Statistical Analysis of Haplotypes, Untyped SNPs, and CNVs in Genome-Wide Association Studies

by
Yijuan Hu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Political Science.

Chapel Hill
2011

Approved by:

Dr. Danyu Lin, Advisor

Dr. Donglin Zeng, Reader

Dr. Wei Sun, Reader

Dr. Fred Wright, Reader

Dr. Patrick Sullivan, Reader

Abstract

YIJUAN HU: Statistical Analysis of Haplotypes, Untyped SNPs, and CNVs in Genome-Wide Association Studies.
(Under the direction of Dr. Danyu Lin.)

Missing data arise in genetic association studies when one is interested in assessing the effects of haplotypes, untyped single nucleotide polymorphisms (SNPs) or copy number variants (CNVs). Haplotypes are combinations of nucleotides at multiple loci along individual homologous chromosomes, and the use of haplotypes tends to yield more efficient analysis of disease association than SNPs. Untyped SNPs are SNPs that are not on the genotyping chips used in the study (i.e., missing on all study subjects), and the analysis of untyped SNPs can facilitate localization of disease-causing variants and permit meta-analysis of association studies with different genotyping platforms. A CNV refers to the duplication or deletion of a segment of DNA sequence compared to a reference genome assembly, and can play a causal role in genetic diseases.

In the first part of the proposal, we provide a general likelihood-based framework for making inference on the effects of haplotypes or untyped SNPs and their interactions with environmental variables. Unlike most of the existing methods, we allow genetic and environmental variables to be correlated. We show that the maximum likelihood estimators are consistent, asymptotically normal, and asymptotically efficient and we develop EM algorithms to implement the corresponding inference procedures. We conduct extensive simulation studies and apply the methods to a genome-wide association study (GWAS) of lung cancer.

In the second part, we focus on comparing two approaches in the analysis of untyped SNPs. The maximum likelihood approach integrates prediction of untyped genotypes and estimation of association parameters into a single framework and yields consistent and efficient estimators of genetic effects and gene-environment interactions with proper variance estimators. The imputation approach is a two-stage strategy which first imputes the untyped genotypes by either the most likely genotypes or the expected genotype counts and then uses the imputed values in downstream association analysis. We conduct extensive simulation studies to compare the bias, type I error, power, and confidence interval coverage between the two methods under various situations. In addition, we provide an illustration with genome-wide data from the Wellcome Trust Case-Control Consortium (WTCCC).

In the third part, we present a general framework for the integrated analysis of CNVs and SNPs in association studies, including the analysis of total copy number as a special case. We use allele-specific copy numbers (ASCNs) to describe both the copy number and allelic variations of a locus. Our approach combines the ASCN calling and association analysis into a single step while allowing for differential errors. We construct likelihood functions that properly account for the case-control sampling and measurement errors. We establish the asymptotic properties of the maximum likelihood estimators and develop EM algorithms to implement the proposed inference procedures. The advantages of the proposed methods over the existing ones are demonstrated through realistic simulation studies and an application to a GWAS of schizophrenia.

Acknowledgments

I owe my gratitude to all the people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my advisor, Dr. Danyu Lin. I have been amazingly fortunate to have an advisor who set an example of a world-class researcher. His immense knowledge and high research standard have always been the major driving forces throughout my graduate study. I also appreciate his generous financial support whenever needed. I hope that one day I would become as good a researcher as Dr. Lin is and as good an advisor to my students as he has been to me.

Dr. Donglin Zeng has always been there to listen and give advice. I am heartily thankful to him for the long discussions that teach me how to grasp the spirit of complicated statistical theories. His patience and invaluable suggestions are indispensable to the completion of this dissertation.

I was extremely delighted to work with Dr. Wei Sun, who has been a motivating colleague as well as a fun friend. I owe big gratitude to him for all the help he can possibly offer.

I am deeply indebted to Dr. Fred Wright, who introduced me to the field of statistical genetics and encouraged me to pursue a career there. I appreciate his numerous pieces of advice throughout these years.

I am greatly thankful to Dr. Patrick Sullivan for reading a draft of my dissertation, commenting on my views and enriching my ideas. I sincerely wish I had more opportunities to work with him.

My sincere thanks also goes to Dr. Lloyd Chambless, for offering me a research assistant job for two years and leading me working on diverse exciting projects. He made my time in Collaborative Studies Coordinating Center a unique working experience of my life.

I owe special thanks to Dr. Jianwen Cai for providing a financial support for me to come to this prestigious department in the first place. None of this would have been possible without her timely help and enormous kindness.

Many friends and schoolmates have helped me stay positive through these years. Their belief in me helped me overcome setbacks and stay focused on my graduate study. I greatly value their friendships.

Finally and most importantly, none of this would have been possible without the love and patience of my husband Xuan. His support and encouragement were undeniably the bedrock upon which the past six years of my life has been built. I thank my parents, Kaicheng and Qin, for their faith in me in all my endeavors. It was under their watchful eyes that I gained so much drive and an ability to tackle challenges head on.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ASCN	Allele-specific copy number
CNV	Copy number variant
EM	Expectation-maximization
GMM	Gaussian mixture model
GWAS	Genome-wide association study
HMM	Hidden Markov model
HWE	Hardy-Weinberg equilibrium
LD	Linkage disequilibrium
MLE	Maximum likelihood estimator
NPMLE	Nonparametric maximum likelihood estimation
SNP	Single nucleotide polymorphism
WTCCC	Wellcome Trust Case-Control Consortium

Chapter 1

Introduction

Many diseases of utmost public health significance, including cancer, hypertension, diabetes, and schizophrenia, are influenced by a variety of genetic and environmental factors, as well as gene-environment interactions. It is widely recognized that genetic dissection of such complex human diseases requires large-scale association studies, which relate disease phenotypes to genetic variants such as single nucleotide polymorphisms (SNPs) and copy number variants (CNVs). In fact, there is now a proliferation of genetic association studies worldwide thanks to the availabilities of dense SNP maps across the human genome and precipitous drops in genotyping costs. An increasing number of these studies survey the entire genome with high-density genotyping chips containing hundred thousand or more SNPs for thousands of individuals; such studies are referred to as genome-wide association studies (GWAS).

There are several options in designing population-based genetic association studies. The simplest is the cross-sectional design, which is preferable if the disease of interest is common or if one is interested in some disease-related traits, such as blood pressure. For rare diseases, it is more cost-effective to adopt the case-control design, which collects genetic and exposure information on each subject retrospectively; as a matter of fact, owing to the strong negative selection and thereby rarity of many complex diseases, most of ongoing studies are case-control. If one is interested in the age at onset of a disease, then it is desirable to follow a cohort of at-risk individuals.

Missing data present a major challenge in genetic association studies. Specifically, missing data arise when one is interested in assessing the effects of haplotypes, untyped SNPs or CNVs on disease phenotypes. The new features of such genetic variants call for the development of novel statistical methods. In addition, the unprecedented scale of GWAS entails new challenges. First, as the enormous number of variants leads to serious multiple comparison problems, it is crucial to develop tests that are optimally powered for the true associations to reach the stringent threshold of statistical significance. Second, the computational burden of GWAS requires methods that can be implemented in computationally efficient algorithms.

The dissertation is organized as follows. In the rest of this chapter, we review the existing literature and identify unresolved problems. In Chapter 2, we provide a general framework for studying the effects of haplotypes or untyped SNPs and/or their interactions with environmental factors. In particular, we relax the assumption of gene-environment independence. In Chapter 3, we focus on comparing two approaches to studying the effects of untyped SNPs, maximum likelihood and single imputation. In Chapter 4, we present a likelihood-based framework for integrated analysis of CNVs and SNPs in association studies, including the analysis of total copy numbers as a special case. In Chapter 5, we outline some ongoing and future work.

1.1 Inference on Haplotype Effects

A haplotype is a specific sequence of nucleotides on the same chromosome of a subject. Because haplotypes incorporate the linkage disequilibrium (LD) information (i.e., correlation structure) of multiple SNPs and correspond to protein sequences, the use of haplotypes tends to yield more efficient analysis of disease association than the use of individual SNPs, especially when the causal variants are not directly measured or when there are strong interactions among multiple mutations on the same chromosome (Akey et al., 2001; Fallin et al., 2001; Li, 2001; Morris and Kaplan, 2002; Schaid et al., 2002; Zaykin et al., 2002; Schaid, 2004). Unfortunately, current genotyping technologies do not separate a subject's two homologous chromosomes, so that we can only observe the combination of the two haplotypes, which is referred to as the (unphased) genotype.

Many papers have focused on inferring haplotypes or estimating haplotype frequencies from unphased genotype data alone, regardless of the phenotype. Excoffier and Slatkin (1995) proposed maximum-likelihood estimation of haplotype frequencies via the expectation-maximization (EM) algorithm. Their model for haplotype frequencies makes no assumption about the LD structure among multiple loci. Their method can only be applied to a small number of markers at a time, because the haplotype frequencies become too low to be estimated with any accuracy when more than a handful of markers are considered. Stephens et al. (2001) developed a Bayesian approach to inferring haplotypes via a Markov chain-Monte Carlo (MCMC) algorithm. They made explicit assumptions about the LD patterns, exploiting ideas from population genetics and coalescent theory. Their method can cope with a large number of linked SNPs simultaneously, but the running time will increase greatly as the number of SNPs increases. Given the probabilistically inferred haplotypes by Excoffier and Slatkin (1995) or Stephens et al. (2001), one can then relate them to the phenotype through a regression model in the downstream association analysis (e.g., Zaykin et al., 2002; Kraft et al., 2005; Cordell, 2006). This two-stage strategy is a form of imputation, and has several potential problems. Lin and Huang (2007) discussed in the context of case-control studies that the haplotype phasing algorithms do not acknowledge the selective-sampling feature of the case-control design and do not take into account the phenotype. Kraft et al. (2005) also noted that the variance estimators do not account for the uncertainty of haplotype phasing. As a result, this imputation strategy can yield substantial bias of estimated genetic effects, poor coverage of confidence intervals, significant inflation of type I error and diminished power of risk haplotype detection (Kraft et al., 2005; Cordell, 2006; Lin and Huang, 2007).

A large number of papers have been published in genetic journals on how to make proper inference about the effects of haplotypes on disease phenotypes. Virtually all of these methods pertain to likelihood and most of them deal with case-control studies, so we first make a distinction between the prospective and retrospective likelihoods. For case-control studies, in which the sampling is conditional on the case-control status, it is appropriate to use the retrospective likelihood. Although Prentice and Pyke (1979) established the equivalence of the retrospective and prospective likelihoods in making inference on the odds ratios, the equivalence requires the distribution of the covariates to be unrestricted and does not hold when the covariate of interest is the haplotype pair (diplotype), the distribution of which has to be restricted for the sake of identifiability. In light of this, the method of Zhao et al. (2003), which uses an estimating function approximating the expectation of the complete-data prospective-likelihood score function given the observable data, is not statistically efficient, compared to the method of Epstein and Satten (2003), which is based on a proper retrospective likelihood. Indeed, Satten and Epstein (2004) compared the methods of Zhao et al. (2003) and Epstein and Satten (2003) via simulation studies and concluded that the retrospective-likelihood method has increased efficiency with respect to the prospective method. However, Epstein and Satten (2003) did not allow environmental factors as Zhao et al. (2003) did. Stram et al. (2003) described an approach based on the joint likelihood of disease and genotype data, after accounting for the ascertainment scheme of the case-control design. This approach requires the sampling probabilities of cases and controls to be known and does not allow environmental factors either. Spinka et al. (2005) accommodated environmental factors in the proposed retrospective maximum-likelihood method and showed that the method is equivalent to an extension of the method by Stram et al. (2003), which can incorporate environmental factors. Meanwhile, Schaid et al. (2002) and Lake et al. (2003) discussed likelihood-based inference for cross-sectional studies under generalized linear models. Lin (2004) showed how to perform the Cox (1972) regression when potentially censored age-at-onset of the disease observations are collected in cohort studies. In a general framework, Lin and Zeng (2006) provided appropriate likelihoods for all commonly used study designs (i.e., cross-sectional, case-control and cohort) and a variety of disease phenotypes (i.e., quantitative traits, disease indicators and potentially censored age-at-onset). The effects of haplotypes on the phenotype are formulated through flexible regression models, which can accommodate various genetic mechanisms and gene-environment interactions. Later, Zeng et al. (2006) extended the framework of Lin and Zeng (2006) to case-cohort and nested case-control designs.

To be specific, we outline the method of Lin and Zeng (2006) for case-control studies. Let H and G denote the pair of haplotypes and the genotype for an individual based on M biallelic SNPs. We write $H = (h_k, h_l)$ if the individual's haplotypes are h_k and h_l, representing the kth and lth of total K possible haplotypes in the sample. Let Y be the disease status, and let $\mathbf{X}$ be the environmental factors. The conditional density of Y given $(H = (h_k, h_l), \mathbf{X})$, denoted by $P_{\alpha,\boldsymbol{\beta}}(Y|H = (h_k, h_l), \mathbf{X})$, can be formulated by the logistic regression with linear predictor, $\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(H = (h_k, h_l), \mathbf{X})$, or more specifically,

$$\alpha + \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X} + \boldsymbol{\beta}_3^{\mathrm{T}}\{I(h_k = h^*) + I(h_l = h^*)\}\mathbf{X},$$

where h^* is the target haplotype of interest. Note that an additive genetic effect and a gene-environment interaction are assumed in the example above, although any genetic mechanisms can be similarly formulated. Write $\pi_k = P(h = h_k)$. Lin and Zeng (2006) demonstrated that it is generally impossible to make inference about haplotype effects without imposing any structure on $P\big(H=(h_k,h_l)\big)$. Thus they considered the assumption of Hardy-Weinberg equilibrium (HWE), in which case

$$P\big(H=(h_k,h_l)\big)=\pi_k\pi_l,$$

and two specific forms of departure from HWE,

$$P\big(H=(h_k,h_l)\big)=(1-\rho)\pi_k\pi_l+\delta_{kl}\rho\pi_k,$$

and

$$P\big(H=(h_k,h_l)\big)=\frac{(1-\rho+\delta_{kl}\rho)\pi_k\pi_l}{1-\rho+\rho\sum_{j=1}^K\pi_j^2},$$

where $\delta_{kk}=1$ and $\delta_{kl}=0$ $(k\neq l)$. Denote $\boldsymbol{\pi}=(\pi_1,\dots,\pi_K)$ under HWE or $\boldsymbol{\pi}=(\rho,\pi_1,\dots,\pi_K)$ under the two forms of Hardy-Weinbery Disequilibrium (HWD). Lin and Zeng (2006) allowed the distribution function of $\mathbf{X}$ to be fully nonparametric, denoted as $F(\mathbf{x})$; let $f(\mathbf{x})$ be the corresponding density. When the disease is rare, considerable simplicity arises because of the approximation $P_{\alpha,\boldsymbol{\beta}}(Y|H,\mathbf{X})\approx\exp\big\{Y\big(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(H,\mathbf{X})\big)\big\}$. Lin and Zeng (2006) additionally assumed that $\mathbf{X}$ is independent of H, so the retrospective likelihood based on $(G_i,\mathbf{X}_i,Y_i)$, $i=1,\dots,n$, can be approximated by

$$L(\boldsymbol{\theta},F)=\prod_{i=1}^n\frac{\sum_{(h_k,h_l)\in\mathcal{S}(G_i)}e^{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(H,\mathbf{X}_i)}P_{\boldsymbol{\pi}}(h_k,h_l)f(\mathbf{X}_i)}{\int_x\sum_{(h_k,h_l)}e^{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(H,\mathbf{x})}P_{\boldsymbol{\pi}}(h_k,h_l)dF(\mathbf{x})},$$

where n is the number of study subjects, $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\pi})$, and $\mathcal{S}(G)$ denotes the set of diplotypes that are consistent with genotype G. Note that missing genotype values can be incorporated by expanding the set $\mathcal{S}(G)$ accordingly. Lin and Zeng (2006) adopted the nonparametric maximum likelihood estimation (NPMLE) approach, in which $F(.)$ is treated as a right-continuous function with jumps at the observed $\mathbf{X}$. The objective function to be maximized is obtained from $L(\boldsymbol{\theta},F)$ by replacing $f(\mathbf{x})$ with the jump size of $F(.)$ at $\mathbf{x}$. In this case, the profile likelihood, derived by maximizing $L(\boldsymbol{\theta}, F)$ with respect to the jump sizes of $F(.)$ for fixed values of $\boldsymbol{\theta}$, has a closed form

$$L^*(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{\sum_{(h_k,h_l)\in\mathcal{S}(G_i)} e^{Y_i\left\{\mu+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(H,\mathbf{X}_i)\right\}} P_{\boldsymbol{\pi}}(h_k,h_l)}{\int_{y=0,1}\sum_{(h_k,h_l)} e^{Y_i\left\{\mu+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(H,\mathbf{x})\right\}} P_{\boldsymbol{\pi}}(h_k,h_l)},$$

where μ is an unknown constant and should be treated as a free parameter. Lin and Zeng (2006) carried out the maximization by the Newton-Raphson algorithm, and established the identifiability of the parameters and the consistency, asymptotic normality, and efficiency of the maximum likelihood estimators (MLE).

All the aforementioned work assumes HWE (or certain 1-parameter extensions) and independence of genetic and environmental factors (or absence of environmental factors). The assumption of gene-environment independence fails in many applications. For example, certain genes may influence both environmental exposure and disease occurrence. Violation of the independence assumption can cause serious bias in the analysis (e.g., Spinka et al., 2005).

