From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Statistical Genetics Analysis of Family Data

Benjamin Yip

Stockholm 2008

Published by Karolinska Institutet. Printed by

ISBN 978-91-7409-137-3

ABSTRACT

The importance of genetic determinants and risk factors of diseases has been consistently recognized in genetic epidemiology, which is one of the fastest growing areas in genomic medicine. Familial clustering is a common characteristic of genetic related phenotypes, providing vital insights into the etiology of diseases by establishing the relative contribution of genetic and environmental factors. The availability of family data has opened up new opportunities for studying genetic and environmental contributions to diverse diseases. Family data overcome the limitations of statistical power common in twin data analysis, but also enhance the breadth of genetic information.

The generalized linear mixed model has provided a central conceptual framework that allows estimation of the genetic and environmental contributions with adjustment for various epidemiological risk factors. However, estimation often requires high-dimensional integrals to integrate out the random effects and in the models that we considered this is general analytically intractable. Since we have to deal with large datasets with sparse binary outcomes, computation has been another stumbling block in the analysis of realistic models.

This thesis focuses on the analysis of population-based family data, for application in cancer, perinatal diseases and psychiatric disorders. We have closely investigated the marginal and hierarchical-likelihood approaches, and also considered ascertainment approaches for both binary traits and age-at-onset traits. We demonstrate that the newly developed methodologies for the analysis of family data are highly flexible and allow straightforward handling of covariates.

Keywords: Variance component, population-based family data, Generalized linear mixed model, ascertainment, binary trait, quantitative genetics

LIST OF PUBLICATIONS

This thesis is based on the following papers, which will be referred to in the the text by their Roman numerals (I-IV):

I Noh M, **Yip BH**, Lee Y, Pawitan Y. Multicomponent Variance Estimation for Binary Traits in Family-Based Studies. Genetic Epidemiology 30:37-47, 2006.

II **Yip BH**, Björk C, Lichtenstein P, Hultman C, Pawitan Y. Covariance component models for multivariate binary traits in family data analysis. Statistics in Medicine 27:1086-1105, 2008.

III **Yip BH**, Reilly M, Cnattingius S, Pawitan Y. Matched ascertainment of informative families for complex genetic modelling. *Submitted*

IV **Yip BH**, Moger TA, Pawitan Y. Genetic analysis of age-at-onset traits based on case-control family data. *Submitted*

Contents

1 INTRODUCTION

Evidence of familial clustering provides the first indication of a genetic basis for a disease. Table below (from Pawitan et al., 2004) shows the distribution of the number of preeclampsia (PE) occurrences among how many women who have had two or three pregnancies. Among these subjects, 570 of the women were preeclamptic pregnancies. However, we would expect only 68 such cases if PE occurs purely as random a Bernoulli event. This observation also holds for women who have had three pregnancies; the number of repeated PEs exceeds what is predicted for random Bernoulli events. Since most mothers have children with the same fathers, it is clear that we cannot separate the maternal and paternal genetic contributions based only on disease clustering from nuclear families. Furthermore, familial clustering may be due to common environmental effects. In general, separating these effects requires investigation of larger family structures and appropriate models. Analysis is complicated by the diversity of the types of trait commonly encountered in genetic research. These include: (1) Gaussian continuous traits; (2) binary traits; and (3) age-at-onset traits.

Table 1: *Familial clustering of PE, summarized for women who had two or three pregnancies in Sweden between 1987 and 1997. The values under "Random" are the corresponding expected number if PE occurs randomly; this is computed using the estimated binomial probabilities.*

Number pregnancies	Number of PE	Number women	Random
2	0	100590	100088
2	1	4219	5223
2	2	570	68
3	0	20530	20438
3	1	943	1206
3	2	124	24
3	3	21	0

1.1 Binary traits

For Gaussian outcomes, analysis is relatively straightforward with the linear mixed models approach (Guo and Want, 2002; Pawitan, 2001). However, the corresponding methodology for binary traits proves to be more limited, particularly in the handling of extended family structures. Conventionally, quantitative analysis of non-Mendelian binary traits tends to rely heavily on the twin method and methodologies are well-developed for continuous and binary phenotypes (Neale and Car-

don, 1992; Sham 2007). However the low rate of twinning frequently leads to sample-size problems, especially in the study of rare conditions. In contrast, the use of family data not only overcomes the sample-size problems encountered with twin data but also potentially provides richer genetic information (Pawitan et al., 2004). However, in quantitative genetic studies using family data, analytical and computational complexity arises from (i) the correlations between family members and (ii) the large datasets typically required to obtain meaningful results. In Sweden, the well-developed health system and good traditions of handling registry data make such analysis feasible. With such valuable resources, the development of a statistical model to handle large volume of family data is much desired.

Arbitrary family types can be structured by specifying the a correlation matrix of random effects and adopt it into the framework of generalized linear mixed models (GLMM). GLMM can be used extensively in many applications in genetic epidemiology, since they accommodate various outcomes by a flexible specification of the link function and response-variable distribution. However, this often requires high-dimensional integrals to integrate out the random effects and the process is in general analytically intractable and computationally intensive. To avoid the need for extensive integration across the joint distribution of random effects, approximation methods have been suggested. Schall (1991), Breslow and Clayton (1993) proposed the use of restricted maximum likelihood (REML) to estimate parameters in GLMM by assuming normal random effects. However, estimation can be severelu biased. Various approximation methods have been proposed by Drum and McCullagh (1993), Shun and McCullagh (1995), Lee and Nelder (1996), Lin and Breslow (1995) and Shun (1997) . However, for binary data these approximation methods have been criticized for giving biased estimates. To overcome the bias, various simulation methods such as Monte Carlo EM (MCEM) (Chen et al., 2002; Mcculloch, 1994), Monte Carlo Newton Raphson (MCNR), simulated ML method (Mcculloch, 1997) and the Gibbs sampling (Zeger and Karim, 1991) have been proposed. All of these simulation-based methods are computationally intensive and can result in incorrect estimates, which may go undetected (Hobert and Casella, 1996).

1.2 Comorbidity

Certain diseases have similarities in their epidemiological features, risk factor patterns and intermediate phenotypes, which strongly indicate that there might be a common etiology. To investigate the common genetic and environmental factors driving the comorbidity of diseases, a statistical model for a multivariate binary

traits (MBT) is required. Nevertheless, in MBT analysis, the dimension of the parameter space increases (at least doubles) and using the Monte Carlo approximation in the computation of the likelihood can lead to computational hurdles. The high-dimensional parameter space, combined with the random events in likelihood computation, makes it more likely that the optimization algorithm might get stuck in a local maximum. Thus, determining good starting values is a crucial and necessary step in the optimization procedure.

1.3 Ascertainment

For rare diseases, large population-based family data is required to observe some disease clustering in families. Such data are obtained primarily from the linkage of population-based registers. Due to the size of the registers, handling such a vast amount of data is computationally very time-consuming. For examples of intensive computation in familial survival data analysis, see Moger et al. (2008).

While ascertainment methods have traditionally been used for efficient data collection, it is clear that they can also be useful for computational purposes. If disease prevalence is low, say 1%, it seems inefficient to randomly sample 10,000 individuals in order to obtain 100 affected cases. Instead, it is more convenient to collect data from families with at least one affected member. Intuitively, these 'genetically-loaded' families contain most of the information about the genetic properties of the disease. Thus, it is more efficient to ascertain these families rather than the whole cohort. For example, Moger et al. (2008) suggested case-cohort methods as a way of dealing with survival traits, especially in large population-based family data.

Non-random ascertainment is commonly used in genetic research to maximize the amount of information in the data for a given sample size (Elstion and Soble, 1979). It has also been suggested for variance components models (Epstein et al., 2002; Andrade and Amos, 2008; Burton, 2003). The most common method of non-random ascertainment in family studies is to include families with at least one affected member, but this may not be optimal. From a genetic model perspective, families with at least *two* affected members are more 'genetically-loaded' than families with only one affected member. Thus, the definition of case families needs to be reexamined.

1.4 Age-at-onset traits

Age-at-onset traits, such as cancer occurrence, are often of interest in genetic studies, but the general family-based methodology to disentangle the genetic and environmental effects has its own limitations. Traditionally, frailty models have been used for correlated survival data, particularly for twins or full siblings or other exchangeable structures (Klein et al., 1999; Lambert et al., 2004). A frailty variable describes the unobserved random variation and creates dependence between lifetimes. Another approach using the copula models, integrates out the frailty and acts more like a marginal model. The difference between the two models is that the copula model provides a population average effect for the covariates, while in a frailty model the regression coefficients are calculated conditional on the value of the frailty variable (Hougaard, 2000). However, when the dependence structure goes beyond full siblings or twins, the likelihood and estimation for frailty models becomes very complicated (Ha and Lee, 2005; Klein et al., 1999; Lambert et al., 2004). An alternative to frailty models is the accelerated-failure time (AFT) model where fixed and random effects act linearly on the log-survival time. Although the AFT model provides several advantages (Ha and Lee, 2005; Klein et al., 1999; Lambert et al., 2004), it has received relatively little attention in the analysis of correlated survival data. Again, this is partly due to the intractable nature of the integration required to obtain the marginal likelihood.

