**: 1016-1026.

Lemire M, Roslin NM, Laprise C, Hudson TJ, Morgan K (2004) Transmission-ratio distortion and allele sharing in affected sib pairs: a new linkage statistic with reduced bias, with application to chromosome 6q25.3 *Am. J. Hum. Genet.* **75**: 571586.

McCulloch CE (1994) Maximum likelihood variance components estimation for binary data, *Journal of the American Statistical Association* **89**: 330-335.

McCulloch CE, Searle SR (2001) *Generalized, Linear, and Mixed Models*, Wiley, New York.

McLachlan GJ, Krishnan K (1997) *The EM Algorithm*, Wiley, New York.

McPeek S. (1999) Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genet. Epidemiol* **16**: 225-249.

Menard S. (1995) Applied logistic regression analysis. Sage publications. Series: quantitative applications in the social sciences, No. 106.

Monks SA, Kaplan NL, Weir BS (1998) A comparative study of sibship tests of linkage and/or association. *Am J Hum Genet* **63**: 1507-1516.

Moran P (1971) Maximum likelihood estimators in nonstandard conditions, *Proc. Cambridge Philos. Soc.* **70**: 441-450.

Morton NE (1955) Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7**: 277-318.

Muthén B (1984) A general structural model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika* **49**: 115-132.

Muthén B (1987) LISCOMP: a computer program. Chicago: Chicago Scientific Software.

Nielsen DM, Ehm MG, Weir BS (1998) Detecting marker disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *Am J Hum Genet* **63**: 15311540.

Owerbach D, Naya FJ, Tsai MJ, Allander SV, Powell DR, Gabbay KH (1997) Analysis of candidate genes for suceptibility to type I diabetes - a case-control and family-association study of genes on chromosome 2q31-35. *Diabetes* **46**: 1069-1074.

Parfrey P, Davidson W, Green J (2002) Clinical and genetic epidemiology of inherited renal disease in Newfoundland. *Kidney International* **61**: 1925-1934.

Pawitan Y (2001) *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

Pericak-Vance MA, Bass MP, Yamaoka LH, Gaskell PC, Scott WK, Terwedo HA, Menold MM, Conneally PM, Small GW, Vance JM, Saunders AM, Roses AD, Haines JL (1997) Complete genomic screen in late-onset familial Alzheimer diseaseevidence for a new locus on chromosome 12. *JAMA* **278**: 12371241.

Prentice RL (1988) Correlated binary regression with covariates specific to each binary observations. *Biometrics* **4**: 1033-1048.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. *American Journal of Human Genetics* **67**: 170-181.

Rabinowitz D (1997) A transmission disequilibrium test for quantitative traits. *Hum Hered* **47**: 342-350.

Rabinowitz D, Laird NM (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity* **50**: 211-223.

Rao CR (1973) *Linear Statistical Inference and Its Applications (2nd ed.)*, New York: John Wiley.

Riordan JR, Rommens JM, Kerem B, et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of cDNA. *Science* **245**: 1066-1073.

Rommens JM, Iannuzzi MC, Kerem B, et al. (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**: 1059-1065.

Rubin DB (1992) Computational aspects of analyzing random effects/longitudinal models. *Statist. Med.* **11**: 1809-1821.

Sammel MD, Ryan LM (1996) Latent variable models with fixed effects. *Biometrics* **52**: 220-243.

Sammel MD, Ryan LM, Legler JM (1997) Latent variable models for mixed discrete and continuous outcomes. *J. R. Statist. Soc.* B **59**: 667-678.

Schork NJ (1993) Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. *Am J Hum Genet* **53**: 1306-1319.

Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82**: 605-610.

Self SG, Longton G,Kopecky KJ, Liang KJ (1991) On estimating HLA/disease association with application to a study of aplastic anaemia. *Biometrics* **47**: 53-61.

Sham P (1998) *Statistics in Human Genetics*. Arnold, London.

Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multi-allelic marker loci. *Ann Hum Genet* **59**: 323-336.

Sham PC, Cherny SS, Purcell S, Hewitt JK (2000) Power of linkage versus association analysis of quantitative traits, by use of variance components models, for sibship data. *Am. J. Hum. Genet.* **66**: 1616-1630.

Sham PC, Rijsdijk FV, Knight J, Makoff A, North B, Curtis D (2004) Haplotype association analysis of discrete and continuous traits using mixture of regression models. *Behav. Genet.* **34**: 207-214.

Shapiro A (1985) Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints *Biometrika* **72**: 133-144.