Recently, Chen et al. (2008) relaxed the assumption of gene-environment independence by postulating a polytomous logistic regression model for the distribution of the haplotypes conditional on the environmental factors. Specifically, they assumed

$$\log\left\{\frac{P\big(H=(h_k,h_l)|\mathbf{X}\big)}{P\big(H=(h_K,h_K)|\mathbf{X}\big)}\right\} = \zeta_{0,k,l} + \zeta_{1,k,l}\mathbf{X},$$

where $H=(h_K,h_K)$ is chosen as the reference diplotype. For the purpose of identifiability, they imposed further constraints on $\boldsymbol{\zeta}_0$ and $\boldsymbol{\zeta}_1$, which are vectorized forms for the parameters $\zeta_{0,k,l}$ and $\zeta_{1,k,l}$. Because the odds ratio associated with the distributions $P(\mathbf{X}|H)$ and $P(H|\mathbf{X})$ are the same, $\boldsymbol{\zeta}_1$ can be interpreted as measures of diplotype effects on the distribution of $\mathbf{X}$. Thus it is natural to specify $\boldsymbol{\zeta}_1$ according to certain model of effects of the underlying haplotypes. For example, assuming an additive effect for the haplotypes, one can write $\zeta_{1,k,l} = \zeta_{1,k} + \zeta_{1,l}$, which allows the diplotype effects to be determined by a reduced set of haplotype effect parameters $\zeta_{1,k}$. Note that $\boldsymbol{\zeta}_0$ defines the diplotype frequencies for a baseline value of $\mathbf{X}$. It is common to use population genetics models, such as HWE, to specify a relationship between diplotype and haplotype frequencies. However, if the diplotypes can influence certain environmental factors, the frequencies of the diplotypes within $\mathbf{X}$ categories may not follow HWE although the underlying population, as a whole, may be in HWE. Thus, the parameter $\boldsymbol{\zeta}_0$ can be defined to entail that the marginal diplotype frequencies follow HWE. Denoting the marginal haplotype frequencies by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, HWE means that

$$P\big(H = (h_k, h_l)\big) = \pi_k \pi_l.$$

Thus, $\boldsymbol{\zeta}_0$ is defined as an implicit function of $\boldsymbol{\zeta}_1$, $\boldsymbol{\pi}$ and $F(\mathbf{X})$, denoted as $\Psi(\boldsymbol{\zeta}_1, \boldsymbol{\pi}, F)$, through the relationship

$$\pi_k \pi_l = \int_{\mathbf{X}} P(H|\mathbf{X}, \boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1) dF(\mathbf{X}),$$

where $F(.)$ is treated nonparametrically. To make inference, Chen et al. (2008) first derived the profile log-likelihood by profiling $F(.)$ out of the complete-data likelihood which assumes that the underlying haplotype information is known, and then replaced $\boldsymbol{\zeta}_0$ by $\Psi(\boldsymbol{\zeta}_1, \boldsymbol{\pi}, \widetilde{F})$, where $\widetilde{F}$ is the empirical distribution of $\mathbf{X}$. After the substitution, they obtained the complete-data estimating equation for $(\boldsymbol{\beta}, \boldsymbol{\zeta}_1, \boldsymbol{\pi})$. Then they incorporated the uncertainty of the phase information by constructing a weighted version of the complete-data estimating equation, which is solved by an EM-like algorithm. Using their method, Chen et al. (2008) were able to detect an interaction between smoking and a NAT2 haplotype in the development of colorectal adenoma that was undetected under the assumption of gene-environment independence.

Chen et al. (2008) decomposed the joint density function $P(\mathbf{X}, H)$ as $P(H|\mathbf{X})P(\mathbf{X})$. Given that genetic susceptibility may influence environmental exposures and not vice versa, for causal interpretation of parameters it is more natural to consider a model for $\mathbf{X}$ given H. Chen et al. (2008) assumed HWE in the general population. Because $P(H|\mathbf{X})$ generally does not follow HWE when $P(H)$ is in HWE, Chen et al. (2008) defined the intercepts in their polytomous logistic model for $P(H|\mathbf{X})$ as implicit functions of all other parameters so as to impose HWE on $P(H)$. Those constraints complicate the estimation process. In addition, the estimating equations of Chen et al. (2008) are not likelihood score equations, the convergence properties of their EM-like algorithms are unclear and their estimators are not asymptotically efficient. As last, their work is confined to case-control studies.

1.2 Inference on Untyped SNP Effects

Untyped SNPs are SNPs that are not on the genotyping chip used in the study and are thus missing on all study subjects. Because current genotyping platforms assay only a small fraction of SNPs in the human genome, many disease-susceptibility loci will inevitably be untyped. Conducting association analysis at untyped SNPs is highly desirable because it can facilitate the selection of SNPs to be genotyped in follow-up studies and enable investigators to compare or combine results from multiple studies with different genotyping chips. Indeed, this analysis has been successful in finding associations that would not have been found using only the original genotypes. For example, Zeggini et al. (2008) imputed 2.20 million HapMap SNPs (Altshuler, 2005) in three studies of type 2 diabetes. Two of the studies had been genotyped on the Affymetrix 500K GeneChip, while the third had been genotyped on the Illumina 317K chip. The imputation of untyped SNPs resulted in two significant results that would not have been found using only the original genotypes. One of these was a known association with PPARG, while the second was a novel association with CDC123-CAMK1D, which has been confirmed through genotyping in replication samples. These associations are not among the top hits in any one study, but show a trend in each component study.

Because untyped SNPs are not measured on any study subject, the missing information cannot be recovered from the study data alone. Fortunately, the LD structure observed in an external reference panel can be used to predict untyped SNPs from typed ones. The most common reference is the HapMap, because of the dense level of genotyping, including over 3.1 million SNPs, on these samples. Once a reference panel is chosen, "untyped SNPs" are often redefined to be SNPs that are not genotyped in the study sample, but are characterized in the reference panel.

The analysis of untyped SNPs is closely related to the concept of "tagging". Specifically, Carlson et al. (2004) selected a single tag SNP as a proxy for every untyped SNP such that the correlation coefficient r^2 between the two SNPs in the reference set is higher than a certain threshold. de Bakker et al. (2005) proposed a multimarker method, acknowledging the fact that some groups of SNPs as a whole work better to predict the untyped SNP than does any single SNP. They selected a specific haplotype to serve as a proxy by r^2 criteria and compared the frequency of that haplotype between the cases and the controls. Their method results in a 1 d.f. χ^2 test. Although multimarker methods are a considerable advancement, de Bakker's method does not fully take advantage of the correlation structure between SNPs and their multimarker tags by ignoring the additional information given by the other haplotypes other than the proxy haplotype. Zaitlen et al. (2007) proposed a new criteria r_h^2 for tag SNP selection that measures the LD between a weighted combination of all haplotypes and the untyped SNP. The weights are chosen to maximize r_h^2. Zaitlen et al. (2007) also proposed a new test statistic that computes a weighted sum of all haplotype frequency differences between the cases and controls. Their method is similar to that of Stram (2004), but they did not restrict the tag SNP selection within regions of haplotype blocks as Stram (2004) did. Their method is also similar to that of Nicolae (2006), but they formulated a much broader set of tests than Nicolae (2006) by choosing different weights, which encompasses the previous single-marker and multimarker approaches involving one haplotype. Nevertheless, the three criteria, Zaitlen's r_h^2, Stram's R_s^2 and Nicolae's M_D, are equivalent. All the aforementioned methods, although simple and intuitive, are not statistically efficient and are confined to case-control comparisons without environmental factors. In a similar spirit of tagging, Lin et al. (2008) proposed a likelihood-based method for the analysis of untyped SNPs in case-control studies with or without environmental factors. The likelihood integrates the study and reference data while reflecting the biased nature of the case-control sampling. This method yields consistent and efficient estimators of genetic effects and gene-environment interactions, and the variance estimators fully account for the uncertainty in inferring the unknown variants.

In what follows, we outline the method of Lin et al. (2008). First, the LD information from a reference panel is used to select a set of $(M-1)$ typed SNPs that provide the most accurate prediction of the untyped SNP, where M is a small number, which is set to five here. The accuracy of prediction is measured by R_s^2 of Stram (2004). Given H and G defined on the set of M SNPs, with one of the component in G always missing at the untyped locus, Lin et al. (2008) extended the framework of Lin and Zeng (2006) from the analysis of haplotypes to untyped SNPs. Specifically, the conditional density $P_{\alpha,\boldsymbol{\beta}}(Y|H,\mathbf{X})$ is formulated by the logistic regression with linear predictor $\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(H, \mathbf{X})$, and $\mathcal{Z}(H, \mathbf{X})$ now models the untyped SNP effects or SNP-environment interactions. For example, under additive mode of inheritance with gene-environment interaction,

$$\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(H, \mathbf{X}) = \beta_1 \mathcal{G}_u(H) + \boldsymbol{\beta}_2^{\mathrm{T}} \mathbf{X} + \boldsymbol{\beta}_3^{\mathrm{T}} \mathcal{G}_u(H) \mathbf{X},$$

where $\mathcal{G}_u(H)$ denotes the genotype induced by the diplotype H at the untyped locus. Under the assumptions of rare disease, HWE and gene-environment independence, the likelihood for $(\boldsymbol{\beta}, \boldsymbol{\pi}, F)$ for the n study subjects takes the form

$$\prod_{i=1}^{n} \frac{\sum_{(h_k, h_l) \in \mathcal{S}(G_i)} e^{Y_i \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(H, \mathbf{X}_i)} \pi_k \pi_l f(\mathbf{X}_i)}{\int_x \sum_{(h_k, h_l)} e^{Y_i \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(H, \mathbf{x})} \pi_k \pi_l dF(\mathbf{x})}.$$

Lin et al. (2008) used the profile-likelihood arguments of Lin and Zeng (2006) to eliminate the distribution of X, so the MLE of $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ can be equivalently obtained by maximizing the profile likelihood

$$L_S(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{\sum_{(h_k, h_l) \in \mathcal{S}(G_i)} e^{Y_i \left\{\mu + \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(H, \mathbf{X}_i)\right\}} \pi_k \pi_l}{\sum_{y=0,1} \sum_{(h_k, h_l)} e^{y \left\{\mu + \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(H, \mathbf{X}_i)\right\}} \pi_k \pi_l},$$

where $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}, \boldsymbol{\pi})$ and μ is an unknown constant. If there are no environmental factors, the likelihood is simply

$$L_S(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{\sum_{(h_k, h_l) \in \mathcal{S}(G_i)} e^{Y_i \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(H)} \pi_k \pi_l}{\sum_{(h_k, h_l)} e^{Y_i \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(H)} \pi_k \pi_l}.$$

Unlike Lin and Zeng (2006), the likelihood for study subjects alone does not contain any information about $\boldsymbol{\beta}$ because $\boldsymbol{\beta}$ will be factored out of the likelihood when the values of the variant of interest are completely missing. Fortunately, the likelihood of the reference panel, denoted as $L_R(\boldsymbol{\pi})$, can be used. It is natural to assume that the study and reference panel are generated from the same underlying population, so the haplotype frequencies in the reference panel can be denoted by the same parameter $\boldsymbol{\pi}$. Lin et al. (2008) maximized the combined likelihood $L_C(\boldsymbol{\theta}) = L_S(\boldsymbol{\theta})L_R(\boldsymbol{\pi})$ through the EM algorithm. The resulting MLE is statistically efficient in that it has the smallest variance among all valid estimators and the corresponding test of association is the most powerful among all valid tests based on the same data and same assumptions. Indeed, Lin et al. (2008) showed through simulation studies that their method is uniformly more powerful than that of Nicolae (2006).

It is worth noting that the tagging-based methods, such as the method of Lin, Hu and Huang (2008), can only be applied to a small number of tags at a time, because the haplotype frequencies become too low to be estimated with good accuracy when more than a handful of tags are considered. Also note that there is no mediating model for the LD structure of the haplotypes, but the haplotype frequencies are estimated directly (nonparametrically).

Recently, gaining popularity are a group of imputation methods which take advantage of Hidden Markov Models (HMMs) (Scheet and Stephens, 2006; Marchini et al., 2007; Li et al., 2010; Browning and Browning, 2007, 2009). As opposed to the aforementioned tagging-based methods, these HMM-based methods exploit population-genetic theory for the LD structure and use information from all markers in LD with the untyped SNP. They are based on variants of the “product of approximate conditionals” (PAC) models described in Li and Stephens (2003). In these models, a subset of haplotypes comprising all SNPs on one chromosome is selected as a reference set, and each reference haplotype represents a hidden state of the HMM at each marker. The true haplotypes underlying the observed genotype data are assumed to be imperfect mosaics of the reference haplotypes. Points of change from one reference haplotype to another allow for historical recombination. The observed alleles may differ from the alleles on the underlying true haplotypes to allow for historical mutations and genotype errors. As part of the model fitting process, parameters such as historical recombination rates between adjacent markers, and mutation rates may be estimated. Once the haplotypes are inferred, the untyped genotypes can be imputed. The imputed values are then be used as known quantities in downstream association analysis.

Like the imputation methods for haplotype analysis, the imputation approach for untyped SNPs is, statistically speaking, less satisfactory than maximum likelihood methods such as that of Lin et al. (2008) because of its bias and inefficiency (Little, 1992). Imputing missing data for cases and controls together can lead to a bias toward the null hypothesis of no association and therefore a loss of power, whereas imputing missing genotypes for cases and controls separately can inflate type I error rates (Balding, 2006; Lin and Huang, 2007). The HMM-based imputation methods tend to extract more LD information from the typed SNPs than maximum likelihood, and there is likely a strong relationship between the amount of information for the untyped SNP and the performance of the association test. However, it is not guaranteed that more information leads to more powerful tests as the imputation approach often uses the information inefficiently. Nevertheless, imputation has several practical advantages over maximum likelihood. First, once the missing data are imputed, the association analysis can be readily carried out for any traits and study designs in standard software packages. Second, for each additional dataset included, it is not necessary to conduct imputation again for existing datasets. Third, analyses regarding secondary and tertiary phenotypes do not require specific imputation. Given the operational convenience of imputation and the statistical optimality of maximum likelihood, comprehensive comparisons of these two approaches are sorely needed.

1.3 Inference on Joint Effects of CNVs and SNPs

A single nucleotide polymorphism (SNP) is a DNA sequence variation that occurs when a single nucleotide in the DNA sequence is altered. SNPs account for a majority of human genetic variation and have been shown to have a significant impact on disease susceptibility. A copy number variant (CNV) refers to the amplification or deletion of a segment of DNA sequence compared to a reference genome assembly. Recent studies have documented the extensive presence of CNVs in the human genome (Sebat et al., 2004; Iafrate et al., 2004; Tuzun et al., 2005; Redon et al., 2006; Kidd et al., 2008; McCarroll et al., 2008). Changes in copy number can have dramatic phenotypic consequences by altering gene dosage, disrupting coding sequences, or perturbing long-range gene regulation. Indeed, CNVs, in particular common copy-number polymorphisms, have been reported to be associated with several complex disease phenotypes, including HIV acquisition and progression (Gonzalez et al., 2005), lupus glomerulonephritis (Aitman et al., 2006), and three systemic autoimmune diseases: systemic lupus erythematosus, microscopic polyangiitis and Wegener's granulumatosis (Yang et al., 2007; Fanciulli et al., 2007).

Because CNVs and SNPs coexist throughout the human genome and may both contribute to phenotype variation, it is desirable to consider both types of variations in association studies of complex human diseases, characterized by allele-specific copy numbers (ASCNs). Ignoring CNVs during SNP genotype calling can lead to erroneous genotypes that appear to violate Mendelian inheritance (MI) or Hardy-Weinberg equilibrium (HWE). For this reason, SNPs in the CNV regions are typically filtered out. In addition, CNVs and SNPs may act in concert to influence disease phenotypes. For example, several cancer studies have shown evidence of the joint effects of CNVs and SNPs (e.g., Van Loo et al., 2010).

SNP genotyping arrays, such as those from Affymetrix and Illumina, hold the promise to study CNVs and SNPs simultaneously. SNP arrays capture ASCN information by generating quantitative two-dimensional measurements. Specifically, Affymetrix arrays provide a pair of raw allele-specific intensities for each SNP while Illumina arrays transform the pair of raw intensities to a measurement of total copy number and a measurement of allelic contrast.

Since the underlying ASCNs are not directly observed, an intuitive approach is to call ASCNs first and then use the ASCN calls in downstream association analysis. Various calling algorithms have been proposed to dissect copy number states from SNP genotyping arrays. For example, several methods, such as QuantiSNP (Colella et al., 2007), PennCNV (Wang et al., 2007) and GenoCNV (Sun et al., 2009), rely on Hidden Markov Models (HMMs) to segment the intensity measurements along the genome. They were designed for Illumina array data for a single sample. PennCNV assumes that the parameters of the HMM are known. QuantiSNP imposes some common priors for these parameters so that only a few hyper-parameters need to be estimated. GenoCNV allows these parameters to be estimated from the data. GenoCNV directly estimate ASCNs. PennCNV and QuantiSNP only output calls for total copy numbers, though ASCNs can be obtained by applying appropriate thresholds for the allelic contrast measurements. For the Affymetrix 6.0 array, a commonly used software is Birdsuite (Korn et al., 2008). While assuming prior information are available for common CNVs but not for rare ones, rare CNVs and common CNVs are handled differently in Birdsuite. Rare CNVs are discovered by an HMM. For common CNVs, Birdsuite makes use of their prior knowledge, such as the locations and copy number states, and reduces the identification of CNVs to CNV "genotyping", which is analogous to SNP genotyping. Specificly, at each known common CNV region, a univariate Gaussian mixture model (GMM) is used to cluster total copy number measurements across individuals and assign each individual a total copy number state. At last, the ASCNs are derived at SNP sites by two-dimensional GMMs informed by the total copy number assignments.

After ASCN calling, the downstream association analysis can be carried out in standard software packages. This two-step strategy is a form of "imputation" in missing data literature. This imputation approach is not optimal for two reasons. First, the association testing may not be robust to the differential errors between cases and controls caused by differences in DNA quality or handling; see Figure 1.1. For example, differential errors arise when batch effects in array processing are correlated with the disease status. Differential errors are prevalent and difficult to exclude, as case and control samples can rarely be obtained in strictly comparable circumstances to ensure identical DNA handling. In the presence of such errors, calling ASCNs with cases and controls combined will lead to differential misclassification and will generate excessive false-positive findings of association. Second, imputation *per se* carries serious flaws because it ignore the phenotype which may be informative about the missing data and the association analysis does not account for the uncertainty in inferring missing data. In general, imputation may yield biased parameter estimators and incorrect variance estimators, which may result in inflated type I error (Hu and Lin, 2010).