1.5 Genetic concepts

To understand our effort in the general context of genetic epidemiology, it is worthwhile to outline the standard steps in genetic analysis of a disease.

- In a quantitative genetic analysis we first establish if a genetic effect is present, by analyzing the co-occurrence of a phenotype among family members.
- A quantitative genetic analysis tells us whether or not a condition is genetic, but it does not tell us what genes are involved and where they are in the genome. For this, a linkage analysis is needed, where some markers can be genotyped and correlated with disease occurrence. Linkage studies are performed on families.
- An association study attempts to establish association between phenotypes and genotypes, with the same purpose of finding the genes involved in a

disease as in linkage analysis, but it is typically performed on unrelated individuals. Both linkage and association studies are also called gene mapping studies.

Quantitative genetic analysis is a necessary first step in establishing the genetic basis of a disease. The methods we describe in this thesis cover only quantitative genetic analysis only. Linkage analysis requires much more detailed probabilistic modeling of the gene transmission from parents to offspring, and is beyond the scope of this thesis. However, the mixed model method is also used in linkage analysis, for example, in quantitative trait linkage analysis (Amos, 1994; Blangero et al., 2001).

All of the phenotypes which we study are the so-called complex or non-Mendelian phenotypes. A Mendelian phenotype is determined by one or two genes that have strong effects, sd that the genotype of a person can be inferred simply by looking at the occurrences of the phenotype inside a family (e.g. Huntington's disease). In contrast, non-Mendelian phenotypes are potentially determined by many genes, each with typically small effects, and possibly also by the environment (Fisher, 1918; Falconer, 1965).

2 AIMS

This work was inspired by the two main challenges that we face in quantitative genetic analysis using family data, namely, statistical inference and efficient computation. Both of these are subject to the problem of intensive computation time, especially during the process of model building. The general aim of this thesis is to develop efficient and flexible statistical models for family data analysis. The specific aims are:

- To develop a methodology for joint analysis of the association of a single binary trait with standard epidemiological risk factors and with genetic and environmental effects;
- To develop a methodology for joint analysis of multivariate binary traits;
- To develop a methodology for analysis of ascertained family data;
- To develop a methodology for family-based survival analysis (age-at-onset traits).

3 MATERIALS

The studies in this thesis were all based on data from Swedish population-based registers. By combining different registers and using the unique personal identification number, one can construct a linkage of data that provides a opportunity to study whether a disease shows a significant familial aggregation, and whether this aggregation can be explained by some covariates. Here, three registers, used in all four papers mentioned below, are briefly described.

The Multi-Generation Register (All papers)

The Multi-Generation Register was created by different data sources and contains information of the recorded (or index) person identity number and his/her first degree relatives for residents in Sweden from 1932 or later. To be included in the register, the index person had to be alive in 1960 or born thereafter. In 2002, the registry contained around 9 million index persons and a total of 13.5 million individuals.

The Hospital Discharge Register

The Hospital Discharge Register became nationwide in 1973, held by the National Board of Health and Welfare, contains data on all inpatient care in Sweden with nationwide coverage for psychiatric diagnosis. For all individuals, the primary discharge diagnosis (and secondary diagnoses if applicable) was assigned by the treating physician and recorded according to the International Classification of Diseases: eighth revision (ICD-8) through 1986, ninth revision (ICD-9) from 1987 to 1996 and 10th revision (ICD-10) from 1997 to 2001 (Lichtenstein et al., 2006). The diagnostic assessment is then forwarded to the Hospital Discharge Register with standardized algorithms across Sweden. The psychiatric diagnoses have been given conservatively, and validation studies have confirmed a very low rate of false-positive diagnoses (Ekholm et al., 2005) . Family members were not diagnosed at a specific time by one psychiatrist, but independently at their hospital discharge across Sweden. Repeated diagnoses were likely performed by independent psychiatrists, especially if the time interval between episodes was large. Therefore in Paper III, to increase the diagnostic specificity, we only defined individuals diagnosed at least twice for the same psychiatric disorder (e.g. schizophrenia or bipolar disorder) as cases.

The Swedish Cancer Registry (Paper IV)

The population-based Swedish Caner Registry (SCR) was established in 1958. It contains individual data on all malignant tumours newly diagnosed within Sweden (*http://www.sos.se/epc/cancereng.html*). Tumours are reported to the SCR separately by both the diagnosing clinicians and the responsible pathologists or cytologists. Nearly 100% of all diagnosed cancers were reported, with histological verification of 97% of the tumours. For malignant tumours, the registration rate has always been high. For benign and *in situ* conditions, there was an initial trend of underreporting, but reporting improved after the 1960s.

Other registers that were also used for the applications are *The Cause of Death Register* (all Papers) and *The Medical Birth Register* (Papers I and III).

4 STATISTICAL MODELS

Traditionally, quantitative analysis of non-Mendelian traits exhibits a heavy dependence on twins data. Related methodologies are well-developed for continuous and binary phenotypes (Neale and Cardon, 1992; Sham, 2007), but the low rate of twinning often leads to sample-size problems. For example, in the study of preeclampsia (PE), Ros et al. (2000) ascertained the pregnancy outcomes of 917 monozygotic (MZ) female twinpairs and 1199 dizycotic (DZ) pairs in the Swedish birth registry from 1973 to 1997. While there is an indication of genetic heritability of 0.57, the result is not significant with a wide 95% confidence interval (0.0 to 0.71). There is clearly a lack of power in this analysis even though PE is not a rare condition, with a prevalence of around 4% in all pregnancies (Table 1). Predictably, the development of many current methods is focused on analyzing the full family data directly. Furthermore, family data also potentially provide richer genetic information (Pawitan et al., 2004), which will be discussed further in later sections.

4.1 Liability-threshold model

The family-based analysis of genetic and environmental contributions to continuous or Gaussian traits is straightforward using the linear mixed models approach (Laird and Ware, 1982; Guo and Wang, 2002), whereas the corresponding analysis of complex binary traits for family data remains rather limited. In the latter, it is common to assume the *liability-threshold* model, introduced by Pearson and Lee (1901). The liability-threshold model postulates an unobserved normal liability π, which generally assumes to have a normal distribution with mean 0 and variance 1 in the general population. The observed binary phenotype (e.g. disease outcome) y is assumed to be present in all individuals whose liability is above a certain *threshold* level, t, and to be absent in all other individuals. In a population-based family study, the level of the threshold can be easily estimated from the population frequency, p, of the disease. Let $F(t)$ be a distribution function of the threshold. Then the value of the threshold t is such that

$$F(t) = 1 - p \tag{1}$$

which is equivalent to

$$t = F^{-1}(1 - p) \tag{2}$$

where $F^{-1}$is the inverse of F. Since the assumption of the liability is $\varphi_i \sim N(0,1)$, then F is simply the standard normal distribution (i.e probit) function Φ. Let φ_i be the liability for individual i, if $t < \Phi^{-1}(1-\varphi_i)$ then this individual by definition will have the disease.

4.2 Twins model

The concept of liability-threshold model fits naturally into the framework of generalized linear mixed models (GLMMs). GLMM can be used extensively for many applications in genetic epidemiology, as it accommodates various outcomes by a flexible specification of the link function and response-variable distribution. Since the liability-threshold model is traditionally based on twins study, demonstration of the general GLMM model by considering twin data is first presented.

Let (y_{i1}, y_{i2}) be the binary phenotypes of interest, measured from twin pair i. The simplest model assumes that y_{ij} is a Bernoulli outcome with probability π_{ij} and

$$\eta(\pi_{ij}) = \beta + g_{ij} + c_{ij} \tag{3}$$

where $\eta()$ is the link function, β is a fixed model parameter associated with prevalence, the additive genetic effect g_{ij} is assumed to be $N(0, \sigma_g^2)$ and the common childhood environment effect c_{ij} is $N(0, \sigma_c^2)$. In the example of liability-threshold model, we assume the distribution of liability is $N(0,1)$, thus the link function $\eta() = \Phi^{-1}()$, so that

$$\pi_{ij} = \Phi(\beta + g_{ij} + c_{ij}). \tag{4}$$

In GLMMs, the canonical link function for Bernoulli outcomes is the logit and not the probit. It is analytically more convenient in deriving the computational algorithm by using the canonical link function. But in biometric genetics applications (e.g. liability-threshold model), the probit link is preferred. Another motivation to use the probit link is that it will facilitate computation of the marginal likelihood in terms of multivariate normal probabilities (Pawitan et al., 2004). However, as demonstrated in Paper I, the logit and probit models can be easily compared after adjustment by a simple scale factor (Noh et al., 2006).