Silvapulle M, Silvapulle P (1995) A score test against one-sided alternatives. *Journal of the American Statistical Association* **90**: 342-349.

Sinha SK (2009) Bootstrap tests for variance components in generalized linear mixed models. *The Canadian Journal of Statistics* **37**: 219-234.

Skrondal M, Rabe-Hesketh S (2004) *Generalized Latent Variable Modeling*. Chapman & Hall/CRC.

Sladek R, Rocheleau G, Rung J, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diseases. *Nature* **445**: 881-885.

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet* **59**: 983-989.

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**: 506-516.

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* **62**: 450-458.

Sullivan PF, Eaves LJ, Kendler KS, Neale MC (2001) Genetic case-control association studies in neuropsychiatry. *Arch. Gen. Psychiat.* **58**: 1015-1024.

Sullivan PF, Neale BM, Neale MC, van den Oord E, Kendler KS (2003) Multipoint and single point non-parametric linkage analysis with imperfect data. *Am. J. Med. Genet. Ser. B* **121**: 89-94.

Sun L, Cox NJ, McPeak MS (2002) A statistical method for identification of polymorphisms that explain a linkage result. *Am J Hum Genet* **70**: 399-411.

Terwilliger JD, Ding Y, Ott J (1992) On the relative importance of marker heterozygosity and intermarker distance in gene mapping. *Genomics* **13**: 951-956.

Thornsberry J, Goodman M, Doebley J, Kresovich S, Nielsen D,Buckler E (2001) Dwarf polymorphisms associate with variation in flowering time. *Nature Genetics* **28**: 286- 289.

Tikhonoff V, Kuznetsova T, Stolarz K, et al. (2003) Beta-adducin polymorphisms, blood pressure, and sodium excretion in three European populations. *Am. J. Hypertens.* **16**: 840-846.

Tom S, Andrew PR (1999) *Human Molecular Genetics,* 2nd ed. John Wiley & Sons Inc.

Tsai SJ, Wang YC, Hong CJ (2001) Allelic variants of the a1a adrenoceptor and the promotor region of the a1a adrenoceptor and temperament factors. *J. Med. Genet.* **105**: 96-98.

Tsai SJ, Wang YC, Chen JY, Hong CJ (2003) Allelic variants of the tryptophan hydroxylase (A218C) and serotonin 1B receptor (A-161T) and personality traits. *Neuropsychobiology* **48**: 68-71.

Tsakiri KD, Cronkhite JT, Kuan PJ, Xing X, Raghu G, Weissler JC, Rosenblatt RL, Shay JM, Garcia CK (2007) Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc Natl Acad Sci USA* **104**: 7552-7.

Van Steen k, Lange C (2005) PBAT: a comprehensive software package for genome-wide association analysis of complex family-based studies. *Hum. Genom.* **2**: 67-69.

Grover VK, Cole DEC, Hamilton DC (2010) Attributing hardy-weinberg disequilibrium to population stratification and genetic association in case-control studies *Annals of Human Genetics* **74**: 7787

Verbeke G, Molenberghs G (2000) *Linear mixed models for longitudinal data*, Springer Series in Statistics, Springer-Verlag, New-York.

Wedderburn RWM (1974). Quasi-likelihood functions, generalized linear models, and the GaussNewton method. *Biometrika* **61 (3)**: 439447.

Weinberg C (1999) Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet* **64**: 1186-1193.

Whittemore A (1996) Genome scanning for linkage: An overview. *Biometrics* **59**: 704-716.

Williams DA (1975) The analysis of binary responses from toxico-logical experiments involving reproduction and teratogenicity, *Biometrics* **31**: 949-952.

Williams JT, Duggirala R, Blangero J (1997) Statistical properties of a variance-components method for quantitative trait linkage analysis in nuclear families and extended pedigrees. *Genet Epidemiol* **14**: 1065-1070.

Wright S (1922) Coefficients of inbreeding and relationship. *The American Naturalist* **56**: 330-338.

Xie X, Ott J (1993) Testing linkage disequilibrium between a disease gene and marker loci. *Am. J. Hum. Genet.* **53**: 1107

Xu GF, O'connell P, Viskochil D, et al. (1990) The neurofibromatosis type 1 gene encodes a protein related to GAP. *Cell* **62**: 599-608.

Young T, Woods M, Parfrey P, Green J, Hefferton D, Davidson D (1999) A founder effect in the Newfoundland population reduces the Bardet-Biedl Syndrome I (BBS1) interval to 1 cM. *Am J Hum Genet*. **65**: 1680-1687.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.* **53**: 79-91.