Barnes et al. (2008) described a likelihood-based method for association studies with total CNVs, which accounts for differential errors and avoids imputation. We illustrate the method of Barnes et al. (2008) in the following. Let R denote the quantitative copy number measurement, K the unobserved true copy number, which is an integer, Y the phenotype and $\mathbf{X}$ the environmental factors. Barnes et al's method is based on the following factorization:

$$P_{\boldsymbol{\gamma},\boldsymbol{\delta},\alpha,\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\pi}}(R, Y, K|\mathbf{X}) = P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R|Y, K, \mathbf{X}) P_{\alpha,\boldsymbol{\beta},\boldsymbol{\xi}}(Y|K, \mathbf{X}) P_{\boldsymbol{\pi}}(K|\mathbf{X}),$$

where the three component parts are referred to as the "signal model", "phenotype model" and "copy number model", respectively. For the signal model, R is assumed to be normally distributed with mean and variance depending on K as well as on $(Y, \mathbf{X})$ to allow for differences in DNA sources and batch effects. The "signal mean" can be modeled through the linear regression with parameters $\boldsymbol{\gamma}$, and the "signal variance" can be linked to the $\boldsymbol{\delta}$-indexed linear predictor by a logarithmic link function. The phenotype model can be any generalized linear model (GLM), and in particular logistic regression for case-control studies. The copy number model $P_{\boldsymbol{\pi}}(K|\mathbf{X})$ can be simplified by assuming gene-environment independence. Then the distribution of K is assumed to be multinomial with $\boldsymbol{\pi}$ denoting the frequencies. Barnes et al. (2008) maximized the likelihood of n study subjects

$$\prod_{i=1}^{n} P_{\boldsymbol{\gamma},\boldsymbol{\delta},\alpha,\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\pi}}(R_i, Y_i, K_i|\mathbf{X}_i)$$

by a variation of the EM algorithm, termed the ECM algorithm, and tested the null hypothesis $\boldsymbol{\beta} = 0$ by the likelihood ratio test.

The method of Barnes et al. (2008) has important limitations. First, it is confined to the total copy number and ignores possible allelic effects. It collapses the allele-specific copy number measurements at SNP sites into total copy number measurements, which may lose information and reduce power. In addition, it adopts a prospective likelihood, which may not be appropriate for case-control studies with missing data or measurement errors.

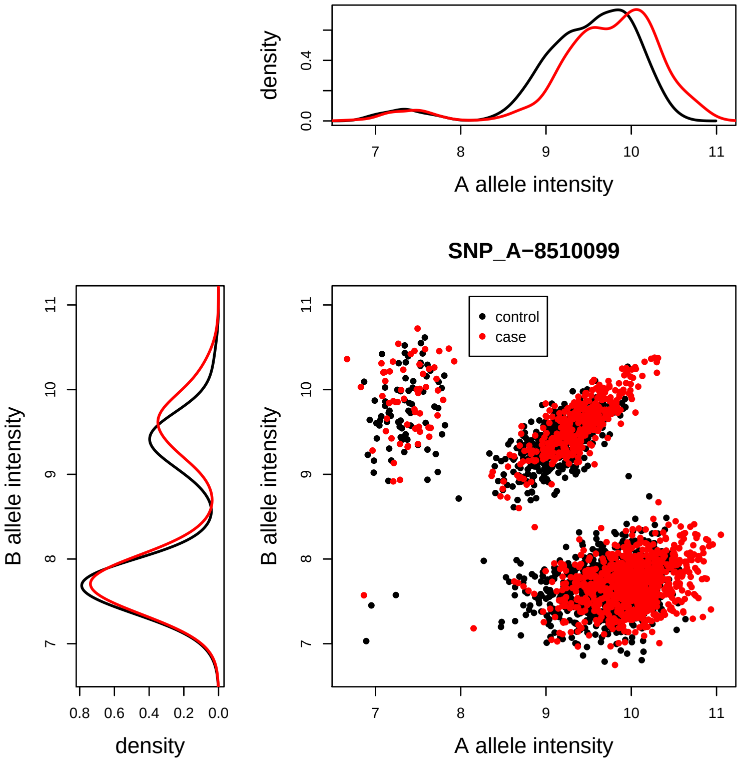

Figure 1.1: An example of Affymetrix intensity data at a SNP site showing differential errors. The data are from a GWAS of schizophrenia (Shi et al., 2009).

Chapter 2

A General Framework for Studying Genetic Effects and Gene-Environment Interactions with Missing Data

2.1 Introduction

In this chapter, we extend the work of Lin and Zeng (2006) to allow gene-environment dependence and to handle untyped SNPs. We provide a unified framework for assessing the roles of individual SNPs (including untyped SNPs) or their haplotypes in the development of disease. The effects of genetic and environmental factors on disease phenotypes are formulated through flexible regression models that incorporate appropriate genetic mechanisms and gene-environment interactions. The dependence between genetic and environmental factors is characterized by a class of odds-ratio functions. The marginal distribution of environmental factors is completely unspecified, while genetic variables may be in Hardy-Weinberg equilibrium or disequilibrium. We construct appropriate likelihoods for all commonly used study designs (including cross-sectional, case-control, and cohort designs) and a variety of disease phenotypes/traits. Unlike the case of gene-environment independence, the likelihoods involve the (potentially infinite-dimensional) distribution of environmental variables even under cross-sectional and cohort designs and are thus difficult to handle both theoretically and numerically. We establish the theoretical properties of the maximum likelihood estimators by appealing to modern asymptotic techniques, and develop efficient and stable numerical algorithms to implement the corresponding inference procedures. We evaluate the proposed methods through extensive simulation studies and apply them to a major GWAS of lung cancer (Amos et al., 2008).

2.2 Methods

2.2.1 Notation and Assumptions

We consider a set of SNPs that are in linkage disequilibrium (i.e., correlated). We may have a direct interest in the haplotypes of these SNPs or wish to use the haplotype distribution to infer the unknown value of one SNP from the observed values of the other SNPs. Let H and G denote the diplotype (i.e., the pair of haplotypes on the two homologous chromosomes) and genotype, respectively. We write $H = (h, h')$ if the diplotype consists of h and h', in which case $G = h + h'$. We allow the values in G to be missing at random. Note that H cannot be determined with certainty on the basis of G if the two constituent haplotypes differ at more than one position or if any SNP genotype is missing.

Let $\mathbf{Y}$ and $\mathbf{X}$ denote, respectively, the phenotype of interest and the environmental factors or covariates. We allow $\mathbf{X}$ to include both covariates that are potentially correlated with H and those known to be independent of H. For cross-sectional and case-control studies, the effects of $\mathbf{X}$ and H on $\mathbf{Y}$ are characterized by the conditional density of $\mathbf{Y} = \mathbf{y}$ given $\mathbf{X} = \mathbf{x}$ and $H = (h, h')$, denoted by $P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{y}|\mathbf{x}, (h, h'))$, where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ pertain to intercept(s), regression parameters, and nuisance parameters (e.g., variance and overdispersion parameters), respectively. The regression effects are specified through the design vector $\mathcal{Z}(\mathbf{X}, H)$, which is a vector-function of $\mathbf{X}$ and H. For example, if we are interested in the additive genetic effect of a risk haplotype h^* and its interactions with $\mathbf{X}$, then we may specify

$$\mathcal{Z}(\mathbf{x}, (h, h')) = \begin{bmatrix} I(h = h^*) + I(h' = h^*) \\ \mathbf{x} \\ \{I(h = h^*) + I(h' = h^*)\}\mathbf{x} \end{bmatrix}, \tag{2.1}$$

where $I(\cdot)$ is the indicator function. For dominant and recessive models, we replace $I(h = h^*) + I(h' = h^*)$ by $I(h = h^* \text{ or } h' = h^*)$ and $I(h = h' = h^*)$, respectively; the co-dominant model contains both additive and recessive effects. If we are interested in the additive effect of a particular SNP, then we replace $I(h = h^*) + I(h' = h^*)$ by the value of $(h + h')$ at that SNP position; dominant, recessive and co-dominant effects are defined similarly.

Let K be the total number of haplotypes that exist in the population. For $k = 1, \ldots, K$, we denote the kth haplotype by h_k. Define $\pi_{kl} = \Pr(H = (h_k, h_l))$ and $\pi_k = \Pr(h = h_k)$, $k, l = 1, \ldots, K$. Under HWE,

$$\pi_{kl} = \pi_k \pi_l, \quad k, l = 1, \ldots, K. \tag{2.2}$$

We also consider two forms of Hardy-Weinberg disequilibrium (HWD),

$$\pi_{kl} = (1 - \rho)\pi_k \pi_l + \delta_{kl} \rho \pi_k, \tag{2.3}$$

and

$$\pi_{kl} = \frac{(1-\rho+\delta_{kl}\rho)\pi_k\pi_l}{1-\rho+\rho\sum_{j=1}^{K}\pi_j^2}, \tag{2.4}$$

where $0 < \pi_k \leq 1$, $\sum_{k=1}^{K}\pi_k = 1$, $\delta_{kk} = 1$, and $\delta_{kl} = 0$ $(k \neq l)$ (Lin and Zeng, 2006). Both (2.3) and (2.4) reduce to (2.2) if $\rho = 0$. Excess homozygosity (i.e., $\pi_{kk} > \pi_k^2, k = 1, \ldots, K$) and excess heterozygosity (i.e., $\pi_{kk} < \pi_k^2, k = 1, \ldots, K$) arise when $\rho > 0$ and $\rho < 0$, respectively, although the range of heterozygosity is restrictive. Denote the probability function of H by $P_{\boldsymbol{\gamma}}(\cdot)$, where $\boldsymbol{\gamma}$ consists of $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^{\mathrm{T}}$ under (2.2) and $\boldsymbol{\pi}$ and ρ under (2.3) or (2.4).

We formulate the dependence of $\mathbf{X}$ on H through the conditional density function $P(\mathbf{X}|H)$. Because of missing genetic data, $P(\mathbf{X}|H)$ cannot be completely nonparametric. Mimicking Chen (2004)'s idea, we define the general odds-ratio function

$$\eta(\mathbf{X}, \mathbf{x}_0, H, (h_0, h_0')) = \frac{P(\mathbf{X}|H)P(\mathbf{x}_0|h_0, h_0')}{P(\mathbf{X}|h_0, h_0')P(\mathbf{x}_0|H)},$$

where (h_0, h_0') and $\mathbf{x}_0$ are fixed points in the sample spaces of H and $\mathbf{X}$, respectively. Then

$$P(\mathbf{X}|H) = \frac{\eta(\mathbf{X}, \mathbf{x}_0, H, (h_0, h_0'))P(\mathbf{X}|h_0, h_0')}{\int_{\mathbf{x}} \eta(\mathbf{x}, \mathbf{x}_0, H, (h_0, h_0'))P(\mathbf{x}|h_0, h_0')d\mathbf{x}},$$

so the conditional density function is represented by the odds ratio function η and the conditional density at a fixed point $P(\mathbf{X}|h_0, h_0')$. We abbreviate $P(\mathbf{x}|h_0, h_0')$ as $f(\mathbf{x})$ and denote the corresponding distribution function by $F(\mathbf{x})$.

Without loss of generality, set $(h_0, h_0') = (h_K, h_K)$. If $\mathbf{X}$ consists of S components that are either continuous or dichotomous, then we may specify that

$$\log \eta(\mathbf{x}, \mathbf{x}_0, (h_k, h_l), (h_K, h_K)) = \sum_{s=1}^{S} \zeta_{s,k,l}(x_s - x_{0,s}),$$

where $\mathbf{x} = (x_1, \ldots, x_S)^{\mathrm{T}}$, $\mathbf{x}_0 = (x_{0,1}, \ldots, x_{0,S})^{\mathrm{T}}$, and $\zeta_{s,k,l}$ $(s = 1, \ldots, S;\ k, l = 1, \ldots, K)$

are log odds ratios with $\zeta_{s,K,K} = 0$. Any categorical covariate of l levels can be represented by $(l-1)$ dichotomous variables. Specific mode of inheritance is imposed on $\zeta_{s,k,l}$ $(k, l = 1, \ldots, K)$ to ensure identifiability. Under the additive model, $\zeta_{s,k,l} = \zeta_{s,k} + \zeta_{s,l}$ with $\zeta_{s,K} = 0$. If a certain component of $\mathbf{X}$, indexed by s', is known to be independent of H, then we set the corresponding $\zeta_{s',k,l}$ $(k, l = 1, \ldots, K)$ to 0. In general, $\log \eta(\mathbf{x}, \mathbf{x}_0, (h_k, h_l), (h_K, h_K)) = \boldsymbol{\zeta}^{\mathrm{T}} \mathcal{D}(\mathbf{x}, h_k, h_l)$, where $\boldsymbol{\zeta}$ is a set of log-odds ratio parameters, and $\mathcal{D}(\mathbf{x}, h_k, h_l)$ is a set of distance measures. This formulation encompasses all generalized linear models for $\mathbf{X}$ with canonical links to H.

REMARK 2.1 Chen et al. (2008) assumed HWE and decomposed the joint density function $P(\mathbf{X}, H)$ as $P(H|\mathbf{X})P(\mathbf{X})$. Because $P(H|\mathbf{X})$ generally does not follow HWE when $P(H)$ is in HWE, Chen et al. (2008) defined the intercepts in their polytomous logistic model for $P(H|\mathbf{X})$ as implicit functions of all other parameters so as to impose HWE on $P(H)$. Those constraints complicate the estimation process. By contrast, we decompose $P(\mathbf{X}, H)$ as $P(\mathbf{X}|H)P(H)$, so that the population genetics assumption on $P(H)$ can be incorporated directly and there are no constraints on other parameters. The odds ratios associated with $P(\mathbf{X}|H)$ and $P(H|\mathbf{X})$ are the same and can be interpreted as the effects of H on $\mathbf{X}$ or the effects of $\mathbf{X}$ on H.

In the sequel, $\mathcal{S}(G)$ denotes the set of diplotypes that are compatible with genotype G, $h^{\dagger}$ denotes a haplotype that differs from h at only one SNP site, and $\nabla_{\mathbf{u}} f(\mathbf{u}, \mathbf{v}) = \partial \mathbf{f}(\mathbf{u}, \mathbf{v}) / \partial \mathbf{u}$. For any parameter $\boldsymbol{\theta}$, we use $\boldsymbol{\theta}_0$ to denote its true value when the distinction is necessary. We assume that the true value of any Euclidean parameter $\boldsymbol{\theta}$ belongs to the interior of a known compact set within the domain of $\boldsymbol{\theta}$ and that F_0 is twice-continuously differentiable with positive derivatives in its support.

2.2.2 Cross-Sectional Studies

In a cross-sectional study, we measure the phenotype $\mathbf{Y}$, genotype G and covariates $\mathbf{X}$ on a random sample of n subjects, so the data consist of $(\mathbf{Y}_i, \mathbf{X}_i, G_i)$ $(i = 1, \ldots, n)$. The phenotype or trait $\mathbf{Y}$ can be any type (e.g., binary or continuous) and possibly multivariate. As mentioned in Section 2.2.1, the conditional density of $\mathbf{Y}$ given $\mathbf{X}$ and H is given by $P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}|\mathbf{X}, H)$, which can be formulated by generalized linear models for univariate traits and by generalized linear mixed models for multivariate traits.

Write $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\zeta})$. The likelihood for $\boldsymbol{\theta}$ and F is

$$L_n(\boldsymbol{\theta}, F) = \prod_{i=1}^{n} \sum_{H \in \mathcal{S}(G_i)} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}_i|\mathbf{X}_i, H) P_{\boldsymbol{\zeta},F}(\mathbf{X}_i|H) P_{\boldsymbol{\gamma}}(H), \tag{2.5}$$

where

$$P_{\boldsymbol{\zeta},F}(\mathbf{x}|h, h') = \frac{\exp\{\boldsymbol{\zeta}^{\mathrm{T}} \mathcal{D}(\mathbf{x}, h, h')\} f(\mathbf{x})}{\int_{\widetilde{\mathbf{x}}} \exp\{\boldsymbol{\zeta}^{\mathrm{T}} \mathcal{D}(\widetilde{\mathbf{x}}, h, h')\} dF(\widetilde{\mathbf{x}})}.$$

We use the NPMLE approach. In this approach, the distribution function $F(\cdot)$ is treated as a right-continuous function with jumps at the observed $\mathbf{X}$. The objective function to be maximized is obtained from (2.5) by replacing $f(\mathbf{x})$ with the jump size of F at $\mathbf{x}$. The maximization can be carried out by the EM algorithm described in Section 2.6.1.

2.2.3 Case-Control Studies

In a case-control study, we measure $\mathbf{X}$ and G on n_1 cases $(Y = 1)$ and n_0 controls $(Y = 0)$. It is natural to formulate the effects of $\mathbf{X}$ and G on Y through the logistic regression model

$$P_{\alpha,\boldsymbol{\beta}}(Y|\mathbf{X}, H) = \frac{\exp\{Y(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}, H))\}}{1 + \exp\{\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}, H)\}}, \tag{2.6}$$

where α is an intercept and $\boldsymbol{\beta}$ is a set of log odds ratios.

Write $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta})$. To reflect case-control sampling, we employ the retrospective likelihood:

$$L_n(\boldsymbol{\theta}, F) = \prod_{i=1}^{n} \frac{\sum_{H \in \mathcal{S}(G_i)} P_{\alpha,\boldsymbol{\beta}}(Y_i|\mathbf{X}_i, H) P_{\boldsymbol{\zeta},F}(\mathbf{X}_i|H) P_{\boldsymbol{\gamma}}(H)}{\int_{\mathbf{x}} \sum_H P_{\alpha,\boldsymbol{\beta}}(Y_i|\mathbf{x}, H) P_{\boldsymbol{\zeta},F}(\mathbf{x}|H) P_{\boldsymbol{\gamma}}(H) d\mathbf{x}}. \tag{2.7}$$

There is very little information about α in case-control data, so the problem is virtually non-identifiable. We focus on two tractable situations: when the disease is rare, and when the disease rate is known. Under such conditions, the haplotype distribution of the general population can be estimated reliably from case-control data.

Rare Disease

When the disease is rare, model (2.6) simplifies to $P_{\alpha,\boldsymbol{\beta}}(Y|\mathbf{X}, H) \approx \exp\big\{Y\big(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}, H)\big)\big\}$. Then the likelihood given in (2.7) becomes

$$L_n(\boldsymbol{\theta}, F) = \prod_{i=1}^{n} \left\{ \frac{\sum_{H \in \mathcal{S}(G_i)} \exp\{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}_i, H)\} P_{\boldsymbol{\zeta},F}(\mathbf{X}_i|H) P_{\boldsymbol{\gamma}}(H)}{\int_{\mathbf{x}} \sum_H \exp\{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{x}, H)\} P_{\boldsymbol{\zeta},F}(\mathbf{x}|H) P_{\boldsymbol{\gamma}}(H) d\mathbf{x}} \right\}^{Y_i}$$

$$\times \left\{ \sum_{H \in \mathcal{S}(G_i)} P_{\boldsymbol{\zeta},F}(\mathbf{X}_i|H) P_{\boldsymbol{\gamma}}(H) \right\}^{1-Y_i}, \tag{2.8}$$

in which $\boldsymbol{\theta}$ consists of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\zeta}$ only. We again adopt the NPMLE approach, which is implemented via the EM algorithm described in Section 2.6.2.