Let $g_i = (g_{i1}, g_{i2})$ and $c_i = (c_{i1}, c_{i2})$ and the random effects are independent of each other. Between-pair genetic effects are independent, but within-pair values are not. For MZ twins it is commonly assumed that the

$$cor(g_{i1}, g_{i2}) = 1,$$

$$cor(c_{i1}, c_{i2}) = 1.$$

For DZ twins

$$\begin{aligned} cor(g_{i1}, g_{i2}) &= 0.5, \\ cor(c_{i1}, c_{i2}) &= 1. \end{aligned}$$

This is the main idea in quantitative genetics studies using twins or family data: *The discrepancy in correlations between family members (twins, full-siblings, half-siblings, parent-child etc) allows us to separate the genetic from the common environmental factor*s. Thus, family data are potentially richer in genetic/environmental information than twins data, we shall see in the next section.

Heritability

For the purpose of interpretation, it is convenient to define the quantity

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_c^2 + 1},$$

known as the *narrow heritability*. Since we assume the probit link (which explains the "1" in the denominator), the heritability measures the proportion of the variance (of liability or predisposition to the phenotype under study) due to additive genetic effects. The additive genetic effect assumes that some alleles contribute a fixed value to the metric value of quantitative value. Another heritability quantity is the so-called broad-sense heritability and reflects all possible genetic contributions, such as dominance variation and multi-genes interaction, to a population's phenotypic variance. Quantitative studies generally deal with non-Mendelian diseases and assume absence of dominant genetic effect and model the genetic effect only in terms of the additive genetic effect.

4.3 Family model

4.3.1 Single binary traits (SBT)

For the next order of complexity after the twins model, suppose we have an arbitrary family structure i (not necessary a nuclear family), let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$ be the outcome vector measured on family members, and conditional on random

effects v_i, y_{ij} assumes to be independent Bernoulli with parameters π_{ij} so that $E(y_{ij}|v_i) = \pi_{ij}$. Assume the probit link function to have

$$\Phi^{-1}(\pi_{ij}) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{v}_i \tag{5}$$

where $\boldsymbol{x}_{ij}$ is a vector of fixed covariates, and $\boldsymbol{\beta}$ a vector of fixed-effect parameters. The random effects $\boldsymbol{v}_i$ captures all random effects that are potentially responsible for family aggregation of the disease and $\boldsymbol{z}_{ij}$ is a known vector describing how these effects shared by family members. Generally, $\boldsymbol{v}_i$ is assumed to be multivariate normal distributed with mean zeroes, and covariance matrix $D_i(\boldsymbol{\theta})$. The parameters of interest in (5) are the regression parameters $\boldsymbol{\beta}$ and variance-covariance components in vector $\boldsymbol{\theta}$.

To help model specification and estimation, the random effects $\boldsymbol{v}_i$ can be separated into genetic and environmental effects. Let $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{in_i})^{'}$ and we can rewrite model (5) as

$$\Phi^{-1}(\boldsymbol{\pi}_i) = \boldsymbol{x}_i^{'}\boldsymbol{\beta} + \boldsymbol{g}_i + \boldsymbol{f}_i + \boldsymbol{c}_i, \tag{6}$$

where the random effect, $\boldsymbol{f}_i$, is the shared family environmental effect and $(\boldsymbol{g}_i, \boldsymbol{c}_i)$ are the genetic and childhood environmental effects, as mentioned earlier. Generally, the random effects assumed to be multivariate normal distributed:

$$\begin{aligned} \boldsymbol{g}_i &\sim MVN(0, \Sigma_g) \\ \boldsymbol{f}_i &\sim MVN(0, \Sigma_f) \\ \boldsymbol{c}_i &\sim MVN(0, \Sigma_c). \end{aligned}$$

In (6), the random effects are assumed to be independent, so $\boldsymbol{v}_i = \boldsymbol{g}_i + \boldsymbol{f}_i + \boldsymbol{c}_i$ and

$$\begin{aligned} D_i(\boldsymbol{\theta}) &= \Sigma_g + \Sigma_f + \Sigma_c \\ &= \sigma_g^2 R_g + \sigma_f^2 R_f + \sigma_c^2 R_c. \end{aligned}$$

where R_k, k=(g, f, c), is the correlation matrix. Various correlations exist within each effect as specified by the covariance matrices, but the effects are assumed to be independent of each other and between families.

Correlation assumptions

Table 2 illustrates the general correlation assumptions between different relationships. On average, first degree of relatives share half of their segregating genes

(Sham, 2007). The environmental correlation depends on the structure of the family data. For example, in nuclear family data (two parents and their offspring), the most common assumption is that there is a common environmental effect, $\boldsymbol{f}$, shared by all family members and a childhood environmental effect, $\boldsymbol{c}$, shared only between siblings. Economics status, diet, living environment or a combination of these, are for example, assumed to be shared by all family members. But in some cases the genetic-family-childhood (gfc) mixed model may not be relevant. For example, there are no bipolar disorder studies that suggest any common environmental effect shared between parents and offspring. Instead, we have found in the schizophrenia and bipolar disorder study (paper II) that spouses are highly correlated in disease occurrence and indicate adult-only environmental effect (a). One potential explanation for such effect is diagnosed patients tend to mate with other diagnosed patients, this phenomena is the so-called assortative mating (Neale and Cardon, 1992).

Table 2: *Genetic and environmental correlations in families.*

	Correlation			
Relationship	Genetic	Childhood	Adult	Family
Parent-child	0.5	0	0	1
Sib-sib	0.5	1	0	1
Spouse-Spouse	0	0	1	1
Half-sibs, paternal	0.25	0	0	0
Haf-sibs, maternal	0.25	1	0	1
Cousins	0.125	0	0	0

Family data not only help overcome the sample size problem in twins study, but also capture more genetic information. In the study of PE, Pawitan et al. (2004) used models that account for maternal effect (m) and foetal genetic effect, and showed that these effects could be readily estimated by using family data. The family data which they used contained pairs of families in which, the mothers are full sisters and the fathers are unrelated men. The outcome vector is the pregnancies of the women. Consequently, the paternal effect was found to be expressed as the foetal effect. The GLMM model is then expressed as:

$$\Phi^{-1}(\boldsymbol{\pi}_i) = \boldsymbol{X}_i^{'}\boldsymbol{\beta} + \boldsymbol{m}_i + \boldsymbol{g}_i + \boldsymbol{f}_i + \boldsymbol{s}_i$$

where $\boldsymbol{m}_i$ is the maternal effect, $\boldsymbol{f}_i$ is the common family environmental effect, and $\boldsymbol{s}_i$ is the common sibling environmental effect. Note, since the underlying correlation assumption of the foetal effect is same as the genetic effect, $\boldsymbol{g}_i$ is also used for the foetal effect. The common family environment is the unique environment cre-

ated by the father and mother, and sibling environment is the common childhood environment experienced by the sisters. To illustrate the differences in the correlation matrices for the random effects, let each sister from the sib-pairs families had two pregnancies, so that the outcome $\boldsymbol{y}_i$ is a binary vector of length 4, where the first two entries are the pregnancies of the first sister, while the the last two entries are the other sister.

$$R_m = \begin{pmatrix} 1 & 1 & 0.5 & 0.5 \\ 1 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 1 \\ 0.5 & 0.5 & 1 & 1 \end{pmatrix}, \quad R_g = \begin{pmatrix} 1 & 0.5 & 0.125 & 0.125 \\ 0.5 & 1 & 0.125 & 0.125 \\ 0.125 & 0.125 & 1 & 0.5 \\ 0.125 & 0.125 & 0.5 & 1 \end{pmatrix},$$

$$R_f = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad R_s = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

The (1,2)-elements of R_m is equal to 1 since the first two pregnancies come from the same mother. The (1,2)-element of R_g is 0.5 since the first two foetuses are full siblings and the (1,3)-element of R_g is 0.125 since it refers to a cousin pair. Similar reasoning applies to R_f and R_s. This illustrates how well GLMM can handle arbitrary family data by constructing proper correlation matrices.

4.3.2 Multiple binary traits (MBT)

Analysis for multiple binary traits (MBT) is a natural extension of the analysis of SBT. Certain clinical conditions, such as schizophrenia and bipolar disorder, have similarities in epidemiological features, risk factor patterns and intermediate phenotypes. This strongly suggests that such disorders may share common genetic loci or common environmental risk factors. To achieve a better etiological understanding, it is therefore important to investigate the common genetic and environmental factors driving the comorbidity of the diseases.