Zeger SL, Liang KY (1986) Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**: 121-130.

Zelterman D, Chen CF (1988) Homogeneity tests against central-mixture alternatives, *Journal of the American Statistical Association* **83**: 179-182.

Zhao LP, Aragaki C, Hsu L, Quiaoit F (1998) Mapping complex traits by single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **63**: 225-240.

Zhao JH, Lissarrague S, Essioux L, Sham PC (2002) GENECOUNTING: haplotype analysis with missing genotypes. *Bioinformatics.* **18**: 1694-1695.

Zhao XZ, Li HF, Shi YY, et al. (2006) Significant association between the genetic variations in the 5' end of the N-methyl-D-aspartate receptor subunit gene GRIN1 and schizophrenia. *Biol. Psychiat.* **59**: 747-753.

Zollner S, Pritchard JK (2005) Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**: 1071-1092.

Appendix A

Laplace's Method

In mathematics, Laplace's method is a technique used to approximate integrals of the form

$$\int_a^b e^{Mf(x)} dx$$

Assume that the function $f(x)$ has a unique global maximum at x_0. If the limits of integration go from $-\infty$ to $+\infty$, then

$$\int_a^b e^{Mf(x)} dx \approx \sqrt{\frac{2\pi}{M|f(x_0)''|}}\, e^{Mf(x_0)} \text{ as } M \longrightarrow \infty$$

A generalization of this method and extension to arbitrary precision is provided by Fog (2008).

In order to find the value of $\Pr(\mathbf{y}_i \mid \{\mathbf{z}_{ij}\})$, The Laplace approximation goes a follows:

$$\Pr(\mathbf{y}_i \mid \{\mathbf{z}_{ij}\}) = \int_{\mathbb{R}^{n_i}} \Pr(\mathbf{y}_i \mid \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i)\, \mathrm{f}(\boldsymbol{\xi}_i)\, d\boldsymbol{\xi}_i \tag{1.1}$$

define $F_0(\boldsymbol{\xi}_i)$ as

$$F_0(\boldsymbol{\xi}_i) = -\frac{1}{2n_i\vartheta}\, \boldsymbol{\xi}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\xi}_i + \frac{1}{n_i} \log \Pr(\mathbf{y}_i \mid \{\mathbf{z}_{ij}\}, \boldsymbol{\xi}_i) \tag{1.2}$$

then

$$\begin{aligned}\Pr(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}) &= \sqrt{\frac{1}{(2\pi\,\vartheta)^{n_i}\,|\boldsymbol{\Sigma}_i|}}\,\int_{\mathbb{R}^{n_i}} \mathrm{e}^{n_i F_0(\boldsymbol{\xi}_i)}\,d\boldsymbol{\xi}_i \\ &\approx \sqrt{\frac{1}{(n_i\vartheta)^{n_i}\,|\boldsymbol{\Sigma}_i|\,|\mathbf{R}_o|}}\,\mathrm{e}^{n_i F_0(\breve{\boldsymbol{\xi}}_{io})}\end{aligned} \tag{1.3}$$

as $n_i \longrightarrow \infty$, where $\breve{\mathbf{z}}_o$ is the solution to

$$\frac{\partial F_0(\boldsymbol{\xi}_i)}{\partial \boldsymbol{\xi}_i} = \mathbf{0}$$

and $\mathbf{R}_o$ is defined as

$$\mathbf{R}_o = -\frac{\partial^2 F_0(\boldsymbol{\xi}_i)}{\partial \boldsymbol{\xi}_i\,\partial \boldsymbol{\xi}_i'}\Big|_{\boldsymbol{\xi}_i=\breve{\boldsymbol{\xi}}_{io}}$$

i.e., $\breve{\boldsymbol{\xi}}_{io}$ satisfy

$$\breve{\boldsymbol{\xi}}_{io} = \vartheta\,\boldsymbol{\Sigma}_i\,\big(\mathbf{y}_i - \mathrm{E}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{io})\big) \tag{1.4}$$

and

$$\mathbf{R}_o = \frac{1}{n_i\vartheta}\,\boldsymbol{\Sigma}_i^{-1} + \frac{\mathrm{Var}(\mathbf{y}_i \,|\, \{\mathbf{z}_{ij}\}, \breve{\boldsymbol{\xi}}_{io})}{n_i}. \tag{1.5}$$