Known Disease Rate

Let p_1 be the known disease rate. We maximize the likelihood given in (2.7) or equivalently

$$L_n(\boldsymbol{\theta}, F) = \prod_{i=1}^{n} \sum_{H \in \mathcal{S}(G_i)} P_{\alpha,\boldsymbol{\beta}}(Y_i|\mathbf{X}_i, H) P_{\boldsymbol{\zeta},F}(\mathbf{X}_i|H) P_{\boldsymbol{\gamma}}(H)$$

subject to the constraint that $\int_{\mathbf{x}} \sum_H P_{\alpha,\boldsymbol{\beta}}(Y = 1|\mathbf{x}, H) P_{\boldsymbol{\zeta},F}(\mathbf{x}|H) P_{\boldsymbol{\gamma}}(H) d\mathbf{x} = p_1$. We show in Section 2.6.3 that the NPMLEs of $\boldsymbol{\theta}$ and F can be obtained via an EM algorithm.

REMARK 2.2 Chen et al. (2008) also focused on the situations of rare disease and known disease rate. Because their estimating equations are not likelihood score equations and involve constraints for the intercepts of their polytomous logistic model, the convergence properties of their EM-like algorithm are unclear, and their estimators are not asymptotically efficient. By contrast, our objective functions are likelihood functions, which are guaranteed to increase at each step of the EM algorithms, and the resulting estimators are asymptotically efficient.

2.2.4 Cohort Studies

In a cohort study, we follow a random sample of n at-risk subjects to observe their ages at onset of disease. The subjects who are disease-free during the follow-up contribute censored observations. Let Y and C denote the time to disease occurrence and the censoring time, respectively. It is assumed that C is independent of Y and H conditional on $\mathbf{X}$ and G. The data consist of $(\widetilde{Y}_i, \Delta_i, \mathbf{X}_i, G_i),\ i = 1, \ldots, n$, where $\widetilde{Y}_i = \min(Y_i, C_i)$, and $\Delta_i = I(Y_i \leq C_i)$.

We formulate the effects of $\mathbf{X}$ and H on Y through a class of semiparametric transformation models

$$\Lambda(t|\mathbf{X}, H) = Q(\Lambda(t) e^{\beta^{\mathrm{T}} \mathcal{Z}(\mathbf{X}, H)}),$$

where $\Lambda(\cdot|\mathbf{X}, H)$ is the cumulative hazard function of Y given $\mathbf{X}$ and H, $\Lambda(\cdot)$ is an unspecified increasing function, and $Q(\cdot)$ is a three-time differentiable function with $Q(0) = 0$ and $Q'(x) > 0$ and satisfying condition (e) of Zeng and Lin (2007). Here and in the sequel, $g'(x) = dg(x)/dx$ and $g''(x) = d^2 g(x)/dx^2$. The choices of $Q(x) = x$

and $Q(x) = \log(1+x)$ yield the proportional hazards model (Cox, 1972) and the proportional odds model (Bennett, 1983), respectively.

Write $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta})$. The likelihood concerning $\boldsymbol{\theta}$, Λ and F takes the form

$$L_n(\boldsymbol{\theta}, \Lambda, F) = \prod_{i=1}^{n} \sum_{H \in \mathcal{S}(G_i)} \Big\{\Lambda'(\widetilde{Y}_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}_i, H)} Q'(\Lambda(\widetilde{Y}_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}_i, H)})\Big\}^{\Delta_i}$$

$$\times \exp\Big\{-Q(\Lambda(\widetilde{Y}_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}_i, H)})\Big\} P_{\boldsymbol{\zeta}, F}(\mathbf{X}_i | H) P_{\boldsymbol{\gamma}}(H). \tag{2.9}$$

Adopting the NPMLE approach, we regard Λ and F as right-continuous functions and replace $\Lambda'(\widetilde{Y}_i)$ and $f(\mathbf{x})$ in (2.9) with the jump size of Λ at $\widetilde{Y}_i$ and the jump size of F at $\mathbf{x}$. The estimation can be carried out through EM algorithms; see Section 2.6.4.

2.2.5 Asymptotic Properties

The NPMLEs in Sections 2.2.2–2.2.4, denoted by $\widehat{\boldsymbol{\theta}}$, $\widehat{F}$ and $\widehat{\Lambda}$, are consistent, asymptotically normal, and asymptotically efficient; rigorous statements and proofs are provided in Theorems 2.1–2.4 of Section 2.6. The limiting covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be consistently estimated by inverting the information matrix for all parameters (including the jump sizes of nuisance functions) or by using the profile likelihood function (Murphy and van der Vaart, 2000).

2.2.6 Untyped SNPs

When one of the SNPs in G is untyped, i.e., missing on all study subjects, the haplotype distribution $\boldsymbol{\pi}$ cannot be estimated from the study data alone. Fortunately, external databases, such as the HapMap, can be used to estimate $\boldsymbol{\pi}$ provided that the external sample and the study sample are generated from the same underlying population.

Let $L_R(\boldsymbol{\pi})$ denote the likelihood for $\boldsymbol{\pi}$ based on the external sample. If the external sample consists of $\widetilde{n}$ unrelated subjects, then $L_R(\boldsymbol{\pi}) = \prod_{j=1}^{\widetilde{n}} \sum_{(h_k,h_l)\in S(G_j)} \pi_k\pi_l$, where G_j is the genotype of the jth subject. The HapMap database provides genotype information for trios. For an external sample of $\widetilde{n}$ trios, the genotype data for the jth trio consist of $G_j \equiv (GF_j, GM_j, GC_j)$ $(j = 1, \ldots, \widetilde{n})$, where GF_j, GM_j and GC_j denote the genotypes for the father, mother and child, respectively. Then

$$L_R(\boldsymbol{\pi}) = \prod_{j=1}^{\widetilde{n}} \sum_{(h_k,h_l,h_{k'},h_{l'})\in\mathcal{S}(G_j)} \pi_k\pi_l\pi_{k'}\pi_{l'},$$

where $(h_k, h_l, h_{k'}, h_{l'}) \in \mathcal{S}(G_j)$ means that (h_k, h_l) is compatible with GF_j, $(h_{k'}, h_{l'})$ is compatible with GM_j, and $(h_k, h_{k'})$, $(h_k, h_{l'})$, $(h_l, h_{k'})$ or $(h_l, h_{l'})$ is compatible with GC_j.

Denote the likelihood for the study data by $L_S(\boldsymbol{\theta})$, in which $\boldsymbol{\theta}$ consists of $\boldsymbol{\pi}$, as well as all other finite- and infinite-dimensional parameters in the likelihood. The likelihood for $\boldsymbol{\theta}$ that combines the study data and the external data is $L_C(\boldsymbol{\theta}) \equiv L_S(\boldsymbol{\theta})L_R(\boldsymbol{\pi})$. We maximize $L_C(\boldsymbol{\theta})$ in the same manner as in the maximization of $L_S(\boldsymbol{\theta})$; the score function and information matrix for $L_R(\boldsymbol{\pi})$ are provided in Appendix B of Lin et al. (2008). The resulting estimators of $\boldsymbol{\theta}$ are consistent, asymptotically normal and asymptotically efficient.

2.3 Simulation Studies

We conducted extensive simulation studies to assess the operating characteristics of the proposed methods in realistic scenarios. We considered 5 SNPs (rs10519198, rs13180, rs3743079, rs8034191 and rs3885951) in a gene on chromosome 15 that is known to affect both smoking behaviour and lung cancer (Amos et al., 2008). Table 2.1 displays the haplotype frequencies of the 5 SNPs. We simulated genotype data from those haplotype frequencies under HWE.

Our first set of studies was concerned with the inference on haplotype effects and haplotype-environment interactions in case-control studies. We simulated disease status from the logistic regression model with an additive effect of h_2:

$$\begin{aligned}&\text{logitPr}\{Y=1|X,H=(h,h')\}\\&\qquad=\alpha+\beta_1\{I(h=h_2)+I(h'=h_2)\}+\beta_2X+\beta_3\{I(h=h_2)+I(h'=h_2)\}X,\end{aligned}$$

where X is Bernoulli with $\Pr(X=1|(h_K,h_K))=.2$. We let $\log\eta(X,0,(h_k,h_l),(h_K,h_K))=(\zeta_{1,k}+\zeta_{1,l})X$, where $\zeta_{1,2}=0.2$, $\zeta_{1,4}=-0.2$, $\zeta_{1,9}=0.1$ and $\zeta_{1,k}=0$ $(k\neq 2,4,9)$.

For making inference on β_1, we set $\beta_2=.25$ and $\beta_3=.0$ and varied β_1 from $-.5$ to $.5$; for making inference on β_3, we set $\beta_1=\beta_2=.25$ and varied β_3 from $-.5$ to $.5$. We chose $\alpha=-3$ and -2.1 to yield disease rates between 5% and 15%. We let $n_1=n_0=500$ and adopted the rare disease assumption in the analysis. We also included the method of Lin and Zeng (2006), which assumes haplotype-environment independence. The results are summarized in Table 2.2.

The proposed estimator for β_1 is virtually unbiased. The proposed estimator for β_3 seems to be slightly biased downward when the disease rate is close to 15%. The proposed variance estimators accurately reflect the true variabilities, the Wald tests have proper type I error, and the confidence intervals have reasonable coverage probabilities. The rare-disease assumption is a good approximation even when the disease rate is as high as 15%. Under the Lin-Zeng method, the estimators are biased, the type I error is inflated, and the confidence intervals have poor coverage probabilities, especially for interactions.

To assess the efficiency loss of modelling gene-environment dependence when the independence assumption actually holds, we modified the above simulation set-up by letting $\boldsymbol{\zeta}=\mathbf{0}$. For making inference on β_1, we set $\alpha=-3$, $\beta_2=.25$ and $\beta_3=0$

and varied e^{β_1} from 1.3 to 1.6; for making inference on β_3, we set $\beta_1 = \beta_2 = .25$ and varied e^{β_3} from 1.5 to 2.3. As shown in Figure 2.1, the power loss is more substantial in testing interactions than in testing main effects. In practice, one should incorporate the independence assumption into the analysis if it is known to be true. Indeed, our formulation allows one to impose the independence on any subset of $\mathbf{X}$. If the independence is not known to hold or not, then the empirical Bayes-type shrinkage estimation (e.g., Chen et al., 2009) provides a nice trade-off between efficiency and robustness; see Section 2.6.5.

The aforementioned studies pertain to a binary covariate and to risk haplotype h_2, which has a relatively high frequency. Additional simulation studies revealed that the above conclusions continue to hold for other haplotype frequencies and other covariate distributions. For example, the left panel of Table 2.3 shows the results under the logistic regression model

$$\begin{aligned}
&\text{logitPr}\{Y = 1|X_1, X_2, (h, h')\} \\
&\qquad = \alpha + \beta_{h_2}\{I(h = h_2) + I(h' = h_2)\} + \beta_{h_1}\{I(h = h_1) + I(h' = h_1)\} \\
&+\beta_{x_1}X_1 + \beta_{x_2}X_2 + \beta_{x_1h_2}\{I(h = h_2) + I(h' = h_2)\}X_1 + \beta_{x_1h_1}\{I(h = h_1) + I(h' = h_1)\}X_1,
\end{aligned}$$

coupled with the odds ratio function $\log\eta((X_1, X_2), (0, 0), (h_k, h_l), (h_K, h_K)) = (\zeta_{1,k} + \zeta_{1,l})X_1$, where X_1 and X_2 are independent conditional on H, the conditional distribution of X_1 given $H = (h_K, h_K)$ is standard normal, X_2 is Bernoulli with .4 success probability, $\alpha = -3$, $\beta_{h_1} = \beta_{h_2} = .25$, $\beta_{x_1} = \beta_{x_2} = .3$, $\beta_{x_1h_2} = \beta_{x_1h_1} = .0$, $\zeta_{1,2} = 0.2$, $\zeta_{1,4} = -0.2$, $\zeta_{1,9} = 0.1$ and $\zeta_{1,k} = 0$ $(k \neq 2, 4, 9)$.

To assess the robustness of the proposed method, we modified the above setting to simulate a conditional distribution of $\mathbf{X}$ given H that does not fit into the odds ratio formulation. Specifically, we let the conditional density of X_1 given $H = (h_k, h_l)$ be $\zeta_k +$

$\zeta_l + t$, where t follows a 3 d.f. t-distribution truncated at ±5. The results are provided in the right panel of Table 2.3. The proposed method is robust to misspecification of the dependence structure.

We also compared the proposed method to that of Chen et al. (2008). We simulated data from the logistic regression model

$$\begin{aligned} &\text{logitPr}\{Y = 1|X, H = (h, h')\} \\ &\qquad = \alpha + \beta_1\{I(h = h_3) + I(h' = h_3)\} + \beta_2 X + \beta_3\{I(h = h_3) + I(h' = h_3)\}X, \end{aligned}$$

and the odds ratio function $\log \eta(X, 0, (h_k, h_l), (h_K, h_K)) = (\zeta_{1,k} + \zeta_{1,l})X$, where the conditional distribution of X given $H = (h_K, h_K)$ is standard normal, $\zeta_{1,3} = 0.2$, $\zeta_{1,4} = -0.2$, $\zeta_{1,9} = 0.1$ and $\zeta_{1,k} = 0$ $(k \neq 3, 4, 9)$. We set $n_1 = n_0 = 500$ and $\alpha = -3$. For making inference on β_1, we set $\beta_2 = 0.25$ and $\beta_3 = 0$ and varied e^{β_1} from 1.5 to 1.8; for making inference on β_3, we set $\beta_1 = \beta_2 = 0.25$ and varied e^{β_3} from 1.5 to 1.8. For each combination of simulation parameters, we generated 1,000 data sets. Our algorithm always converged, whereas the algorithm of Chen et al. (2008), as implemented in their SAS program, failed to converge in about 3% of the data sets. Figure 2.2 presents the power curves of the two methods based on the data sets in which the algorithm of Chen et al. converged. The proposed method is uniformly more powerful than Chen et al.'s, especially in detecting interactions.

Our final set of studies dealt with analysis of untyped SNPs in cohort studies. We simulated ages at onset of disease from the proportional hazards model $\Lambda(t|X, H) = t^2 e^{\beta_1 \mathcal{G}_4(H) + \beta_2 X + \beta_3 \mathcal{G}_4(H) X}$, where $\mathcal{G}_4(H)$ is the genotype induced by the diplotype H at the 4th locus, and X is the same as in the first set of case-control studies. We generated censoring times from the uniform $(0, \tau)$ distribution, where τ was chosen to yield approximately 250, 500 or 1,000 cases under n =5,000. We set $\beta_1 = \beta_2 = .25$ and varied β_3 from $-.5$ to .5. We set the 4th SNP to be missing in the observed data and generated an external data set of 30 trios from the haplotype distribution of Table 2.1. As shown in Table 2.4, the proposed method performs very well.

2.4 Lung Cancer Study

Lung cancer is the most common type of cancer in terms of both incidence and mortality, with the highest rates in Europe and North America. Although this malignancy is attributable to environmental exposures, primarily cigarette smoking, genetic factors influencing lung cancer susceptibility have been reported in numerous studies. Recently, a genome-wide case-control association study of histologically confirmed non-small cell lung cancer was conducted to identify common low-penetrance alleles influencing lung cancer risk (Amos et al., 2008). Controls were matched to cases according to smoking behavior, age (in 5-year groups) and sex, and former smokers were further matched by years of cessation. The study population was restricted to individuals of self-reported European descent to minimize confounding by ethnic variation.

In the discovery phase of the study, 1,154 ever-smoking cases and 1,137 ever-smoking controls were genotyped for 317,498 tagging SNPs on Illumina HumanHap300 v1.1 BeadChips. Two SNPs, rs1051730 and rs8034191, mapping to a region of strong linkage disequilibrium within 15q25.1 containing PSMA4 and the nicotinic acetylcholine receptor subunit genes CHRNA3 and CHRNA5, were found to be significantly associated with lung cancer risk. The investigators kindly provided us data on a cluster of 37 SNPs surrounding those two SNPs.

We first investigate haplotype effects and haplotype-smoking interactions with sliding windows of 5 SNPs. For each window, we fit a logistic regression model that compares all haplotypes (with observed frequencies greater than 0.2% in the control group) to the most frequent haplotype under the additive mode of inheritance and includes cigarettes per day as a continuous covariate. Because the SNPs in the region are known to be associated with smoking behavior, we allow all haplotypes (with observed frequencies greater than 0.4% in the control group) to be potentially correlated with the smoking variable in the proposed general odds-ratio function. We assume HWE and adopt the rare-disease approximation. For comparisons, we also fit the haplotype-environment independence model of Lin and Zeng (2006).

Table 2.5 presents the results for a window containing SNP rs1051730. Haplotype 11110 is significantly related to smoking. Haplotype 00000 also has a large effect on smoking, although not significant at the 0.05 level. For those two haplotypes, the Lin-Zeng method would declare statistical significance at the 0.05 level for haplotype-smoking interactions, whereas the proposed method would not. These differences are consistent with the simulation results shown in Table 2.2 that the Lin-Zeng method tends to produce false positive results for haplotype-environment interactions when the independence assumption fails.

Next, we investigate the effects of individual SNPs and their interactions with smoking in the development of lung cancer for the 37 typed SNPs and 259 untyped HapMap SNPs in the region. In accordance with the study sample, we choose the HapMap sample of Utah residents with ancestry from northern and western Europe as the reference panel in the analysis of untyped SNPs. For each untyped SNP, we identify a set of 4 typed SNPs within 100,000 base pairs that provides the best prediction (Lin et al., 2008). We apply the proposed and Lin et al. (2008) methods. The former allows gene-environment dependence whereas the latter assumes independence. For typed SNPs, we also perform standard logistic regression analysis, which allows any form of gene-environment dependence and thus serves as a benchmark. The dependence between smoking and SNPs in the region of interest turns out to be very strong; the results are not shown here. Figure 2.3 displays the results for testing SNP effects (adjusted for smoking) and for testing SNP-smoking interactions. For typed SNPs, the results based on the proposed method and standard logistic regression are highly similar, suggesting that our odds ratio formulation is reasonable; the results of the Lin et al. method are different, especially for interactions. For untyped SNPs, the Lin et al. method yields more significant results, especially for interactions, than the proposed method. Because of the strong gene-environment dependence, the results of the Lin et al. method are unreliable.