For an arbitrary family structure i consisting of n_i members, let $\boldsymbol{y}_{ik}$=$\{y_{i1k} \dots y_{in_ik}\}$ be the binary outcome vector of disease k. In Paper II k =s, b where s stands for schizophrenia and b bipolar disorder. Conditional on random effects, the model assumes that y_{ijk} is Bernoulli with parameter π_{ijk}. Then the general MBT model can be written as:

$$\eta(\boldsymbol{\pi}_{ijk}) = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{g}_i + \boldsymbol{f}_i + \boldsymbol{c}_i + \boldsymbol{u}_i, \tag{7}$$

where the newly defined $\boldsymbol{u}_i$ is the unshared environmental effect. The unshared ef-

fect represent environmental exposures that are not shared by other family members, e.g. smoking habits, prenatal environment, head injury, somatic diseases, drug abuse, etc. Again, we assume that the random effects are multivariate normal distributed with means equal to zeroes and with the corresponding covariance matrices Σ_g, Σ_f, Σ_c, Σ_u. The dimension of the covariance matrices are doubled compared to the covariance matrices in (6). For example, the genetic covariance matrix

$$\Sigma_g = \begin{pmatrix} \sigma^2_{sg} R_g & \sigma_{sbg} R_g \\ \sigma_{sbg} R_g & \sigma^2_{bg} R_g \end{pmatrix},$$

where R_g is the same correlation matrix used in SBT analysis. The off-diagonal block is the covariance structure between the two disorders. The covariance components σ_{sbg} explain the covariance between schizophrenia and bipolar disorder. In other words, they explain the genetic variance of the comorbidity of the disorders. When >0, then and 1- can be interpreted as the proportions of schizophrenia genetic variability that are explained by common genetic loci and unique genetic loci (i.e. schizophrenia only), respectively. If the two disorders do not share any genetic loci, then σ_{sbg} is expected to be zero (or close to zero). Similar interpretations can be applied to other random effects.

The main challenge for MBT analysis is the high dimensional nature of variance-covariance components that need to be estimated. Determining good starting values for the optimization procedure is a critical step (especially if the MBT outcomes are rare) and will be explained further in the later section.

4.3.3 Age-at-onset traits

Conventionally, frailty models are used for correlated survival data (Aalen and Tretli, 1999) , but they generally do not accommodate an arbitrary family structure. As an alternative to frailty models, the accelerated-failure time (AFT) model is considered, wherein the fixed and random effects act linearly on the log-survival time. The AFT models presented in this thesis capture the usual genetic and environmental components. It also allows the data to be both left truncated and right censored (LTRC). To emphasize the mixed effects, we will use the term MAFT to mean 'mixed AFT model'.

MAFT specifies a direct relationship between survival time and covariates including fixed and random effects. Let T_{ij} be the survival time for the jth member of the ith family, for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. The survival time T_{ij} are only partially observed due to the left truncation L_{ij} and the right censoring F_{ij}. In general, the

left truncation is defined as the difference in time between start of follow up and birth date. We make the following assumptions:

1. L_{ij} and F_{ij} are assume to be *conditionally independent*, given the random effects (Lai and Ying, 1994).

2. Given the random effects, F_{ij} are *noninformative* about the random effects.

Let $y_{ij} \equiv \min(\log T_{ij}, \log F_{ij})$ be the logarithm of the possibly-censored survival time, $\delta_{ij} \equiv I\left(\log T_{ij} \leq \log F_{ij}\right)$ be the event indicator, and $b_{ij} \equiv \log L_{ij}$. Thus we consider the MAFT

$$\begin{aligned} \log T_{ij} &= \mu_{ij} + e_{ij} \quad (8) \\ \mu_{ij} &= \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + g_{ij} + f_{ij} + c_{ij} \quad (9) \end{aligned}$$

where $x_{ij} = (x_{ij1}, ..., x_{ijp})$ is a vector of known fixed covariates, β a p-vector of regression parameters, g_{ij}, f_{ij} and c_{ij} are the random effects of the *gfc* model. We assume $e_{ij} \sim N(0, \sigma_e^2)$. The random effects are independent between the families, but dependent within the family according the relationship between family members. Model (8) permits a straightforward interpretation; the fixed and random effects are acting linearly on the log of survival time, while frailty models work on the hazard scale.

For the analysis of nuclear families, frailty models are more complex than MAFT models. For example, following the additive model by Korsgaard and Andersen (1998) for a family of parents and one child, four independent frailty terms are needed to capture genetic (two terms), individual (one term) and common environmental (one term) effects. More terms may be needed for larger numbers of children, and these create very complicated formulae for the exact marginal likelihood. By comparison, MAFT is conceptually the same for an arbitrary family structure.

5 LIKELIHOOD INFERENCE

All the models presented in the previous sections require intractable integration or approximation approaches to obtain the marginal likelihood. For instance, although MAFT has simple interpretation, it has received relatively little attention in the likelihood analysis of correlated survival data, due to the intractable integration. To avoid the integration, many methodologies rely either on Monte Carlo approximation of high-dimensional integrals or Laplace approximation.

5.1 Standard marginal likelihood: using Monte Carlo approximation

The Monte Carlo approximation is done either by using the standard marginal likelihood approach (Pawitan et al., 2004), or by using the Gibbs sampling algorithm in the Bayesian approach (Zeger and Karim, 1991; Burton et al., 1999). Chan and Kuk (1997) used a hybrid expectation-maximization (EM)-Gibbs-sampling algorithm, where the E-step computation by Gibbs sampling slows the algorithm. More recently, Rabe-Hesketh et at. (2002) developed an adaptive Gauss-Hermite quadrature (GHQ) method to fit GLMM, but there has been no study showing the feasibility of this approach for large family data.

In quantitative genetic analysis, large data sets are often required, since many traits of interest have low prevalence. For example, in the analysis of preeclampsia data by Pawitan et al (2004), more than 475,000 families were included. In practice, Monte Carlo methods are often computationally too intensive to handle large data sets with general fixed-effect predictors. While the Gibbs sampling approach is flexible, it is prohibitively slow with a large number of families. This is because of one key step in the algorithm that requires sampling from the conditional distribution of random-effect parameters representing the families. Furthermore, as reported by Zeger and Karim (1991), successive samples of random effects tend to be highly correlated, so convergence can be very slow.

In Paper II and Paper III, the standard marginal likelihood approach was used. In general, using the probit model $\Phi^{-1}(p_{ij}) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}^{'}{}_{ij}\boldsymbol{v}_i$, we have

$$\begin{aligned} p_{ij} &= P(Z_j < \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}^{'}{}_{ij}\boldsymbol{v}_i) \\ &= P(Z_j - \boldsymbol{z}_{ij}^{'}\boldsymbol{v}_i < \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}), \end{aligned}$$

where the $\boldsymbol{Z}_j$'s are independent standard normal variates and $\boldsymbol{v}_i$ is the composite random effect with mean zero and variance $D_i(\boldsymbol{\theta})$. Thus we have the marginal probability

$$p(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \int p(\boldsymbol{y}_i|\boldsymbol{v}_i)|D_i(\boldsymbol{\theta}))|^{-q/2} \exp\{-\frac{1}{2}\boldsymbol{v}_i' D_i(\boldsymbol{\theta})^{-1}\boldsymbol{v}_i\} d\boldsymbol{v}_i \quad (10)$$

$$= E_{v_i}\{\prod_j p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}\} \quad (11)$$

$$= P(l_{ij} < W_{ij} < u_{ij}, \text{ for all } j), \quad (12)$$

where q is the dimension of $\boldsymbol{v}_i$, and $W_{ij} \equiv Z_j - \boldsymbol{z}_{ij}^{'}\boldsymbol{v}_i$. The vector $W_i \equiv (W_{i1}, \ldots, W_{in_i})$

is $N(0, \Sigma_i)$ with

$$\Sigma_i = \boldsymbol{z}_i D_i(\boldsymbol{\theta}) \boldsymbol{z}_i' + I_i,$$

where $\boldsymbol{z}_i$ denotes the matrix obtained by stacking the row vectors $\boldsymbol{z}_{ij}$, and I_i is the $n_i \times n_i$ identity matrix. The upper bound $u_{ij} = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}$ if $y_{ij} = 1$, and $u_{ij} = \infty$ if $y_{ij} = 0$. Similarly, the lower bound $l_{ij} = -\infty$ if $y_{ij} = 1$, and $l_{ij} = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}$ if $y_{ij} = 0$. Computation of the normal probability (12) is done using a Monte-Carlo algorithm (Genz, 1993).

Let $f_i(\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}) \equiv P(\boldsymbol{Y}_i = \boldsymbol{y}_i)$, where we make the parameters explicit; the total log-likelihood would then be

$$l = \sum_{i=1}^{N} \log f_i(\boldsymbol{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}). \tag{13}$$

Estimates of the parameters of interest can then be obtained by maximization of (13), as function of $(\boldsymbol{\beta}, \boldsymbol{\theta})$.

In general, we are dealing with population-based databases, where N is extremely large (e.g. in Paper III, N consist of 325,000 family-pairs). Since the evaluation of each probability requires a non-trivial Monte-Carlo approximation, a naive approach is out of the question. Additionally, the likelihood may not be smooth and requires to use derivative-free software for optimization.

5.1.1 Determination of good starting values

The high-dimensional parameter space (especially one as illustrated in Paper II where we were dealing with multiple-binary traits) combined with the noise in the likelihood computation causes the optimization algorithm to get stuck easily in local maxima. We found in our simulations that good starting values are crucial. In practice, it is difficult to guess a good starting value for all the parameters. We have therefore developed a modified method of moments for obtaining good starting values

Our data provide information about variance and covariance components in the form of concordances, measured as odds ratios (ORs), for different family members. In principle, a high OR between parents and children may indicate a common genetic effect. This means that there is some mapping between the variance (and covariance) components with the ORs for different family members. Although the nonlinear mapping between the variance parameters and the ORs is not available

analytically, it can be obtained with good precision by using simulations (see Table 3 in paper III).