2.5 Discussion

This chapter extends the work of Lin and Zeng (2006) to allow gene-environment dependence and to handle untyped SNPs. As demonstrated in the simulation studies and real example, the results of association analysis depend critically on the assumption about gene-environment relationship. If the genetic and environmental factors are known to be independent, then one should impose this structure in the analysis to improve efficiency. If the independence does not hold, then one should avoid this assumption to enhance the validity of inference.

Unlike Lin and Zeng (2006), our likelihood functions involve the (potentially infinite-dimensional) distribution of covariates even for cross-sectional and cohort studies. Also, Lin and Zeng (2006) did not consider case-control studies with known disease rates. Even for case-control studies with rare disease, our likelihood function is more complicated than that of Lin and Zeng (2006) because the distribution of covariates cannot be profiled out due to the modeling of gene-environment dependence. Thus, our numerical algorithms are fundamentally different from those of Lin and Zeng (2006) for all study designs. Although the basic structures of our theoretical proofs are similar to those of Lin and Zeng (2006), the actual techniques employed are novel. Due to the presence of multiple nonparametric conditional distribution functions of $\mathbf{X}$ given H, the proofs of identifiability of parameters and nonsingularity of information matrices are very delicate.

Lin and Zeng (2006) considered the setting in which $\mathbf{X}$ is independent of H conditional on G. It is difficult to construct realistic scenarios in which $\mathbf{X}$ is independent of H conditional on G but not independent of H unconditionally. Indeed, G is equivalent to H if there is only a single SNP or H consists of (h, h) or $(h, h^{\dagger})$. It is more natural to allow direct association between H and $\mathbf{X}$, as is done in this chapter.

We have assumed that $\mathbf{X}$ is completely observed. In practice, the values of certain environmental variables (e.g., smoking history and dietary information) may be unknown on some study subjects. A major advantage of the odds-ratio formulation is that it can readily handle missing covariates (Chen, 2004). Specifically, we express $P(\mathbf{X}|H)$ as $P(X_1|H)P(X_2|X_1, H) \times P(X_3|X_1, X_2, H)\ldots$, and represent each conditional density function in terms of a general odds ratio function and an arbitrary one-dimensional distribution function. In this way, we can accommodate arbitrary missing patterns in $\mathbf{X}$ and easily extend the theory and numerical algorithms of this chapter.

In the genetic and epidemiologic literature, it has become a common practice to infer the haplotypes or the values of untyped SNPs for each subject based on the genotype data alone and then include those imputed values in downstream association analysis. This is single imputation with improper posterior distributions and can yield biased estimates of genetic effects, inflated type I error and reduced statistical power (e.g., Lin and Huang, 2007; Lin et al., 2008).

We infer the unknown value of an untyped SNP nonparametrically from a small set of typed SNPs which is chosen to provide the best prediction among all flanking SNPs. An alternative approach is to use all typed SNPs on the chromosome under a population genetics model. To incorporate the latter approach into our framework, we let G denote all the SNPs on the chromosome and decompose G into the typed component G_t and the untyped component G_t. The joint density of the observed data $(\mathbf{Y}, \mathbf{X}, G_t)$ can be written as

$$P(\mathbf{Y}, \mathbf{X}, G_t) = \sum_{G_u} P(\mathbf{Y}|\mathbf{X}, G_t, G_u) P(\mathbf{X}|G_t, G_u) P(G_t, G_u).$$

We calculate $P(G_t, G_u)$ through a hidden Markov model (e.g., Marchini et al., 2007). It is difficult to correctly specify the regression model $P(\mathbf{Y}|\mathbf{X}, G_t, G_u)$. For estimating the marginal effect of an untyped SNP, we include only that SNP in the regression model. Even when we are interested in the marginal effect of a single SNP, we need to include all the SNPs on the chromosome that are correlated with $\mathbf{X}$ in $P(\mathbf{X}|G_t, G_u)$. Inclusion of a large number of SNPs is computationally infeasible and statistically inefficient, whereas omission of important SNPs can bias the association analysis. We prefer the flanking SNPs approach because it is computationally simpler and yield more robust and possibly more efficient inference.

Table 2.1: Observed haplotype frequencies from a lung cancer study

Index	Haplotype	Frequency
h_1	00000	.0278
h_2	00010	.2101
h_3	00011	.0923
h_4	01000	.2080
h_5	01001	.0005
h_6	01010	.0026
h_7	10010	.0078
h_8	10011	.0083
h_9	11100	.1465
h_{10}	11110	.0158
h_{11}	10000	.2803

Table 2.2: Simulation results for estimating and testing haplotype effects and haplotype-environment interactions in case-control studies

		Proposed					Lin-Zeng				
α	β_1	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
-2.1	-.5	.001	.138	.137	.989	.861	-.051	.131	.131	.985	.955
	-.25	.000	.132	.132	.989	.250	-.049	.125	.125	.987	.433
	0	.003	.129	.127	.990	.010	-.047	.121	.120	.985	.015
	.25	.002	.123	.125	.993	.287	-.047	.114	.117	.988	.198
	.5	.002	.122	.123	.992	.940	-.046	.114	.114	.982	.918
-3	-.5	-.001	.138	.139	.992	.863	-.052	.131	.132	.988	.951
	-.25	.002	.133	.133	.988	.239	-.048	.126	.126	.985	.416
	0	.003	.127	.128	.993	.007	-.048	.119	.120	.985	.015
	.25	.003	.123	.124	.991	.290	-.047	.116	.116	.982	.203
	.5	.000	.124	.122	.991	.941	-.050	.114	.113	.984	.916
α	β_3										
-2.1	-.5	-.003	.270	.270	.992	.243	.284	.190	.193	.842	.052
	-.25	-.010	.261	.260	.989	.052	.255	.178	.178	.857	.011
	0	-.004	.259	.254	.990	.010	.217	.167	.167	.891	.109
	.25	-.004	.253	.251	.991	.051	.161	.158	.158	.937	.519
	.5	-.017	.257	.252	.989	.250	.082	.149	.151	.981	.899
-3	-.5	-.001	.273	.270	.989	.227	.248	.194	.193	.883	.079
	-.25	-.002	.256	.259	.988	.051	.238	.176	.178	.880	.009
	0	-.002	.255	.251	.988	.012	.221	.164	.165	.882	.118
	.25	-.003	.245	.246	.991	.052	.195	.155	.155	.901	.612
	.5	-.010	.249	.243	.989	.282	.154	.148	.148	.936	.967

NOTE: Bias and SE are the bias and standard error of the parameter estimator. SEE is the mean of the standard error estimator. CP is the coverage probability of the 99% confidence interval. Power pertains to the .01-level test of zero parameter value. Each entry is based on 5,000 replicates.

Table 2.3: Simulation results for estimating and testing haplotype effects and haplotype-environment interactions in case-control studies with two risk haplotypes and two covariates

		Correctly specified $P(\mathbf{X}\|H)$					Misspecified $P(\mathbf{X}\|H)$				
Para.	True Value	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
β_{h_2}	.25	.000	.116	.114	.992	.361	.010	.113	.114	.989	.378
β_{h_1}	.25	.003	.288	.283	.990	.041	.013	.298	.287	.989	.045
β_{x_1}	.3	.003	.084	.083	.991	.859	.014	.060	.059	.988	.997
β_{x_2}	.3	-.005	.129	.130	.991	.377	.001	.131	.132	.989	.385
$\beta_{x_1h_2}$	.0	-.002	.109	.105	.987	.013	-.017	.070	.071	.989	.011
$\beta_{x_1h_1}$	.0	.005	.267	.269	.991	.009	-.008	.181	.182	.990	.010

NOTE: See the Note to Table 2.2.

Table 2.4: Simulation results for the analysis of an untyped SNP in cohort studies

β_3	Cases	Bias	SE	SEE	CP	Power
0	250	-.003	.236	.233	.990	.010
	500	.004	.164	.163	.992	.008
	1,000	.001	.120	.120	.990	.010
-.25	250	-.003	.262	.256	.991	.049
	500	.003	.180	.178	.988	.112
	1,000	.001	.130	.129	.990	.254
-.5	250	-.009	.295	.285	.990	.194
	500	-.000	.203	.197	.991	.491
	1,000	.001	.144	.142	.989	.842
.25	250	.001	.217	.215	.991	.077
	500	.003	.154	.153	.991	.177
	1,000	.000	.114	.115	.992	.345
.5	250	.000	.203	.202	.991	.457
	500	.002	.147	.146	.991	.813
	1,000	-.003	.113	.112	.991	.973

NOTE: See the Note to Table 2.2.

Table 2.5: Estimates of haplotype effects and haplotype-smoking interactions for a set of 5 SNPs in the lung cancer study

Parameters	Proposed	Lin-Zeng
Logistic disease-risk model ($\boldsymbol{\beta}$)		
11110	.249(.069)**	.252(.069)**
11011	-.097(.084)	-.099(.084)
00000	.198(.139)	.201(.139)
11010	-.255(.237)	-.252(.237)
00011	.519(.737)	.536(.748)
smoking	.093(.090)	.021(.071)
11110×smoking	-.013(.069)	.094(.047)*
11011×smoking	-.032(.087)	-.061(.062)
00000×smoking	.108(.132)	.190(.086)*
11010×smoking	-.044(.236)	-.006(.181)
00011×smoking	.289(.349)	.290(.348)
General odds-ratio function ($\boldsymbol{\zeta}$)		
11110	.108(0.050)*	–
11011	-.030(.061)	–
00000	.083(.100)	–
11010	.038(.151)	–

NOTE: Standard error estimates are shown in parentheses. *$P < 0.05$. **$P < 0.001$.

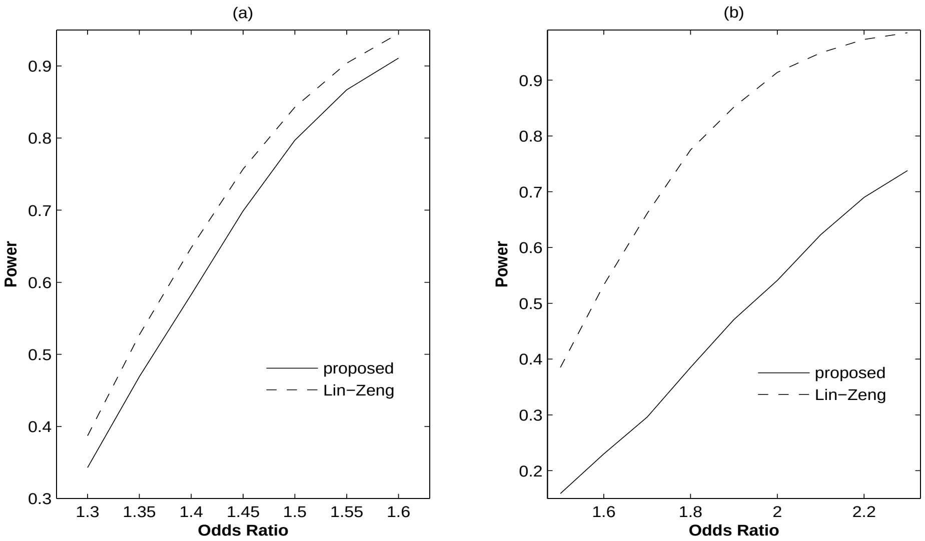

Figure 2.1: Power of testing (a) main effects and (b) interactions at the 1% nominal significance level for the proposed and Lin-Zeng methods when the independence assumption holds.

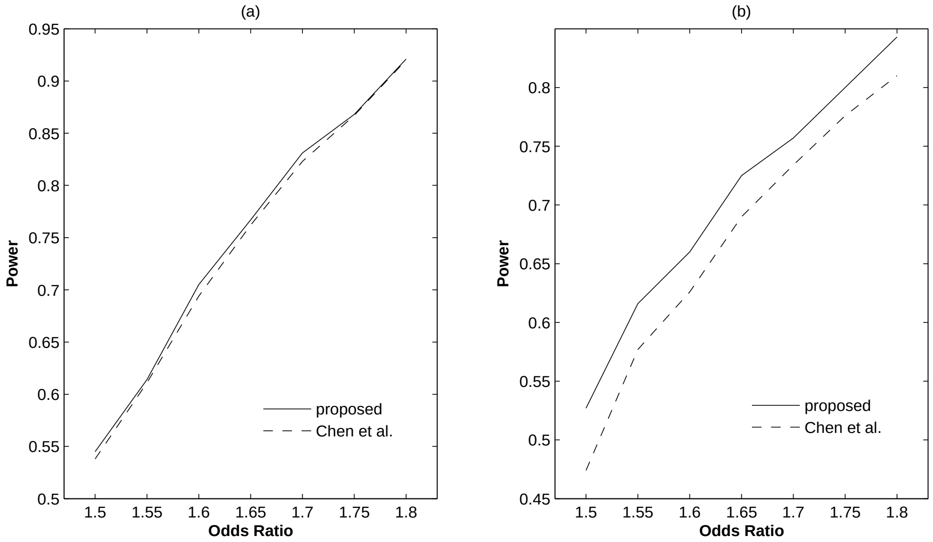

Figure 2.2: Power of testing (a) main effects and (b) interactions at the 1% nominal significance level for the proposed and Chen et al. methods.

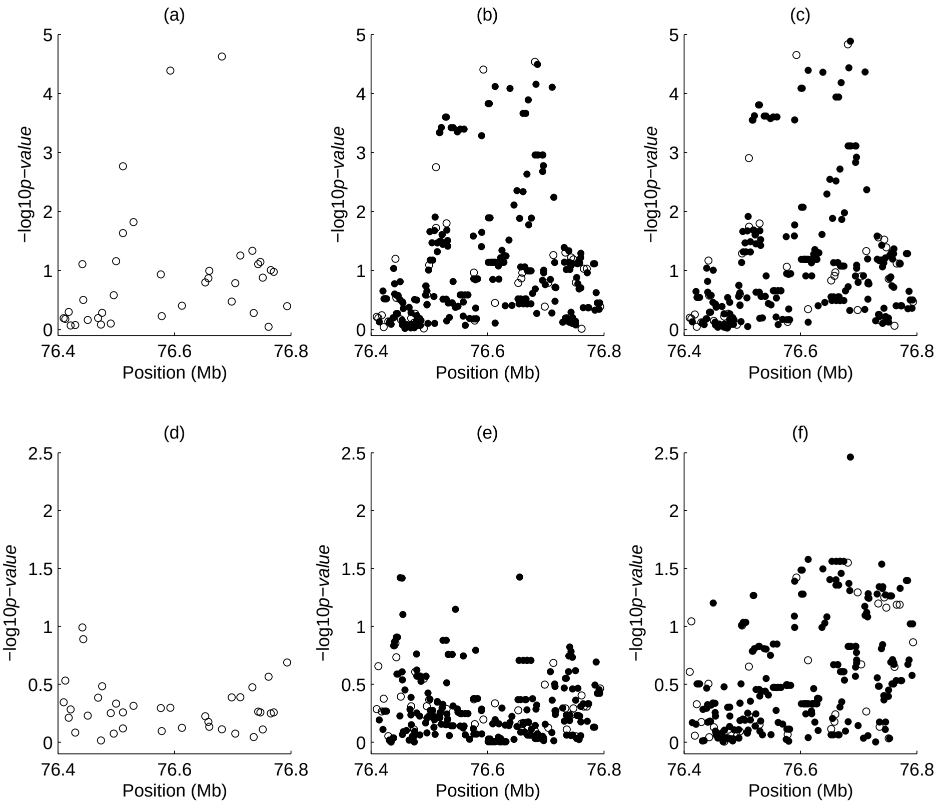

Figure 2.3: Results of association tests for additive effects of individual SNPs in the lung cancer study: the $-\log_{10}$(p-values) for the genotyped and untyped SNPs are shown in circles and dots, respectively; (a), (b) and (c) pertain to testing SNP effects (adjusted for smoking) under the standard logistic regression, the proposed method and the Lin et al. method, respectively; (d), (e) and (f) pertain to testing SNP-smoking interactions under the standard logistic regression, the proposed method and the Lin et al. method, respectively.

2.6 Appendix

In this section, we present the EM algorithms (treating H as missing data) for all the designs considered. We state in Theorems 2.1–2.4 the asymptotic properties of the NPMLEs described in Sections 2.2.2–2.2.4 and provide the proofs of the theorems. For each theorem, it is necessary to verify that the parameters are identifiable and the information matrices along all non-trivial parametric submodels are non-singular. We state those intermediate results in Lemmas 2.1–2.8.