Given the fixed effects in the probit model, multiple datasets were simulated by using different combinations of variance-covariance-components. ORs of different types were obtained for each simulated data. Then a quadratic-loss function based on the log-ORs was used:

$$Q(\boldsymbol{\theta}) = \sum_{j=1}^{K} \frac{(\log(\mathrm{OR}_j^{obs}) - \log(\mathrm{OR}_j^{mod}))^2}{\mathrm{var}(\log(\mathrm{OR}_j^{obs}))} \tag{14}$$

where OR^{obs} are the observed ORs from the real data and OR^{mod} are the ORs from the probit model computed from the simulated data, where different settings of $\boldsymbol{\theta} = (\sigma_1^2, ..., \sigma_q^2)$ were used. K denotes the number of ORs that can be obtained in one particular family data. For example, in nuclear family data with m members and each of them containing outcomes from two diseases, $K = m(2m-1)$. In principle, we can compute $Q(\boldsymbol{\theta})$ by using Monte Carlo simulation, and we can thus optimize it. Once (14) is minimized, we anticipate that the best parameter configuration are those which produce model-based ORs closest to the observed ORs. This set of parameters will then be used as the starting values for the optimization of the total log-likelihood.

5.1.2 Grouped family data

The likelihood computation of (13) will obviously be faster if the data are grouped according to the configurations of the family outcomes and covariates $\{\boldsymbol{y}_i, \boldsymbol{x}_i\}$. The total likelihood can then be written as

$$l = \sum_{j=1}^{M} w_j \log f_j(\boldsymbol{y}_j, \boldsymbol{x}_j, \boldsymbol{\beta}, \boldsymbol{\theta}) \tag{15}$$

where w_j is the number of families with the same configuration j, and M is the total number of configurations,

If the family data consist of information on binary outcomes and p binary covariates from k family members, then $M \leq 2^{k(p+1)}$, and the number of probability computations can be reduced by a factor of N/M. However, for analyses of families with up to 4 members, this grouping will substantially reduce the computation time only when we use one or two covariate(s). As we increase the number of covariates, M increases rapidly, so even the grouped data become too large to be analyzed. If information of general covariates (age, smoking status, gender etc) is

available then in order to make the analysis manageable, ascertainment - a form of data reduction, is required.

5.1.3 Ascertained data

Intuitively, one may expect that information about familial clustering comes from families with at least two affected members (see the previous section about ORs). If the prevalence of the disease is low, then only a small fraction of families will be informative, even when population based-data were used. The following table, from Paper III, shows the distribution of the number of small gestational age (SGA) births observed among the pregnancies within the families of sibpairs (sister-sister, brother-brother and bother-sister pairs). There are only 1,055+58+3=1,116 sibpairs (0.3% of the total), that had SGA more than once.

Number of SGAs	0	1	2	3	4
Number of family pairs	306,706	18,807	1,055	58	3

Hence, if the genetic information in the full data can be preserved, an ascertainment of families with at least two affected members can offer dramatic data reduction. It is more efficient to ascertain these families rather than the whole cohort.

For the grouped data, ascertainment is naturally done on the family configurations rather than on the family units. Let $S = S_0 \cup S_1$ be a set of all family configurations, and can be divided into two disjoint sets, where S_1 is the set of all families with at least k affected members, and S_0 is the set of control families. In line with the usual case-control studies, we suggest that all case-family configurations be kept. Control-family configurations are in general included with probability less than one.

Exact and weighted likelihoods Let $A_j = 1$ if family j is ascertained, and 0 otherwise, and $a_j = P(A_j = 1|Y_j = y_j)$. Typically a_j is a function of the number of affected members, but it can also be a function of covariates. Then, the exact ascertainment-adjusted likelihood contribution from an observed $\boldsymbol{y}_j$ is

$$P(Y_j = y_j|x_j, A_j = 1) = \frac{a_j P(Y_j = y_j|x_j)}{\sum_k a_k P(Y_k = y_k|x_j)},$$

where k runs over all possible configurations from the same covariate x_j, such that $\sum_k P(Y_k = y_k|x_j) = 1$. Note that the denominator requires the evaluation of

probabilities for all families that may get ascertained, even if many of the families are in fact unobserved in the data. Consequently, the computational burden of the exact likelihood remains too demanding for routine analysis.

To this end, we instead consider a weighted-likelihood, or pseudo-likelihood

$$\hat{\ell} = \sum_{j=1}^{M} \frac{A_j}{P(A_j = 1 | Y_j = y_j)} w_j \log f_j(y_j, x_j, \boldsymbol{\theta}, \boldsymbol{\beta}), \tag{16}$$

which is clearly an unbiased estimate of the log-likelihood (15). The main advantage over the exact likelihood is that we only need to evaluate the probabilities for family configurations that are both observed and ascertained. The optimization can then be done in the same way as previously described.

Standard inference in the pseudo-likelihood framework typically relies on the asymptotic normality of the estimates, with the so-called sandwich formula for the variance (Kalbfleisch and Lawless, 1988). Unfortunately, for our problem, deriving the sandwich formula analytically proved complicated. Instead, the bootstrap method was used to compute the standard errors for grouped family data.

5.1.4 Optimal matching

We found in Paper III that sampled data obtained by the family case-control design produces estimates that are ofthen far from the full data estimates, although the weighted-likelihood (16) was used. This is because the full data are grouped so that the sampled data are often too different from the full data with respect to certain features that reflect the parameters of interest. Optimal matching, a new ascertainment scheme, was therefore suggested.

The vector of unknown parameters can be divided into two groups: regression parameters and variance components. Hence, there are two types of statistics that are natural for matching.

- Estimates from an ordinary GLM (without random effects); and
- ORs (between family members) that capture familial risk.

The basic idea is simple: we try to ascertain "balanced" samples, where the balance is determined by the ORs and regression coefficient estimates in the GLM using full data. As we have shown previously that ORs describing risk in relatives are good proxy measures of the magnitude of variance components. If sampled data

have similar ORs and regression coefficient estimates compared to the full data, then we would expect the estimates of variance component and coefficient to be of the same magnitude.

5.2 Hierarchical likelihood

The use of the hierarchical likelihood (h-likelihood) approach to analysis of family data is highly flexible and allows straightforward handling of covariates. Less-than-full-likelihood or approximate-likelihood approaches for analysis of clustered binary outcomes, such as the penalized quasi-likelihood (PQL) method by Breslow and Clayton (1993), are known to produce biases. Such biases can be severe when the cluster size is small. Bias-correction method-for-PQL (CPQL) estimators were proposed by Lin and Breslow (1995), but they still could not eliminate biases. For GLMMs, the h-likelihood of Lee and Nelder (1996) allows estimation of fixed regression estimates together with the random effects estimates directly, without the need of integration to find the marginal likelihood. The PQL method is similar to this h-likelihood method, but it ignores some derivative terms in the dispersion estimation, causing non-negligible biases Lee and Nelder (2001).

5.2.1 The h-likelihood procedure

The h-likelihood for a GLMM is

$$h(\beta, \theta, v) = \log f_\beta(y|v) + \log f_\theta(v). \tag{17}$$

where $f_\beta(y|v)$ is the density of $y|v$ with $E(y|v) = X\beta + Zv$, $f_\theta(v)$ is the density of the composite random effect. This is strictly speaking not a likelihood in the Fisherian sense because of the presence of unobserved random effects. By eliminating some parameters from the h-likelihood, two forms of profile likelihood will become available: (1) the classical marginal likelihood and (2) the adjusted profile likelihood. The marginal likelihood, m, can be obtained from the h-likelihood by integrating out the random parameters,

$$m(\beta, \theta) = \log \int \exp(h(\beta, \theta, v))dv. \tag{18}$$

In normal mixed models, we can further eliminate β via conditioning to obtain the restricted likelihood (Patterson and Thomsson, 1971)

$$r(\theta) = \log f_\theta(y|\hat{\beta}_\theta).$$

where $\hat{\beta}_\theta$ is the maximum likelihood estimate (MLE) of β, given θ. The restricted likelihood was proposed for inferences about θ to reduce bias, especially in finite samples. To eliminate a generic nuisance parameter α from a log-likelihood ℓ, Lee and Nelder (2001) considered an adjusted profile likelihood function $p_\alpha(\ell)$, defined by

$$p_\alpha(\ell) = \left[\ell - \frac{1}{2}\log\left(\det\left\{D(\ell,\alpha)/2\pi\right\}\right)\right]|\alpha = \hat{\alpha}. \tag{19}$$

where $D(\ell,\alpha) = -\partial^2\ell/\partial\alpha^2$, and $\hat{\alpha}$ solves $\partial\ell/\partial\alpha = 0$, given a fixed value of the parameter of interest. For fixed effects β, the use of $p_\beta(m)$ is equivalent to conditional on $\hat{\beta}_\theta$, i.e., $p_\beta(m) \simeq r(\theta)$ to the first order (Cox and Reid, 1987), while for random effects v the use of $p_v(h)$ is equivalent to integrating them out by using first-order Laplace approximation, i.e., $p_b(h) \simeq m$. In normal linear mixed model (LMM) we can show the exact relationships

$$m(\beta,\theta) = p_v(h) \quad \text{and} \quad r(\theta) = p_\beta(m) = p_{\beta,v}(h).$$

In principle, we should use the h-likelihood h for inference about v, the marginal likelihood m for β, and the restricted likelihood $p_\beta(m)$ for dispersion parameters. Thus for estimation of family data, $p_v(h)$ is used to estimate both β and θ.