2.6.1 Cross-Sectional Studies

Numerical Algorithm

Suppose that there are J distinct values of $\mathbf{X}$, denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_J$. Let $F\{\mathbf{x}_j\}$ be the jump size of F at $\mathbf{x}_j$. To incorporate the restriction that $\sum_j F\{\mathbf{x}_j\} = 1$, we estimate $\log(F\{\mathbf{x}_j\}/F\{\mathbf{x}_J\})$ $(j = 1, \ldots, J-1)$ instead. Define $\mathcal{D}_{jkl} = \mathcal{D}(\mathbf{x}_j, h_k, h_l)$, $\mathcal{Z}_{jkl} = \mathcal{Z}(\mathbf{x}_j, h_k, h_l)$,

$$\mathcal{W}_{kl} = \begin{pmatrix} I(h_k = h_1) + I(h_l = h_1) \\ \vdots \\ I(h_k = h_{K-1}) + I(h_l = h_{K-1}) \end{pmatrix}, \mathcal{M}_{jkl} = \begin{pmatrix} \mathcal{D}_{jkl} \\ I(j=1) \\ \vdots \\ I(j = J-1) \end{pmatrix}, \boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\zeta} \\ \log(F\{\mathbf{x}_1\}/F\{\mathbf{x}_J\}) \\ \vdots \\ \log(F\{\mathbf{x}_{(J-1)}\}/F\{\mathbf{x}_J\}) \end{pmatrix}.$$

To incorporate the constraint that $\sum_k \pi_k = 1$, we define $\nu_k = \log(\pi_k/\pi_K)$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{K-1})^{\mathrm{T}}$, so $P_{\gamma}(H = (h_k, h_l)) = \exp(\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl})/\sum_{k,l}\exp(\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl})$. Under $\mathbf{X} = \mathbf{x}_j$ and $H = (h_k, h_l)$,

$$\frac{\exp\{\boldsymbol{\zeta}^{\mathrm{T}}\mathcal{D}(\mathbf{X}, H)\}f(\mathbf{X})}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}^{\mathrm{T}}\mathcal{D}(\mathbf{x}, H)\}dF(\mathbf{x})} = \frac{\exp(\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{jkl})}{\sum_{j'}\exp(\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{j'kl})}.$$

The complete-data log-likelihood is

$$l_n^c = \sum_{i,j,k,l} I\{\mathbf{X}_i = \mathbf{x}_j, H_i = (h_k, h_l)\}\Big\{\log P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}_i|\mathbf{x}_j, (h_k, h_l)) + \boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl} + \boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{jkl}$$

$$- \log \sum_{j'} \exp(\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{j'kl})\Big\} - n \log \sum_{k,l} \exp(\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl}).$$

In the E-step, we evaluate $E\{I(\mathbf{X}_i = \mathbf{x}_j, H_i = (h_k, h_l))\big|\mathbf{X}_i, \mathbf{Y}_i, G_i\}$, which can be shown to be

$$\omega_{ijkl} \equiv \frac{I\{\mathbf{X}_i = \mathbf{x}_j, (h_k, h_l) \in \mathcal{S}(G_i)\}P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}_i|\mathbf{x}_j, (h_k, h_l))e^{\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl}+\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{jkl}} / \sum_{j'} e^{\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{j'kl}}}{\sum_{(h_{k'},h_{l'})\in\mathcal{S}(G_i)} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}_i|\mathbf{x}_j, (h_{k'}, h_{l'}))e^{\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{k'l'}+\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{jk'l'}} / \sum_{j'} e^{\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{j'k'l'}}}.$$

In the M-step, we maximize l_n^c with $I\{\mathbf{X}_i = \mathbf{x}_j, H_i = (h_k, h_l)\}$ replaced by ω_{ijkl}. The maximization is carried out by the quasi-Newton algorithm. Starting with $\boldsymbol{\alpha} = \mathbf{0}$, $\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{\delta} = \mathbf{0}$ and $\nu_k = \log(\widetilde{\pi}_k/\widetilde{\pi}_K)$ $(k = 1, \ldots, K-1)$, where the $\widetilde{\pi}_k$'s are the MLEs of the π_k's based on G_i $(i = 1, \ldots, n)$, we iterate between the E-step and M-step until the change in the observed-data log-likelihood is negligible.

We can estimate the limiting covariance matrix of $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}$ by inverting the (observed-data) information matrix for all the parameters including the jump sizes of $\widehat{F}$. The information matrix is obtained via the Louis (1982) formula. We can also estimate the limiting covariance matrix of $\widehat{\boldsymbol{\theta}}$ by using the profile likelihood function $pl_n(\boldsymbol{\theta}) \equiv \max_F \log L_n(\boldsymbol{\theta}, F)$. Particularly, the (s,t)th element of the inverse covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by $-\epsilon_n^{-2}\Big\{pl_n(\widehat{\boldsymbol{\theta}}+\epsilon_n\mathbf{e}_s+\epsilon_n\mathbf{e}_t)-pl_n(\widehat{\boldsymbol{\theta}}+\epsilon_n\mathbf{e}_s)-pl_n(\widehat{\boldsymbol{\theta}}+\epsilon_n\mathbf{e}_t)+pl_n(\widehat{\boldsymbol{\theta}})\Big\}$, where ϵ_n is a constant of order $n^{-1/2}$, and $\mathbf{e}_s$, and $\mathbf{e}_t$ are the sth and tth canonical vectors. We calculate $pl_n(\boldsymbol{\theta})$ via the EM algorithm by holding $\boldsymbol{\theta}$ constant in both the E-step and M-step.

Theoretical Results

We impose the following conditions.

CONDITION 2.1 If $P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}|\mathbf{X},H) = P_{\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\xi}}}(\mathbf{Y}|\mathbf{X},H)$ for any $H=(h,h)$ and $H=(h,h^{\dagger})$, then $\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}$, $\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}$, and $\boldsymbol{\xi}=\widetilde{\boldsymbol{\xi}}$.

CONDITION 2.2 If there exists a constant vector $\boldsymbol{\nu}$ such that $\boldsymbol{\nu}^{\mathrm{T}}\nabla_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\log P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}|\mathbf{X},H) = 0$ for any $H=(h,h)$ and $H=(h,h^{\dagger})$, then $\boldsymbol{\nu}=\mathbf{0}$.

CONDITION 2.3 If there exists a function $a(H)$ and a constant vector $\mathbf{b}$ such that $a(H)+\mathbf{b}^{\mathrm{T}}\mathcal{D}(\mathbf{X},H)=0$ with probability one, then $a=0$ and $\mathbf{b}=\mathbf{0}$.

REMARK 2.3 Condition 2.1 ensures that $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi})$ are identifiable from the genotype data while Condition 2.2 ensures nonsingularity of the information matrix. All commonly used regression models, particularly generalized linear (mixed) models with design vectors in the form of (2.1), satisfy these two conditions. Condition 2.3 pertains to the identifiability of $\boldsymbol{\zeta}$. This condition holds under all common modes of inheritance for the $\zeta_{s,k,l}$ provided that $\mathbf{X}$ is linearly independent given H.

LEMMA 2.1 If two sets of parameters $(\boldsymbol{\theta},F)$ and $(\widetilde{\boldsymbol{\theta}},\widetilde{F})$ yield the same joint distribution of the data, then $\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}$ and $F=\widetilde{F}$.

Proof: Suppose that

$$\sum_{H\in\mathcal{S}(G)} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(\mathbf{Y}|\mathbf{X},H)P_{\boldsymbol{\zeta},F}(\mathbf{X}|H)P_{\boldsymbol{\gamma}}(H) = \sum_{H\in\mathcal{S}(G)} P_{\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\xi}}}(\mathbf{Y}|\mathbf{X},H)P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H)P_{\widetilde{\boldsymbol{\gamma}}}(H).$$

Letting $G=2h$ or $G=h+h^{\dagger}$ and integrating over $\mathbf{Y}$ on both sides, we obtain

$$P_{\boldsymbol{\zeta},F}(\mathbf{X}|H)P_{\boldsymbol{\gamma}}(H) = P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H)P_{\widetilde{\boldsymbol{\gamma}}}(H).$$

Integrating over $\mathbf{X}$ on both sides then yields that $P_{\boldsymbol{\gamma}}(H)=P_{\widetilde{\boldsymbol{\gamma}}}(H)$. By Lemma 1 of Lin and Zeng (2006), $\boldsymbol{\gamma}=\widetilde{\boldsymbol{\gamma}}$. Thus, $P_{\boldsymbol{\zeta},F}(\mathbf{X}|H)=P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H)$. It follows from the definition of $P_{\boldsymbol{\zeta},F}(\mathbf{X}|H)$ that

$$\exp\{(\boldsymbol{\zeta}-\widetilde{\boldsymbol{\zeta}})^{\mathrm{T}}\mathcal{D}(\mathbf{X},H)\}\frac{f(\mathbf{X})}{\widetilde{f}(\mathbf{X})}=\frac{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF(\mathbf{x})}{\int_{\mathbf{x}}\exp\{\widetilde{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}d\widetilde{F}(\mathbf{x})}.$$

By setting $H=(h_0,h_0')$, we obtain $\mathcal{D}(\mathbf{X},H)=\mathbf{0}$, so the above equation reduces to $f(\mathbf{x})=\widetilde{f}(\mathbf{x})$ for any $\mathbf{x}$. It then follows from Condition 2.3 that $\boldsymbol{\zeta}=\widetilde{\boldsymbol{\zeta}}$. Therefore, $P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}(Y|\mathbf{X},H)=P_{\widetilde{\boldsymbol{\alpha}},\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\xi}}}(Y|\mathbf{X},H)$ for any $H=(h,h)$ or $H=(h,h^{\dagger})$. By Condition 2.1, $\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}$, $\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\xi}=\widetilde{\boldsymbol{\xi}}$.

LEMMA 2.2 If there exist a vector $\boldsymbol{\mu}_{\boldsymbol{\theta}}\equiv(\boldsymbol{\mu}_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}^{\mathrm{T}},\boldsymbol{\mu}_{\gamma}^{\mathrm{T}},\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}})^{\mathrm{T}}$ and a function $\psi(\mathbf{x})$ with $E[\psi(\mathbf{X})]=0$ such that $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}}l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0,F_0)+l_{F_0}(\boldsymbol{\theta}_0,F_0)[\int\psi\ dF_0]=0$, where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$, and $l_{F_0}[\int\psi\ dF_0]$ is the score function for F along the submodel $F_0+\epsilon\int\psi\ dF_0$ with scalar ϵ, then $\boldsymbol{\mu}_{\boldsymbol{\theta}}=\mathbf{0}$ and $\psi=0$.

Proof: We wish to verify that if there exist a vector $\boldsymbol{\mu}_{\boldsymbol{\theta}}\equiv(\boldsymbol{\mu}_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}^{\mathrm{T}},\boldsymbol{\mu}_{\gamma}^{\mathrm{T}},\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}})^{\mathrm{T}}$ and a function $\psi(\mathbf{x})$ with $E[\psi(\mathbf{X})]=0$ such that

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}}l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0,F_0)+l_{F_0}(\boldsymbol{\theta}_0,F_0)[\int\psi\ dF_0]=0, \qquad (2.10)$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$, and $l_{F_0}[\int\psi\ dF_0]$ is the score function for F along the submodel $F_0+\epsilon\int\psi\ dF_0$ with scalar ϵ, then $\boldsymbol{\mu}_{\boldsymbol{\theta}}=\mathbf{0}$ and $\psi=0$. To this end, we set $G=2h$ or $G=h+h^{\dagger}$. Then (2.10) becomes

$$\begin{aligned}&\boldsymbol{\mu}_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}^{\mathrm{T}}\nabla_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\log P_{\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0,\boldsymbol{\xi}_0}(\mathbf{Y}|\mathbf{X},H)+\boldsymbol{\mu}_{\gamma}^{\mathrm{T}}\nabla_{\gamma}\log P_{\gamma_0}(H)\\ &+\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{X},H)-\frac{\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}}\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}\mathcal{D}(\mathbf{x},H)dF_0(\mathbf{x})}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})}\\ &+\psi(\mathbf{X})-\frac{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}\psi(\mathbf{x})dF_0(\mathbf{x})}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})}=0. \qquad (2.11)\end{aligned}$$

Taking the expectation with respect to $P_{\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0,\boldsymbol{\xi}_0}(\mathbf{Y}|\mathbf{X},H)$ yields

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}}\nabla_{\boldsymbol{\gamma}}\log P_{\boldsymbol{\gamma}_0}(H)+\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{X},H)-\frac{\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}}\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}\mathcal{D}(\mathbf{x},H)dF_0(\mathbf{x})}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})}$$

$$+\psi(\mathbf{X})-\frac{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}\psi(\mathbf{x})dF_0(\mathbf{x})}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})}=0. \qquad (2.12)$$

Since $\mathcal{D}(\mathbf{x},H)=\mathbf{0}$ for any $\mathbf{x}$ under $H=(h_0,h_0')$, we have

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}}\nabla_{\boldsymbol{\gamma}}\log P_{\boldsymbol{\gamma}_0}(h_0,h_0')+\psi(\mathbf{X})-\int_{\mathbf{x}}\psi(\mathbf{x})dF_0(\mathbf{x})=0.$$

This implies that $\psi(\mathbf{x})$ is constant over $\mathbf{x}$, so $\psi=0$. Thus, (2.12) reduces to

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}}\nabla_{\boldsymbol{\gamma}}\log P_{\boldsymbol{\gamma}_0}(H)+\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{X},H)-\frac{\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}}\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}\mathcal{D}(\mathbf{x},H)dF_0(\mathbf{x})}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})}=0.$$

By Condition 2.3, $\boldsymbol{\mu}_{\boldsymbol{\zeta}}=\mathbf{0}$. It then follows from Lemma 1 of Lin and Zeng (2006) that $\boldsymbol{\mu}_{\boldsymbol{\gamma}}=\mathbf{0}$. Hence, (2.11) reduces to $\boldsymbol{\mu}_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}^{\mathrm{T}}\nabla_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\log P_{\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0,\boldsymbol{\xi}_0}(\mathbf{Y}|\mathbf{X},H)=0$. By Condition 2.2, $\boldsymbol{\mu}_{\boldsymbol{\alpha}}=\mathbf{0}$, $\boldsymbol{\mu}_{\boldsymbol{\beta}}=\mathbf{0}$, and $\boldsymbol{\mu}_{\boldsymbol{\xi}}=\mathbf{0}$.

THEOREM 2.1 Under Conditions 2.1-2.3, $|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0|+\sup_{\mathbf{x}}|\widehat{F}(\mathbf{x})-F_0(\mathbf{x})|\to 0$ almost surely. In addition, $n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

Proof: We first prove the consistency of $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}$. Because $\widehat{\boldsymbol{\theta}}$ is bounded and $\widehat{F}$ is a distribution function, it follows from Helly's selection theorem that, for any subsequence of $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}$, there exists a further subsequence, still denoted as $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}$, such that $\widehat{\boldsymbol{\theta}}\to\boldsymbol{\theta}^*$ and $\widehat{F}\to F^*$ in distribution. It suffices to show $\boldsymbol{\theta}^*=\boldsymbol{\theta}_0$ and $F^*=F_0$. Since $\widehat{F}$ maximizes the likelihood function and its jump sizes are positive, there exists a Lagrange multiplier $\widehat{\lambda}$ such that

$$\frac{1}{\widehat{F}\{\mathbf{X}_k\}} - \sum_{i=1}^{n} \frac{\sum_{H\in\mathcal{S}(G_i)} P_{\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\xi}}}(\mathbf{Y}_i|\mathbf{X}_i,H)P_{\widehat{\boldsymbol{\gamma}}}(H)\frac{\exp\{\widehat{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{X}_i,H)\}\exp\{\widehat{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{X}_k,H)\}}{[\int_{\mathbf{x}}\exp\{\widehat{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}d\widehat{F}(\mathbf{x})]^2}}{\sum_{H\in\mathcal{S}(G_i)} P_{\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\xi}}}(\mathbf{Y}_i|\mathbf{X}_i,H)P_{\widehat{\boldsymbol{\gamma}}}(H)\frac{\exp\{\widehat{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{X}_i,H)\}}{\int_{\mathbf{x}}\exp\{\widehat{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}d\widehat{F}(\mathbf{x})}} - \widehat{\lambda} = 0,$$

where $\widehat{F}\{\mathbf{X}_k\}$ is the jump size of $\widehat{F}$ at $\mathbf{X}_k$. Due to the constraint that $\sum_k \widehat{F}\{\mathbf{X}_k\} = 1$, the above equation implies that $\widehat{\lambda} = 0$. Define $\widetilde{F}$ as a distribution function with jumps at the $\mathbf{X}_k$'s such that the jump size is proportional to

$$\left[\sum_{i=1}^{n} \frac{\sum_{H\in\mathcal{S}(G_i)} P_{\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0,\boldsymbol{\xi}_0}(\mathbf{Y}_i|\mathbf{X}_i,H)P_{\boldsymbol{\gamma}_0}(H)\frac{\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{X}_i,H)\}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{X}_k,H)\}}{[\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})]^2}}{P_{\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0,\boldsymbol{\xi}_0}(\mathbf{Y}_i|\mathbf{X}_i,H)P_{\boldsymbol{\gamma}_0}(H)\frac{\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{X}_i,H)\}}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})}}\right]^{-1}.$$

By the Glivenko-Cantelli theorem, $\widetilde{F}$ uniformly converges to F_0. In addition, $\widehat{F}$ is absolutely continuous with respect to $\widetilde{F}$, and $d\widehat{F}/d\widetilde{F}$ converges uniformly to some positive function g. Finally, since $n^{-1}\log\{L_n(\widehat{\boldsymbol{\theta}},\widehat{F})/L_n(\boldsymbol{\theta}_0,\widetilde{F})\} \geq 0$, we can take the limit as $n \to \infty$. Thus, the Kullback-Leibler information for $(\boldsymbol{\theta}^*, F^*)$ is non-positive, so the density under $(\boldsymbol{\theta}^*, F^*)$ is the same as the true density. It then follows from Lemma 2.1 that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $F^* = F_0$. This establishes the consistency of $(\widehat{\boldsymbol{\theta}}, \widehat{F})$. The weak convergence of $\widehat{F}$ to F_0 can be strengthened to the uniform convergence since F_0 is a continuous distribution function.

To derive the asymptotic distribution, we consider the score equation along the submodel $(\widehat{\boldsymbol{\theta}} + \epsilon\mathbf{v}, d\widehat{F} + \epsilon(\psi - \int \psi d\widehat{F}))$, where $\mathbf{v}$ is a vector with norm bounded by 1, and ψ is any function with $\int \psi dF_0 = 0$ and with total variation bounded by 1. The score equation takes the form

$$\sqrt{n}\,\boldsymbol{\Omega}_1(\mathbf{v},\psi)^{\mathrm{T}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \sqrt{n}\int \Omega_2(\mathbf{v},\psi)d(\widehat{F} - F_0) = \mathbf{G}_n\left\{l_{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{v} + l_F[\psi]\right\} + o_p(1),$$

where $\mathbf{G}_n$ denotes the empirical measure, $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}_0$, l_F is the score operator for F_0, $(\boldsymbol{\Omega}_1, \Omega_2)$ is a linear operator of the first-order Fredholm-type which maps $(\mathbf{v}, \psi)$ to the same space as $(\mathbf{v}, \psi)$, and $o_p(1)$ means a random variable converging in probability to zero uniformly in $\mathbf{v}$ and ψ. By some algebra, $(\boldsymbol{\Omega}_1, \Omega_2)[\mathbf{v}, \psi] = 0$ implies that the Fisher information along the submodel is zero, so $\mathbf{v} = \mathbf{0}$ and $\psi = 0$ by Lemma 2.2. Thus, $(\boldsymbol{\Omega}_1, \Omega_2)$ is invertible. We then verify all the conditions in Theorem 3.3.1 of van der Vaart and Wellner (1996). Hence, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{F} - F_0)$ weakly converges to a mean-zero Gaussian process.

In light of the above derivation, the influence function for $\widehat{\boldsymbol{\theta}}$ is a linear combination of some $l_{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{v} + l_F[\psi]$. Thus, the influence function lies on the tangent space spanned by the score functions and thus must be the efficient influence function. This means that $\widehat{\boldsymbol{\theta}}$ is asymptotically efficient in that its limiting covariance matrix attains the semiparametric efficiency bound.

2.6.2 Case-Control Studies with Rare Disease

Numerical Algorithm

We adopt the notation of Section 2.6.1. The E-step of the EM algorithm is the same as in Section 2.6.1. In the M-step, the objective function to be maximized is

$$\begin{aligned}\widetilde{l}_n(\boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{\delta}) = \sum_{i,j,k,l} \omega_{ijkl} \Big\{ Y_i \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}_{jkl} + \boldsymbol{\nu}^{\mathrm{T}} \mathcal{W}_{kl} + \boldsymbol{\delta}^{\mathrm{T}} \mathcal{M}_{jkl} - \log\Big(\sum_{j'} e^{\boldsymbol{\delta}^{\mathrm{T}} \mathcal{M}_{j'kl}} \Big) \Big\} \\ - n_1 \log \Big\{ \sum_{j,k,l} e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}_{jkl} + \boldsymbol{\nu}^{\mathrm{T}} \mathcal{W}_{kl}} \frac{e^{\boldsymbol{\delta}^{\mathrm{T}} \mathcal{M}_{jkl}}}{\sum_{j'} e^{\boldsymbol{\delta}^{\mathrm{T}} \mathcal{M}_{j'kl}}} \Big\} - n_0 \log \Big\{ \sum_{k,l} e^{\boldsymbol{\nu}^{\mathrm{T}} \mathcal{W}_{kl}} \Big\},\end{aligned}$$

where ω_{ijkl} is defined in Section 2.6.1. We use the Louis formula to calculate the observed-data information matrix, whose inverse is used to estimate the asymptotic covariance matrix of the NPMLEs; the profile likelihood method can also be used to estimate the covariance matrix of $\widehat{\boldsymbol{\theta}}$.

Theoretical Results

We impose the following identifiability condition.

CONDITION 2.4 If $\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X}, H) = \widetilde{\alpha} + \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(\mathbf{X}, H)$ for any $H = (h, h)$ and $H = (h, h^{\dagger})$, then $\alpha = \widetilde{\alpha}$ and $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$.

LEMMA 2.3 If two sets of parameters $(\boldsymbol{\theta}, F)$ and $(\widetilde{\boldsymbol{\theta}}, \widetilde{F})$ yield the same joint distribution, then $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$ and $F = \widetilde{F}$.

Proof: Suppose that

$$\left\{\frac{\sum_{H\in\mathcal{S}(G)}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X},H)\}P_{\boldsymbol{\zeta},F}(\mathbf{X}|H)P_{\boldsymbol{\gamma}}(H)}{\int_{\mathbf{x}}\sum_{H}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{x},H)\}P_{\boldsymbol{\zeta},F}(\mathbf{x}|H)P_{\boldsymbol{\gamma}}(H)d\mathbf{x}}\right\}^{Y}\left\{\sum_{H\in\mathcal{S}(G)}P_{\boldsymbol{\zeta},F}(\mathbf{X}|H)P_{\boldsymbol{\gamma}}(H)\right\}^{1-Y}$$

$$=\left\{\frac{\sum_{H\in\mathcal{S}(G)}\exp\{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(\mathbf{X},H)\}P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H)P_{\widetilde{\boldsymbol{\gamma}}}(H)}{\int_{\mathbf{x}}\sum_{H}\exp\{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(\mathbf{x},H)\}P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{x}|H)P_{\widetilde{\boldsymbol{\gamma}}}(H)d\mathbf{x}}\right\}^{Y}\left\{\sum_{H\in\mathcal{S}(G)}P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H)P_{\widetilde{\boldsymbol{\gamma}}}(H)\right\}^{1-Y}. \tag{2.13}$$

Setting $Y = 0$ and $G = 2h$ or $G = h + h^{\dagger}$ in (2.13), we obtain

$$P_{\boldsymbol{\zeta},F}(\mathbf{X}|H)P_{\boldsymbol{\gamma}}(H) = P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H)P_{\widetilde{\boldsymbol{\gamma}}}(H).$$

Integrating over $\mathbf{X}$ on both sides yields $P_{\boldsymbol{\gamma}}(H) = P_{\widetilde{\boldsymbol{\gamma}}}(H)$, so $\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}}$. Thus, $P_{\boldsymbol{\zeta},F}(\mathbf{X}|H) = P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H)$. By the arguments in the proof of Lemma 2.1, $f = \widetilde{f}$ and $\boldsymbol{\zeta} = \widetilde{\boldsymbol{\zeta}}$. Letting $Y = 1$ and $G = 2h$ or $G = h + h^{\dagger}$ in (2.13), we see that $\exp\{(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})^{\mathrm{T}}\mathcal{Z}(\mathbf{X}, H)\}$ must be a constant. It then follows from Condition 2.4 that $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$.

LEMMA 2.4 If there exist a vector $\boldsymbol{\mu}_{\boldsymbol{\theta}} \equiv (\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}, \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}}, \boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}})^{\mathrm{T}}$ and functions $\psi(\mathbf{x})$ with $E[\psi(\mathbf{X})] = 0$ such that

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, F_0) + l_F(\boldsymbol{\theta}_0, F_0)[\int \psi \, dF_0] = 0,$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$, and $l_F[\int \psi \, dF_0]$ is the score function for F along the submodel $F_0 + \epsilon \int \psi \, dF_0$, then $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{0}$ and $\psi = 0$.

Proof: We wish to show that if there exist a vector $\boldsymbol{\mu}_{\boldsymbol{\theta}} \equiv (\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}, \boldsymbol{\mu}_{\gamma}^{\mathrm{T}}, \boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}})^{\mathrm{T}}$ and functions $\psi(\mathbf{x})$ with $E[\psi(\mathbf{X})] = 0$ such that

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, F_0) + l_F(\boldsymbol{\theta}_0, F_0)[\int \psi \, dF_0] = 0, \tag{2.14}$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$, and $l_F[\int \psi \, dF_0]$ is the score function for F along the submodel $F_0 + \epsilon \int \psi \, dF_0$, then $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{0}$ and $\psi = 0$. To this end, we choose $Y = 0$ and $G = 2h$ or $G = h + h^{\dagger}$. Then (2.14) becomes

$$\boldsymbol{\mu}_{\gamma}^{\mathrm{T}} \nabla_{\gamma} \log P_{\gamma_0}(H) + \boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}} \mathcal{D}(\mathbf{X}, H) - \frac{\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}} \int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}} \mathcal{D}(\mathbf{x}, H)\} \mathcal{D}(\mathbf{x}, H) dF_0(\mathbf{x})}{\int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}} \mathcal{D}(\mathbf{x}, H)\} dF_0(\mathbf{x})}$$

$$+\psi(\mathbf{X}) - \frac{\int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}} \mathcal{D}(\mathbf{x}, H)\} \psi(\mathbf{x}) dF_0(\mathbf{x})}{\int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}} \mathcal{D}(\mathbf{x}, H)\} dF_0(\mathbf{x})} = 0. \tag{2.15}$$

With $H = (h_0, h_0')$, (2.15) reduces to $\boldsymbol{\mu}_{\gamma}^{\mathrm{T}} \nabla_{\gamma} \log P_{\gamma_0}(h_0, h_0') + \psi(\mathbf{X}) - \int_{\mathbf{x}} \psi(\mathbf{x}) dF_0(\mathbf{x}) = 0$. This implies that $\psi(\mathbf{x})$ is constant, so it must be zero. Thus, (2.15) reduces to

$$\boldsymbol{\mu}_{\gamma}^{\mathrm{T}} \nabla_{\gamma} \log P_{\gamma_0}(H) + \boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}} \mathcal{D}(\mathbf{X}, H) - \frac{\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}} \int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}} \mathcal{D}(\mathbf{x}, H)\} \mathcal{D}(\mathbf{x}, H) dF_0(\mathbf{x})}{\int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}} \mathcal{D}(\mathbf{x}, H)\} dF_0(\mathbf{x})} = 0.$$

By Condition 2.3, $\boldsymbol{\mu}_{\boldsymbol{\zeta}} = \mathbf{0}$, so (2.15) further reduces to $\boldsymbol{\mu}_{\gamma}^{\mathrm{T}} \nabla_{\gamma} \log P_{\gamma_0}(H) = 0$. By Lemma 1 of Lin and Zeng (2006), $\boldsymbol{\mu}_{\gamma} = \mathbf{0}$. Setting $Y = 1$ and $G = 2h$ or $G = h + h^{\dagger}$, we see that $\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}, H)$ must be a constant. By Condition 2.4, $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$.

We provide a mathematical definition of rare disease in Condition 2.5 and state the asymptotic results in Theorem 2.2.

CONDITION 2.5 $\Pr(Y_i = 1|\mathbf{X}_i, H_i) = a_n \exp\{\boldsymbol{\beta}_0^{\mathrm{T}} \mathcal{Z}(\mathbf{X}_i, H_i)\}/[1+a_n \exp\{\boldsymbol{\beta}_0^{\mathrm{T}} \mathcal{Z}(\mathbf{X}_i, H_i)\}]$, $i = 1, \ldots, n$, where $a_n = o(n^{-1/2})$.

THEOREM 2.2 Assume that Conditions 2.3-2.5 hold and $n_1/n \to q \in (0, 1)$. Then $|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sup_{\mathbf{x}} |\widehat{F}(\mathbf{x}) - F_0(\mathbf{x})| \to 0$ almost surely, and $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

Proof: Let $\widetilde{P}_n$ be the probability measure generated by the likelihood function given in (2.8) and let P_{n0} be the true likelihood function. Since $a_n = o(n^{-1/2})$, we have $\log \widetilde{P}_n/P_{n0} \to_{\widetilde{P}_n \text{ or } P_{n0}} 1$. By LeCam's lemma, $\widetilde{P}_n$ and P_{n0} are equivalent. Thus, the asymptotic properties under the true likelihood is equivalent to those under the the approximate likelihood given in (2.8). In other words, we can assume that data are generated from (2.8). Hence, the conclusion of the theorem follows from the arguments in the proof of Theorem 2.1.

2.6.3 Case-Control Studies with Known Disease Rate

Numerical Algorithm

The E-step is similar to that of Section 2.6.1. In the M-step, we use the Lagrange multiplier λ for the constraint

$$\sum_{j,k,l} P_{\alpha,\boldsymbol{\beta}}(Y = 1|\mathbf{x}_j, h_k, h_l) \frac{\exp(\boldsymbol{\nu}^{\mathrm{T}} \mathcal{W}_{kl} + \boldsymbol{\delta}^{\mathrm{T}} \mathcal{M}_{jkl})}{\sum_{j'} \exp(\boldsymbol{\delta}^{\mathrm{T}} \mathcal{M}_{j'kl})} = p_1 \sum_{k,l} \exp(\boldsymbol{\nu}^{\mathrm{T}} \mathcal{W}_{kl}).$$

The objective function to be maximized in the M-step is

$$\widetilde{l}_n(\alpha, \boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{\delta}, \lambda) = \sum_{i,j,k,l} \omega_{ijkl} \Big\{ \log P_{\alpha,\boldsymbol{\beta}}(Y_i|\mathbf{x}_j, h_k, h_l) + \boldsymbol{\nu}^{\mathrm{T}} \mathcal{W}_{kl} + \boldsymbol{\delta}^{\mathrm{T}} \mathcal{M}_{jkl} - \log\Big(\sum_{j'} e^{\boldsymbol{\delta}^{\mathrm{T}} \mathcal{M}_{j'kl}}\Big) \Big\}$$

$$-\lambda \Big\{ \sum_{j,k,l} P_{\alpha,\boldsymbol{\beta}}(Y = 1|\mathbf{x}_j, h_k, h_l) e^{\boldsymbol{\nu}^{\mathrm{T}} \mathcal{W}_{kl} + \boldsymbol{\delta}^{\mathrm{T}} \mathcal{M}_{jkl}} / \sum_{j'} e^{\boldsymbol{\delta}^{\mathrm{T}} \mathcal{M}_{j'kl}} - p_1 \sum_{k,l} e^{\boldsymbol{\nu}^{\mathrm{T}} \mathcal{W}_{kl}} \Big\}$$

$$-n\log\Big\{\sum_{k,l} e^{\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl}}\Big\}.$$

We can treat λ as a free parameter in $\widetilde{l}_n(\alpha,\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\delta},\lambda)$, so that the constraint is automatically met by setting the derivative with respect to λ to zero. The maximization can be carried out by the quasi-Newton method. The variances and covariances can be estimated by the inverse information matrix or by the profile-likelihood method.

Theoretical Results

LEMMA 2.5 If two sets of parameters $(\boldsymbol{\theta},F)$ and $(\widetilde{\boldsymbol{\theta}},\widetilde{F})$ yield the same joint distribution, then $\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}$ and $F=\widetilde{F}$.

Proof: Suppose that

$$\sum_{H\in\mathcal{S}(G)} P_{\alpha,\boldsymbol{\beta}}(Y|\mathbf{X},H)P_{\boldsymbol{\zeta},F}(\mathbf{X}|H)P_{\boldsymbol{\gamma}}(H)=\sum_{H\in\mathcal{S}(G)} P_{\widetilde{\alpha},\widetilde{\boldsymbol{\beta}}}(Y|\mathbf{X},H)P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H)P_{\widetilde{\boldsymbol{\gamma}}}(H).$$

Letting $G=2h$ or $G=h+h^{\dagger}$, we have

$$P_{\alpha,\boldsymbol{\beta}}(Y|\mathbf{X},H)P_{\boldsymbol{\zeta},F}(\mathbf{X}|H)P_{\boldsymbol{\gamma}}(H)=P_{\widetilde{\alpha},\widetilde{\boldsymbol{\beta}}}(Y|\mathbf{X},H)P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H)P_{\widetilde{\boldsymbol{\gamma}}}(H). \tag{2.16}$$

Set $Y=0$ or 1 in (2.16). The summation of the two resulting equations yields

$$P_{\boldsymbol{\zeta},F}(\mathbf{X}|H)P_{\boldsymbol{\gamma}}(H)=P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H)P_{\widetilde{\boldsymbol{\gamma}}}(H).$$

By the arguments in the proof of Lemma 2.3, $\boldsymbol{\gamma}=\widetilde{\boldsymbol{\gamma}}$, $f=\widetilde{f}$, and $\boldsymbol{\zeta}=\widetilde{\boldsymbol{\zeta}}$. Then (2.16) reduces to $\exp\big\{(\alpha-\widetilde{\alpha})+(\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}})^{\mathrm{T}}\mathcal{Z}(\mathbf{X},H)\big\}=1$. By Condition 2.4, $\alpha=\widetilde{\alpha}$ and $\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}$.

LEMMA 2.6 If there exist a vector $\boldsymbol{\mu}_{\boldsymbol{\theta}}\equiv(\mu_\alpha,\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}},\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}},\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}})^{\mathrm{T}}$ and a function ψ with $E[\psi(\mathbf{X})]=0$ such that

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}}l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0,F_0)+l_F(\boldsymbol{\theta}_0,F_0)[\int\psi\;dF_0]=0,$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$, and $l_F[\int \psi \, dF_0]$ is the score function for F along the submodel $F_0 + \epsilon \int \psi dF_0$ that satisfies the constraint $\Pr(Y = 1) = p_1$, then $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{0}$ and $\psi = 0$.

Proof: We wish to show that if there exist a vector $\boldsymbol{\mu}_{\boldsymbol{\theta}} \equiv (\mu_\alpha, \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}, \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}}, \boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}})^{\mathrm{T}}$ and functions ψ with $E[\psi(\mathbf{X})] = 0$ such that

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, F_0) + l_F(\boldsymbol{\theta}_0, F_0)[\int \psi \, dF_0] = 0, \tag{2.17}$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$, and $l_F[\int \psi \, dF_0]$ is the score function for F along the submodel $F_0 + \epsilon \int \psi dF_0$ that satisfies the constraint $\Pr(Y = 1) = p$, then $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{0}$ and $\psi = 0$. With $G = 2h$ or $G = h + h^{\dagger}$, (2.17) becomes

$$(\mu_\alpha + \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}, H)) \left[Y - \frac{\exp\{\alpha_0 + \boldsymbol{\beta}_0^{\mathrm{T}} \mathcal{Z}(\mathbf{X}, H)\}}{1 + \exp\{\alpha_0 + \boldsymbol{\beta}_0^{\mathrm{T}} \mathcal{Z}(\mathbf{X}, H)\}} \right] + \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}} \nabla_{\boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}_0}(H) + \boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}} \mathcal{D}(\mathbf{X}, H)$$
$$- \frac{\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}} \int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}} \mathcal{D}(\mathbf{x}, H)\} \mathcal{D}(\mathbf{x}, H) dF_0(\mathbf{x})}{\int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}} \mathcal{D}(\mathbf{x}, H)\} dF_0(\mathbf{x})} + \psi(\mathbf{X}) - \frac{\int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}} \mathcal{D}(\mathbf{x}, H)\} \psi(\mathbf{x}) dF_0(\mathbf{x})}{\int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}} \mathcal{D}(\mathbf{x}, H)\} dF_0(\mathbf{x})} = 0.$$

The difference of the two equations under $Y = 1$ and $Y = 0$ yields $\mu_\alpha + \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}, H) = 0$. By Condition 2.4, $\mu_\alpha = 0$ and $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$. It then follows from the arguments in the proof of Lemma 2.4 that $\boldsymbol{\mu}_{\boldsymbol{\zeta}} = \mathbf{0}$, $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \mathbf{0}$, and $\psi = 0$.

THEOREM 2.3 Under Conditions 2.3-2.4, the results of Theorem 2.2 hold.