Computational algorithm The computational algorithm iterates between these two steps

- Given $\hat{\beta}$ and $\hat{\theta}$, estimate the random effects v by solving $\partial h/\partial v = 0$.
- Given $\hat{v}$, updates β and θ by solving $\partial p_v(h)/\partial\beta = 0$ and $\partial p_v(h)/\partial\theta = 0$.

5.2.2 H-likelihood for age-at-onset traits

Given the observed data $(y_{ij}, \delta_{ij}, b_{ij}, x_{ij})$ from all individuals, and the independent assumptions (section 4.2.2), H-likelihood can be easily adapted to LTRC data. The h-likelihood (Lee and Nelder, 1996) for model (8) is

$$h \equiv h(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{v}) = \sum_{ij}\ell_{1ij} + \sum_i \ell_{2i}, \tag{20}$$

where the first term is the sum of conditional log-density of (y_{ij}, δ_{ij}), given the composite random effects $\boldsymbol{v}$ and $y_{ij} > b_{ij}$. Specifically, the individual contribution

is

$$\ell_{1ij} = -\frac{1}{2}\delta\left\{\log\left(2\pi\sigma_e^2\right) + (m_{ij})^2\right\} + (1-\delta_{ij})\log\left\{1-\Phi(m_{ij})\right\} - \log\left\{1-\Phi(m_{ij}^*)\right\},$$

where $m_{ij} \equiv (y_{ij}-\mu_{ij})/\sigma_e$, $m_{ij}^* \equiv (b_{ij}-\mu_{ij})/\sigma_e$, $\Phi(\cdot)$ is the standard normal distribution function and μ_{ij} is given by (9).

The second term in (20) can be given most succinctly in matrix notation. The likelihood contribution of the ith family is

$$\ell_{2i} \;=\; -\frac{1}{2}\{\log|2\pi D_i| + \boldsymbol{v}_i^T D_i^{-1}\boldsymbol{v}_i\}.$$

where $D_i = \sigma_g^2 R_{ig} + \sigma_f^2 R_{if} + \sigma_c^2 R_{ic}$ is the covariance matrix of the random effects for the ith family in the MAFT model (8). The estimation procedure follows the computational algorithm as described in section 5.2.1. General analytical derivatives for $p_v(h)$ is complicated, and therefore we simply use a derivative-free simplex algorithm as implemented in statistical software ℝ. The h-likelihood estimation procedure is summarized in the following steps:

A. Given $\boldsymbol{\theta}$, estimate $(\boldsymbol{\beta}, \boldsymbol{v})$ by by solving $\hat{\beta} = {}^{\partial h}/_{\partial\beta} = 0$; and

B. Given $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{v}})$ from step A, estimate $\boldsymbol{\theta}$ by maximizing the adjusted-profile likelihood function (19)

The two steps are iterated until convergence.

5.2.3 H-likelihood for age-at-onset traits based on case-control family

Ha et al. (2007) fitted the MAFT model for twins data, using the h-likelihood approach. However, the volume of family data makes the h-likelihood approach computationally too demanding for routine analysis. Again, proper data reduction approach is required to facilitate analysis for such vast amount of data. While grouping the configurations of the family outcomes and covariates $\{\boldsymbol{y}_i, \boldsymbol{x}_i\}$ accordingly can efficiently reduce the binary trait data, grouping does not have the same efficient in data reduction for age-at-onset trait. Therefore the ascertainment scheme here is different from that of the binary case. For frailty models of simple sib-only structure, (Moger et al., 2008) showed that one might use the pseudo-likelihood analysis of the case-control data to produce efficient estimates comparable to those from an exact likelihood analysis of the full data. Here, we present

the family-based case-control design and the pseudo-h-likelihood approach in the analysis of MAFTs.

Using the same SGA example and notation, let S_1 be the set of all families with at least 2 affected members, then S_0 is defined as the set of control families, where the number of affected members is less than 2. Here, all case family will be included with probability $p_1 = 1$, and control family with $p_0 < 1$. Since the size of the cohort is known, p_0 is simply a fraction of the total control families. In general, we might want to stratify the families by some characteristics (for example, the number of children), and this will generate finer partitions. Let A_k be the set of families ascertained from S_k; since $p_1 = 1$, trivially $A_1 = S_1$. The pseudo-(log)-likelihood, which is an unbiased estimator of the full log-likelihood, is given by

$$\ell_p = \sum_{k=1}^{K} \frac{1}{p_k} \sum_{i \in A_k} m_i, \tag{21}$$

where m_i is given by (18). The pseudo-likelihood has the same complexity as the marginal likelihood. However, the same Laplace approximation applies immediately to (21), which means that the h-likelihood approach will also work here. Thus, we first define the pseudo-h-likelihood

$$h_p = \sum_{k=1}^{K} \frac{1}{p_k} \sum_{i \in A_k} \left\{ \sum_j \ell_{1ij} + \sum_i \ell_{2i} \right\}, \tag{22}$$

with the same ℓ_{1ij} and ℓ_{2i} as before. Then the estimation becomes straightforward:

- Given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ estimate the random effect $\hat{\boldsymbol{v}}$, by solving $\partial h_p / \partial v = 0$,
- Given $\boldsymbol{\theta}$ and $\hat{\boldsymbol{v}}$ estimate $\hat{\boldsymbol{\beta}}$ by solving $\partial h_p / \partial \beta = 0$,
- Given $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{\beta}}$ estimate $\boldsymbol{\theta}$ by maximizing

$$p_v(h_p) = \sum_{k=1}^{K} \frac{1}{p_k} \sum_{i \in A_k} p_{v_i}(h), \tag{23}$$

  where $p_{v_i}(h)$ is the contribution to previous adjusted profile likelihood from family i. This formula is motivated by the individual Laplace approximation of the marginal likelihood m_i.

6 RESULTS AND DISCUSSION

Family data provide an important analytical basis for exploring variation in human traits. However, when a disease of interest is rare, large datasets will be needed. In Sweden, the long tradition of good population statistics, the structure of the Swedish health care system, and the nationwide healthcare registers collectively provide an excellent basis for quantitative genetic studies. In this thesis, we have developed statistical approaches to deal with two main challenges: (1) statistical inference for binary traits and age-at-onset traits, (2) efficient computation.

6.1 Marginal likelihood versus h-likelihood

Both Monte Carlo approximation and h-likelihood approaches have been studied and applied to tackle the first challenge. In Paper I, we showed that h-likelihood works well with a large number of families and the results are comparable with Monte Carlo approach, but computation of individual random-effect estimates proves time-consuming. By comparison, the run-time using the marginal likelihood approach with prevalence as the only covariates is 60 times faster (Noh et al., 2006)! On the other hand, while the dimension of random effect parameters increases, as in Paper II, the marginal likelihood may get stuck in local maxima, due to the noise from Monte Carlo integration. We found that the method of moments, by using the observed ORs and the model-based ORs, works well and can obtain reasonable satisfactory starting values, when the model is without covariates. For general models, further investigation is required to validate this approach, especially if some covariates are correlated with random effects.

Another limitation with the marginal approach is that it does not provide asymptotic standard errors. Instead, the standard errors for the marginal approach were obtained by smoothing the marginal log-likelihood around each estimated parameter value (Paper II) or by using the bootstrap method (Paper III). Consequently, one might expect these to be less reliable than standard errors obtained from the asymptotic formula. In principle, when the dimension of covariates is high the h-likelihood approach is preferred, which is highly flexible and allows straightforward handling of covariates, as shown in Papers I and IV.

6.2 Ascertainment

To deal with the second challenge, efficient computation, two ascertainment schemes used for analysis of different traits have been put forth. Large datasets are re-

quired to obtain meaningful results in quantitative genetics. But the so-called "genetically-loaded" families are only a small fraction of the population. We have applied and extended the usual case-control study design (which is commonly used in medical studies to reduce costs) and collected information on cases as well as a subsample of control families. It is well-known that a case-control study has high efficiency in terms of estimation. Traditionally, a case-family is defined as such that there is at least one affected family member. For quantitative genetic studies, we have proposed that a case-family should have at least two affected members. This is motivated by simulations studies in Paper II. Using this definition of a case-family, we proposed optimal matching as an ascertainment scheme for binary traits with grouped family configuration (Paper III), and case-control scheme using pseudo-h-likelihood for estimation when the outcome is an age-at-onset trait (Paper IV).

In both ascertainment schemes, we have taken the advantage of the availability of some population statistics. For binary traits, the optimal matching used the observed ORs and regression coefficient from the full cohort. This approach is akin to making balanced considerations in two-stage sampling methodology (Reilly, 1996), but the simulations approach to sample selection allows much more complex stratification. For age-at-onset traits, the use of family-based case-control design and pseudo-h-likelihood is applicable since the fractions for each sampled family are known when the population data are available. While existing family-based case-control methods are still not practical enough for routine use (Moger et al., 2008; Lu and Wang, 2002) for fitting the complex genetic and environmental component models, we have shown in Paper IV that the pseudo-h-likelihood approach is more straightforward to use. The pseudo-h-likelihood approach is also useful for other case-control studies where a mixed-effect model is of interest.

6.3 Twins data versus Family data

While the assumption of genetic correlation is based on underlying biological mechanisms, the assumption of environmental correlation lacks such a foundation and is harder to make and justify. As an example, unlike twins, siblings are different in age and hence may experience different environments during development. Intuitively, we may expect that siblings are less environmentally correlated with each other as the difference in age increases. One way of improving the model is to allow free parameters for the correlations, but it is not clear whether there will be enough evidence in the data to separate this parameter from the corresponding family variance component.

In twins data analysis, twins are assumed to be exchangeable, and the sum of the variance components should be identical for the twins. We found in Paper II that parents and children have different unshared-common variance components. One explanation is that the models for the analysis were simplistic and missed some important covariates which may capture and explain the individual (e.g., parents, child) variation. Besides, existing models did not consider that parents and their children belong to different birth cohorts. The model developed in Paper IV can accommodate multiple covariates and age-of-onset explicitly.

6.4 Applications

The methods developed in this thesis have successfully given some promising results in pediatric, cancer and psychiatric epidemiology studies.

6.4.1 Preeclampsia

In Paper I we first demonstrated that the estimates by using the h-likelihood approach were comparable with the analysis from Pawitan et al. (2004), using the exact marginal likelihood. The GLMM used for comparison included the pregnancies order (first or subsequent) as the only covariates. Then we extended the GLMM model and considered general covariates or risk factors, such as, diabetes status of the mother, whether or not the mother is Nordic, maternal delivery age and the smoking status of the mother. The results shown that being older, Nordic, or diabetic are significantly associated with a higher risk of preeclampsia. However we also shown that some covariates, such as diabetes and smoking, are partly confounded with the genetic effects. The heritability to PE droped from 55% to 45% and indicates familial co-aggregation in PE is partly due to background risk factors (i.e. smoking, diabetes). Our finding suggested that genetic factors may only be fractionally mediated by known risk factors, such that there must be other possibly unknown mechanism involved.

6.4.2 Comorbidiy of schizophrenia and bipolar disorder

The covariance models from Paper II have been used to study etiology of familial co-aggregation in schizophrenia and bipolar disorder (Lichtenstein et al., In Press), and in melanoma and squamous cell carcinoma of the skin (Lindström et al., 2007). A shared origin of schizophrenia and bipolar disorder is well-recognized from their overlapping risk factors and the high comorbidity risk, though attempts

to disentangle the genetic and environmental effects have proved unsuccessful in two earlier family studies (Kendler et al., 1993; Maier et al., 1993). We found that schizophrenia and bipolar disorder have an approximate genetic correlation of 0.66, which agrees with an earlier study (Cardno et al., 1999). The genetic effect explained 65.5% of the variability of comorbidity and supports the results from linkage studies, which have found common susceptibility loci for schizophrenia and bipolar disorder (Potash et al., 2003; Walss-Bass et al., 2005).

6.4.3 Small-to-gestational age

In an earlier study of SGA Svensson et al. (2006) showed that genetic factors, especially the foetal component, account for majority of the liability of having SGA births. However, the model they used only contained birth order as the only covariates, while the genetic liability to SGA may be partly mediated by well-known maternal risk factors for SGA births, such as smoking and preeclampsia. But due to computational constraints, they could not investigate potential confounding effects by including such risk factors into the model. By using the novel approach in Paper III, we could perform analysis for the model including mother smoking status, birth order, preeclampsia history, maternal body mass index (BMI), maternal nationality (Nordic or not). Although still significant, the maternal and foetal genetic variance components drop substantially in more complex models. This reflects some confounding between these components and the risk factors, which is not surprising. For example, preeclampsia is associated with both maternal and fetal genetic effects (Pawitan et al., 2004). This illustrates the potential utility of performing GLMM with general risk factors.

6.4.4 Melanoma

The MAFT models were applied to the survival data of melanoma. Simulation studies were performed and from that results, we learned : (i) with the full data, the h-likelihood proceduces accurate estimate of the parameters of the MAFT-model, and (ii) the variance component parameters can be well estimated using the pseudo-h-likelihood procedure applied to the case-control data, but potentially with some loss of precision. For variance component estimation, this loss can be reduced by increasing the number control families (the number of control families is equal to and twice the number of case families).

For the real data analysis, we obtained total of 125,739 families, with 367 case-families (least 2 affected member). In addition to analysing the full data, two case-

control datasets were obtained, by sampling 2 and 4 control families for every case family. The two case-control datasets were approximatly 0.9% and 1.5% of the families in the full data. The results from the ascertained data, particularly dataset wtih 4 control families, were close to the full data. The standard errors of the regression estimates from the ascertained data were larger compared to full data SEs, those of the variance component estimates were comparable. This is consistent with the simulation results.

7 CONCLUSION

By using GLMM as a framework for quantitative genetic studies, we have successfully developed new approaches that can be applied to data with arbitrary family structures and various traits, namely:

1. single binary traits;
2. multivariate binary traits; and
3. age-at-onset traits,

where general covariates can be included in the analysis. However, when applied to large population-based data, the routine analysis is possible and practical only when data reduction techniques are employed. Based on the case-contol design and the new definition of a case family, non-random ascertaiment were applied and shown a great efficiency in data reduction.

We believe that our main contributions of the developed approaches offer new strategies for a more comprehensive scope of quantitative genetic studies, in which the modelling is more flexible by allowing general covariates and the computation becomes more efficient.

We intend to standardize our programs and offer free use for the wider research community in the future.

8 ACKNOWLEDGEMENT

I would like to express my sincere gratitude to everyone who has contributed to the present study. My time in MEB has been a very rewarding and enjoyable time, thanks to all the people who have shared ideas, experiences and everyday life with me during these years. My special thanks to:

Yudi Pawitan, my supervisor and my friend, for your patience, encouragement, inspiration and support from the first day. Thank you for teaching me the logic of thinking (solving puzzles). Your understanding, belief and guidance, made it possible for me to complete this thesis.

Paul Lichtenstein, my co-supervisor, I truly admire how you approach and solve complicated scientific problems in simple ways. Thank you for all the support and guidance during these years.

Stefano Calza, my mentor and my master in ℝ, for all the fun we have together and for your friendship. "May the force be with you".

Kamila Czene, my unofficial mentor, for sharing your profound knowledge in epidemiology, always in a generous and straightforward way. I always leave your room with new energy and greater confidence in myself.

Linda Lindström, Camilla Björk, Sven Cnattingius, Maengseok Noh, Youngjo Lee, Tron Moger, Christina Hultman for collaborations and co-authorship.

Chuen Seng Tan, my dear friend and "room mate", for all the fruitful discussions we had and for always willing to help.

Marie Jansson, for your invaluable help with various administrative things, especially with the dissertation application.

Marie Reilly, for your constructive comments, which I both hate and love. Thank you also for the last-minute proofread.

Hans-Olov Adami, *Nancy Petterson*, *Henrik Grönberg*, the past and present prefects of MEB, for your vision, creativity and plans of MEB.

The past and present lovely members of Biostatistics group for making everyday work pleasant and creating such a wonderful working environment.

The past and present wonderful IT-crew, for creating and maintaining such a superior system.

The psychiatry group in Hong Kong, for helping me, a Swedish guy, to enjoy the Hong Kong social life.

I must also thank *my parents* and *my sister* for their constant love, prayers, support and belief in me, especially all the pratical help and support when I am in Hong Kong.

Uncle *Wing* and his family, for your care and warmth.

My new big family in Hong Kong: *family Mok* and *family Lui*, for your superb care of my child and treating me as a real family member. Especially Liam, who designed the excellent cover page of my thesis and forced to be the "bad" brother from time to time.

Last, but not least, I would like to thank my beautiful wife *Bobo Mok*, for your love, care, prayers and lots of yummi food. Without you everything is meaningless. Of course I really need to thank you for our adorable daughter, *Iman*, who makes my life more complete. I love you both!

9 REFERENCES

Aalen OO, Tretli S. Analyzing incidence of testis cancer by means of a frailty model. Cancer Causes and Control, 10:285–292, 1999.

Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. American Journal of Human Genetics 54:535–43, 1994.

Blangero J, Williams JT, Almasy L. Variance component methods for detecting complex trait loci. Advances in Genetics, 42:151–81, 2001.

Breslow N and Clayton D. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88:9-25, 1993.

Breslow N and Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika, 82:81–91, 1995.

Burton PR, Tiller KJ, Gurrin LC, Cookson WO, Musk AW, Palmer LJ. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (glmms) and gibbs sampling. Genetic Epidemiology, 17:118–40, 1999.

Burton PR. Correcting for nonrandom ascertainment in generalized linear mixed models (glmms), fitted using gibbs sampling. Genetic Epidemiolology, 24:24– 35, 2003.

Cardno AG, Marshall EJ, Coid B, Macdonald AM, Ribchester TR, Davies NJ, et al. Heritability estimates for psychotic disorders: the maudsley twin psychosis series. Archives of General Psychiatry, 56:162–8, 1999.

Chan J and Kuk A. Maximum likelihood estimation for probit linear mixed models with correlated random effects. Biometrics, 53:86–97, 1997.

Chen J, Zhang D, Davidian M. A monte carlo em algorithm for generalized linear mixed models with flexible random effects distribution. Biostatistics, 3:347–60, 2002.

Cox D and Reid N. Parameter orthogonality and approximate conditional inference. Journal of the Royal Statistical Society. Series B, 49:1–39, Jan 1987.

de Andrade MD and Amos CI. Ascertainment issues in variance component models. Genetic Epidemiology, 19:333–344, 2000.

Drum M and McCullagh P. Reml estimation with exact covariance in the logistic mixed model. Biometrics, 49:677–689, 1993.

Ekholm B, Ekholm A, Adolfsson R, Vares M, Osby U, Sedvall GC, and Jönsson EG. Evaluation of diagnostic procedures in swedish patients with schizophrenia and related psychoses. Nordic Journal of Psychiatry, 59:457–64, 2005.

Elston RC and Sobel E. Sampling considerations in the gathering and analysis of pedigree data. American Journal of Human Genetics , 31:62–9, 1979.

Epstein MP, Lin X, Boehnke M. Ascertainment-adjusted parameter estimates revisited. American Journal of Human Genetics , 70:886–95, 2002.

Falconer DS. Inheritance of liability to certain diseases estimated from incidence among relatives. Annals of Human Genetics, 29:51–76, 1965.

Fisher RA. On the correlation between relatives on the supposition of mendelian inheritance. Translations of the Royal Society, 52:399–433, 1918.

Genz A. Comparison of methods for the computation of multivariate normal probabilities. Computing Science and Statistics, 25:400-405, 1993.

Guo G and Want J. The mixed or multilevel model for behavior geneticanalysis. Behavior Genetics, 32:37–48, 2002.

Ha ID and Lee Y. Multilevel mixed linear models for survival data. Lifetime Data Analysis, 11:131–142, 2005.

Ha ID, Lee Y, Pawitan Y. Genetic mixed linear models for twin survival data. Behavior Genetics, 37:621–30, 2007.

Hobert J and Casella G. The effect of improper priors on gibbs sampling in hierarchical linear mixed models. Journal of the American Statistical Association, 91:1461–1473, 1996.

Hougaard P. Analysis of multivariate survival data. Springer, 2000.

Kalbfleisch JD and Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. Statistics in Medicine, 7:149–60, 1988.

Kendler KS, McGuire M, Gruenberg AM, O'Hare A, Spellman M, Walsh D. The roscommon family study. I. methods, diagnosis of probands, and risk of schizophrenia in relatives. Archives of General Psychiatry, 50:527–40, 1993.

Klein JP, Pelz C, Zhang MJ. Modeling random effects for censored data by a multivariate normal regression model. Biometrics, 55:497–506, 1999.

Korsgaard R and Andersen AH. The additive genetic frailty model. Scandinavian Journal of Statistics, 2:255–270, 1998.

Lai T and Ying Z. A missing information principle and m-estimators in regression analysis with censored and truncated data. The Annals of Statistics, 22:1222–1255, 1994.

Laird N and Ware J. Random-effects models for longitudinal data. Biometrics, 38:963-974, 1982.

Lambert P, Collett D, Kimber A, Johnson R. Parametric accelerated failure time models with random effects and an application to kidney transplant survival. Statistics in Medicine, 23:3177–3192, 2004.

Lee Y and Nelder JA. Hierarchical generalized linear models. Journal of the Royal Statistical Society. Series B, 58:619–656, 1996.

Lee Y and Nelder JA. Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. Biometrika, 987–1006, 2001.

Lichtenstein P, Björk C, Hultman CM, Scolnick E, Sklar P, Sullivan PF. Recurrence risks for schizophrenia in a swedish national cohort. Psychological Medicine, 36:1417–1425, 2006.

Lichtenstein P, Yip BH, Björk C, Pawitan Y, Cannon TD, Sullivan PF, Christina M. Common genetic influences for schizophrenia and bipolar disorder: A population-based study of 2 million nuclear families. Lancet, In press.

Lin X and Breslow N. Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of the American Statistical Association, 19:1007–
1016, 1996.

Lindström L, Pawitan Y, Reilly M, Hemminki K, Lichtenstein P, Czene K. Estimation of genetic and environmental factors for melanoma onset using population-based family data. Statistics in Medicine, 25:3110-23, 2006.

Lindström LS, Yip BH, Lichtenstein P, Pawitan Y, Czene K. Etiology of familial aggregation in melanoma and squamous cell carcinoma of the skin. Cancer Epidemiol Biomarkers Prev, 16:1639–43, 2007.

Lu S and Wang M. Cohort case-control design and analysis for clustered failure-time data. Biometrics, 58:764–772, 2002.

Maier W, Lichtermann D, Minges J, Hallmayer J, Heun R, Benkert O, Levinson DF. Continuity and discontinuity of affective disorders and schizophrenia. Results of a controlled family study. Archives of General Psychiatry, 50:871–83, 1993.

Mcculloch CE. Maximum-likelihood variance-components estimation for binary data. Journal of the American Statistical Association, 89:330–335, 1994.

Mcculloch CE. Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association, 92:162–170, 1997.

Moger TA, Pawitan Y, Borgan O. Case-cohort methods for survival data on families from routine registers. Statistics in Medicine, 27:1062–74, 2008.

Neale M and Cardon L. Methodology for genetic studies of twins and families. Springer 1992.

Noh M, Yip B, Lee Y, Pawitan Y. Multicomponent variance estimation for binary traits in family-based studies. Genetic Epidemiology, 30:37–47, 2006.

Patterson HD and Thomson R. Recovery of inter-block infromation when block sizes are unequal. Biometrika, 58:545–554, 1971.

Pawitan Y. In all likelihood: Statistical modelling and inference using likelihood. Oxford University Press, 2001.

Pawitan Y, Reilly M, Nilsson E, Cnattingius S, Lichtenstein. P Estimation of genetic and environmental factors for binary traits using family data. Statistics in Medicine, 23:449–65, 2004.

Pearson K and Lee A. On the inheritance of characters not capable of exact quantitative measurement. Philosophical Transactions of the Royal Society of Lodon, A., 195:79–150, 1901.

Potash B, Zandi PP, Willour VL, TH Lan, YQ Huo, D Avramopoulos, YY Shugart, et al. Suggestive linkage to chromosomal regions 13q31 and 22q12 in families with psychotic bipolar disorder. The American Journal of Psychiatry, 160:680–686, 2003.

Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal, 2:1-21, 2002.

Reilly M. Optimal sampling strategies for two-stage studies. American Journal of Epidemiology, 143:92–100, 1996.

Ros HS, Lichtenstein P, Lipworth L, Cnattingius S. Genetic effects on the liability of developing preeclampsia and gestational hypertension. American Journal of Medical Genetics, 91:256–60, 2000.

Schall R. Estimation in generalized linear models with random effects. Biometrika, 78:719–727, 1991.

Sham P. Statistics in human genetics. Oxford University Press, 2007.

Shun Z and McCullagh P. Laplace approximation of high dimensional integrals. Journal of the Royal Statistical Society. Series B, 57:749–760, 1995.

Shun Z. Another look at the salamander mating data: A modified laplace approximation approach. Journal of the American Statistical Association, 92:341–349, 1997.

Svensson AC, Pawitan Y, Cnattingius S, Reilly M, Lichtenstein P. Familial aggregation of small-for-gestational-age births: the importance of fetal genetic effects. American Journal of Obstetrics and Gynecology, 194:475–9, 2006.

Walss-Bass C, Escamilla MA, Raventos H, Montero AP, Armas R, Dassori A, et al. Evidence of genetic overlap of schizophrenia and bipolar disorder: Linkage disequilibrium analysis of chromosome 18 in the costa rican population. American journal of medical genetics. Part B, 139B:54–60, 2005.

Zeger S and Karim M. Generalized linear models with random effects; a gibbs sampling approach. Journal of the American Statistical Association, 86:79–86, 1991.