Proof: First, we prove the consistency. Since $\widehat{\boldsymbol{\theta}}$ is bounded and $\widehat{F}$ is a distribution function, for any subsequence of $(\widehat{\boldsymbol{\theta}}, \widehat{F})$, there exists a further subsequence, still denoted as $(\widehat{\boldsymbol{\theta}}, \widehat{F})$, such that $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$, and $\widehat{F}$ weakly converge to F^*. The consistency will hold if we can show that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $F^* = F_0$. We abbreviate $\eta(\mathbf{x}, \mathbf{x}_0, (h, h'), (h_0, h_0'))$ and $P_{\alpha,\boldsymbol{\beta}}(Y|\mathbf{x}, H) P_{\boldsymbol{\gamma}}(H)$ as $\eta(\mathbf{x}, H)$ and $q(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x}, H, Y)$, respectively. After differentiating the log-likelihood function with respect to the jump sizes of F, we see that there exist some Lagrange multipliers $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ such that, for $k = 1, \ldots, n$,

$$\frac{1}{\widehat{F}\{\mathbf{X}_k\}} - \sum_{i=1}^{n} \frac{\sum_{H \in \mathcal{S}(G_i)} q(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \mathbf{X}_i, H, Y_i)\eta(\mathbf{X}_i, H)\eta(\mathbf{X}_k, H)/\{\int_{\mathbf{x}} \eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x})\}^2}{\sum_{H \in \mathcal{S}(G_i)} q(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \mathbf{X}_i, H, Y_i)\eta(\mathbf{X}_i, H)/\int_{\mathbf{x}} \eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x})}$$
$$-\widehat{\lambda}_2 \sum_{H} \left[\frac{q(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \mathbf{X}_k, H, 1)\eta(\mathbf{X}_k, H)}{\int_{\mathbf{x}} \eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x})} - \frac{\eta(\mathbf{X}_k, H) \int_{\mathbf{x}} q(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \mathbf{x}, H, 1)\eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x})}{\{\int_{\mathbf{x}} \eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x}))\}^2} \right] - \widehat{\lambda}_1 = 0.$$

In addition, $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ satisfy the constraint equations

$$\sum_{k=1}^{n} \widehat{F}\{\mathbf{X}_k\} = 1,$$

$$\sum_{k=1}^{n} \sum_{H} q(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \mathbf{X}_k, H, 1) \frac{\eta(\mathbf{X}_k, H)}{\int_{\mathbf{x}} \eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x})} \widehat{F}\{\mathbf{X}_k\} = p_1.$$

It follows that $\widehat{\lambda}_1 = 0$. Thus,

$$\left\{ \sum_{i=1}^{n} \frac{\sum_{H \in \mathcal{S}(G_i)} q(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \mathbf{X}_i, H, Y_i)\eta(\mathbf{X}_i, H)\eta(\mathbf{X}_k, H)/\{\int_{\mathbf{x}} \eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x})\}^2}{\sum_{H \in \mathcal{S}(G_i)} q(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \mathbf{X}_i, H, Y_i)\eta(\mathbf{X}_i, H)/\int_{\mathbf{x}} \eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x})} \right.$$
$$\left. +\widehat{\lambda}_2 \sum_{H} \left[\frac{q(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \mathbf{X}_k, H, 1)\eta(\mathbf{X}_k, H)}{\int_{\mathbf{x}} \eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x})} - \frac{\eta(\mathbf{X}_k, H) \int_{\mathbf{x}} q(\widehat{\alpha}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \mathbf{x}, H, 1)\eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x})}{\{\int_{\mathbf{x}} \eta(\mathbf{x}, H) d\widehat{F}(\mathbf{x}))\}^2} \right] \right\}^{-1} = 1,$$

and each denominator on the left-hand side should be positive. This equation for $\widehat{\lambda}_2$ has a unique solution satisfying the above constraints. In addition, we can show that $\widehat{\lambda}_2/n$ is bounded with probability one. Thus, we can choose a further subsequence such that $\widehat{\lambda}_2/n \to \lambda_2^*$.

We construct a discrete distribution function $\widetilde{F}$ such that $\widetilde{F} \to F_0$ uniformly. The sequence can be constructed along the lines of Lin and Zeng (2006, §A.4.6). Although $\widetilde{F}$ is a distribution function, it may not satisfy the constraint that

$$\int_{\mathbf{x}} \sum_{H} P_{\alpha_0, \boldsymbol{\beta}_0}(Y = 1|\mathbf{x}, H) P_{\boldsymbol{\gamma}}(H) P_{\boldsymbol{\zeta}_0, F}(\mathbf{x}|H) f(\mathbf{x}) d\mathbf{x} = p_1.$$

Thus, we modify the jump size of $\widetilde{F}$ at $\mathbf{X}_k$ as $[\widetilde{F}\{\mathbf{X}_k\} + \xi/n]/(1+\xi)$ for some constant ξ such that ξ satisfies the above constraint. It can be shown that the solution exists and $\xi \to 0$. The modified distribution function $\widetilde{F}$ then satisfies all the constraints. By the Glivenko-Cantelli theorem, $\widehat{F}$ is absolutely continuous with respect to $\widetilde{F}$, and $d\widehat{F}/d\widetilde{F}(\mathbf{x}) \to q(\mathbf{x})$ uniformly in $\mathbf{x}$ for some positive function $q(\cdot)$. Since $n^{-1}\log\{L_n(\widehat{\boldsymbol{\theta}}, \widehat{F})/L_n(\boldsymbol{\theta}_0, \widetilde{F})\} \geq 0$, we take limits. We conclude that the Kullback-Leibler information for $(\boldsymbol{\theta}^*, F^*)$ is non-positive. Hence, Lemma 2.5 entails that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $F^* = F_0$.

We now derive the asymptotic distribution. We obtain score functions by differentiating $\log L_n(\boldsymbol{\theta}, F)$ with respect to $\widehat{\boldsymbol{\theta}}$ along the direction $\mathbf{v}$ and with respect to $\widehat{F}$ along submodels with tangent direction ψ satisfying all the constraints and with the total variation bounded by 1. The linearization of the score functions around the true parameter value, together with the Donsker theorem, yields

$$n^{1/2}\left[\boldsymbol{\Omega}_1(\mathbf{v},\psi)^{\mathrm{T}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \int \Omega_2(\mathbf{v},\psi)d(\widehat{F} - F_0)\right] = n^{-1/2}\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}} l_{\boldsymbol{\theta}} + l_F[\psi]\right) + o_p(1),$$

where $\boldsymbol{\Omega} \equiv (\boldsymbol{\Omega}_1, \Omega_2)$ corresponds to the information operator and has the form of the first-order Fredholm type, and $l_{\boldsymbol{\theta}}$ and l_F are the score operators for $\boldsymbol{\theta}$ and F, respectively. According to Lemma 2.6, $\boldsymbol{\Omega}$ is invertible. Thus, the weak convergence follows from Theorem 3.3.1 of van der Vaart and Wellner (1996). In addition, $\widehat{\boldsymbol{\theta}}$ is an asymptotically linear estimator for $\boldsymbol{\theta}_0$ with the influence function in the score space, so it follows from Proposition 3.3.1 of Bickel et al. (1993) that the limiting covariance matrix of $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ attains the semiparametric efficiency bound.

2.6.4 Cohort Studies

Numerical Algorithm

We present the EM algorithm for the proportional hazards model. Suppose that there are L distinct failure times $t_1,\ldots,t_L$. Let $\Lambda\{t_l\}$ denote the jump size of Λ at t_l, and d_l the number of failures at t_l. In the E-step, we evaluate the conditional expectations

$$\begin{aligned}\omega_{ijkl} \equiv & E\{I(\mathbf{X}_i=\mathbf{x}_j, H_i=(h_k,h_l))\big|\widetilde{Y}_i,\Delta_i,\mathbf{X}_i,G_i\}\\ = & \frac{I(\mathbf{X}_i=\mathbf{x}_j,(h_k,h_l)\in\mathcal{S}(G_i))R_{ijkl}(\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\delta})/\sum_{j'}\exp(\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{j'kl})}{\sum_{(h_{k'},h_{l'})\in\mathcal{S}(G_i)}R_{ijk'l'}(\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\delta})/\sum_{j'}\exp(\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{j'k'l'})},\end{aligned}$$

where $R_{ijkl}(\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\delta})=\exp(\Delta_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}_{jkl}+\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl}+\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{jkl}-e^{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}_{jkl}}\sum_{m:t_m\le\widetilde{Y}_i}\Lambda\{t_m\})$. In the M-step, we maximize

$$\begin{aligned}\widetilde{l}_n(\boldsymbol{\beta},\boldsymbol{\nu},\boldsymbol{\delta},\Lambda)=&\sum_{i,j,k,l}\omega_{ijkl}\Delta_i\log\Lambda\{\widetilde{Y}_i\}+\sum_{i,j,k,l}\omega_{ijkl}\Bigg(\Delta_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}_{jkl}+\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl}+\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{jkl}\\ &-\log\Bigg\{\sum_{j'}\exp(\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{j'kl})\Bigg\}-e^{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}_{jkl}}\sum_{m:t_m\le\widetilde{Y}_i}\Lambda\{t_m\}\Bigg)-n\log\Bigg\{\sum_{k,l}\exp(\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl})\Bigg\}.\end{aligned}$$

The estimate for $\Lambda\{t_m\}$ is given explicitly by $d_m/\sum_{i:\widetilde{Y}_i\ge t_m}\sum_{j,k,l}\omega_{ijkl}e^{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}_{jkl}}$, and the estimate for $\boldsymbol{\beta}$ solves the equation

$$\sum_{i,j,k,l}\omega_{ijkl}\Delta_i\mathcal{Z}_{jkl}-\sum_{m=1}^{L}d_m\frac{\sum_{i:\widetilde{Y}_i\ge t_m}\sum_{j,k,l}\omega_{ijkl}\mathcal{Z}_{jkl}e^{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}_{jkl}}}{\sum_{i:\widetilde{Y}_i\ge t_m}\sum_{j,k,l}\omega_{ijkl}e^{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}_{jkl}}}=0.$$

The remaining parameters can be estimated by maximizing

$$\sum_{i,j,k,l}\omega_{ijkl}\Bigg[\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl}+\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{jkl}-\log\Bigg\{\sum_{j'}\exp(\boldsymbol{\delta}^{\mathrm{T}}\mathcal{M}_{j'kl})\Bigg\}\Bigg]-n\log\Bigg\{\sum_{k,l}\exp(\boldsymbol{\nu}^{\mathrm{T}}\mathcal{W}_{kl})\Bigg\}.$$

We can estimate the asymptotic variances and covariances by the inverse information matrix or the profile-likelihood method. For other transformation models, we may use the Laplace transformation to convert the estimation problem into that of the proportional hazards model with a random effect; see Zeng and Lin (2007).

Theoretical Results

We impose the following conditions:

CONDITION 2.6 There exists a positive constant δ_0 such that $\Pr(C \geq \tau|\mathbf{X}, G) = \Pr(C = \tau|\mathbf{X}, G) \geq \delta_0$ almost surely, where τ corresponds to the end of the study.

CONDITION 2.7 The true value $\Lambda_0(t)$ of $\Lambda(t)$ is a strictly increasing function in $[0, \tau]$ and is continuously differentiable. In addition, $\Lambda_0(0) = 0$ and $\Lambda_0'(0) > 0$.

LEMMA 2.7 If two sets of parameters $(\boldsymbol{\theta}, F, \Lambda)$ and $(\widetilde{\boldsymbol{\theta}}, \widetilde{F}, \widetilde{\Lambda})$ yield the same joint distribution, then $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$, $F = \widetilde{F}$ and $\Lambda = \widetilde{\Lambda}$.

Proof: Suppose that

$$\begin{aligned}
&\sum_{H \in \mathcal{S}(G)} \left[\Lambda'(\widetilde{Y}) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X},H)} Q'(\Lambda(\widetilde{Y}) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X},H)})\right]^{\Delta} \exp\Big\{-Q(\Lambda(\widetilde{Y}) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X},H)})\Big\} P_{\boldsymbol{\zeta},F}(\mathbf{X}|H) P_{\boldsymbol{\gamma}}(H) \\
= &\sum_{H \in \mathcal{S}(G)} \left[\widetilde{\Lambda}'(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}} \mathcal{Z}(\mathbf{X},H)} Q'(\widetilde{\Lambda}(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}} \mathcal{Z}(\mathbf{X},H)})\right]^{\Delta} \exp\Big\{-Q(\widetilde{\Lambda}(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}} \mathcal{Z}(\mathbf{X},H)})\Big\} P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H) P_{\widetilde{\boldsymbol{\gamma}}}(H).
\end{aligned}$$

We choose $\Delta = 1$ and integrate $\widetilde{Y}$ from 0 to y on both sides to obtain the equation

$$\begin{aligned}
&\sum_{H \in \mathcal{S}(G)} \left[1 - \exp\{-Q(\Lambda(y) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X},H)})\}\right] P_{\boldsymbol{\zeta},F}(\mathbf{X}|H) P_{\boldsymbol{\gamma}}(H) \\
= &\sum_{H \in \mathcal{S}(G)} \left[1 - \exp\{-Q(\widetilde{\Lambda}(y) e^{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}} \mathcal{Z}(\mathbf{X},H)})\}\right] P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H) P_{\widetilde{\boldsymbol{\gamma}}}(H). \qquad (2.18)
\end{aligned}$$

We obtain a second equation by setting $\Delta = 0$ and $\widetilde{Y} = y$. The summation of the two equations yields

$$\sum_{H \in \mathcal{S}(G)} P_{\boldsymbol{\zeta},F}(\mathbf{X}|H) P_{\boldsymbol{\gamma}}(H) = \sum_{H \in \mathcal{S}(G)} P_{\widetilde{\boldsymbol{\zeta}},\widetilde{F}}(\mathbf{X}|H) P_{\widetilde{\boldsymbol{\gamma}}}(H).$$

By the arguments in the proof of Lemma 2.1, $\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}}$, $f = \widetilde{f}$ and $\boldsymbol{\zeta} = \widetilde{\boldsymbol{\zeta}}$. By letting $G = 2h$ or $G = h + \widetilde{h}$ in (2.18), we have $\Lambda(y)e^{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X},H)} = \widetilde{\Lambda}(y)e^{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(\mathbf{X},H)}$, which entails $\Lambda = \widetilde{\Lambda}$ and $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$ under Condition 2.4.

LEMMA 2.8 If there exist a vector $\boldsymbol{\mu}_{\boldsymbol{\theta}} \equiv (\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}, \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}}, \boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}})^{\mathrm{T}}$ and functions $\psi(\mathbf{x})$ and $\phi(t)$ with $E[\psi(\mathbf{X})] = E[\phi(Y)] = 0$ such that

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, F_0, \Lambda_0) + l_F(\boldsymbol{\theta}_0, F_0, \Lambda_0)[\int \psi \, dF_0] + l_\Lambda(\boldsymbol{\theta}_0, F_0, \Lambda_0)[\int \phi \, d\Lambda_0] = 0,$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$, $l_F[\int \psi \, dF_0]$ is the score function for F along the sub-model $F_0 + \epsilon \int \psi \, dF_0$, and $l_\Lambda[\int \phi \, d\Lambda_0]$ is the score function for Λ along the sub-model $\Lambda_0 + \epsilon \int \phi \, d\Lambda_0$, then $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{0}$, $\psi = 0$ and $\phi = 0$.

Proof: We wish to show that if there exist a vector $\boldsymbol{\mu}_{\boldsymbol{\theta}} \equiv (\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}, \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}}, \boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}})^{\mathrm{T}}$ and functions $\psi(\mathbf{x})$ and $\phi(t)$ with $E[\psi(\mathbf{X})] = E[\phi(Y)] = 0$ such that

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, F_0, \Lambda_0) + l_F(\boldsymbol{\theta}_0, F_0, \Lambda_0)[\int \psi \, dF_0] + l_\Lambda(\boldsymbol{\theta}_0, F_0, \Lambda_0)[\int \phi \, d\Lambda_0] = 0, \qquad (2.19)$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$, $l_F[\int \psi \, dF_0]$ is the score function for F along the sub-model $F_0 + \epsilon \int \psi \, dF_0$, and $l_\Lambda[\int \phi \, d\Lambda_0]$ is the score function for Λ along the sub-model $\Lambda_0 + \epsilon \int \phi \, d\Lambda_0$, then $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{0}$, $\psi = 0$ and $\phi = 0$. With $\Delta = 1$, (2.19) becomes

$$\sum_{H \in \mathcal{S}(G)} \Lambda_0'(\widetilde{Y}) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}} Q'(\Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}}) \exp\Big\{-Q(\Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}})\Big\} P_{\boldsymbol{\gamma}_0}(H) \frac{\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{X}, H)\} f_0(\mathbf{X})}{\int_{\mathbf{x}} \exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x}, H)\} dF_0(\mathbf{x})}$$

$$\times \Bigg\{ \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z} + \frac{\Big[Q''(\Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}}) - \big(Q'(\Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}})\big)^2\Big] \Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}} \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}}{Q'(\Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}})}$$

$$+\phi(\widetilde{Y}) + \frac{\Big[Q''(\Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}}) - \big(Q'(\Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}})\big)^2\Big] \int_0^{\widetilde{Y}} \phi(t) d\Lambda_0(t) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}}}{Q'(\Lambda_0(\widetilde{Y}) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}})} + \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}} \nabla_{\boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}_0}(H)$$

$$+\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{X},H)-\frac{\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}}\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}\mathcal{D}(\mathbf{x},H)dF_0(\mathbf{x})}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})}$$

$$\left.+\psi(\mathbf{X})-\frac{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}\psi(\mathbf{x})dF_0(\mathbf{x})}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})}\right\}=0. \tag{2.20}$$

In the above equation, we integrate $\widetilde{Y}$ from 0 to τ. We also let $\Delta=0$ and $\widetilde{Y}=\tau$ in (2.19). The summation of these two equations with $G=2h$ or $G=h+h^{\dagger}$ yields

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\mathrm{T}}\nabla_{\boldsymbol{\gamma}}\log P_{\boldsymbol{\gamma}_0}(H)+\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}}\mathcal{D}(\mathbf{X},H)-\frac{\boldsymbol{\mu}_{\boldsymbol{\zeta}}^{\mathrm{T}}\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}\mathcal{D}(\mathbf{x},H)dF_0(\mathbf{x})}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})}$$
$$+\psi(\mathbf{X})-\frac{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}\psi(\mathbf{x})dF_0(\mathbf{x})}{\int_{\mathbf{x}}\exp\{\boldsymbol{\zeta}_0^{\mathrm{T}}\mathcal{D}(\mathbf{x},H)\}dF_0(\mathbf{x})}=0.$$

It follows from the arguments in the proof of Lemma 2.2 that $\boldsymbol{\mu}_{\boldsymbol{\gamma}}=\mathbf{0}$, $\boldsymbol{\mu}_{\boldsymbol{\zeta}}=\mathbf{0}$, and $\psi=0$. By letting $G=2h$ or $G=h+h^{\dagger}$ and $Y=0$ in (2.20), we obtain $\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(\mathbf{X},H)+\phi(0)=0$. It then follows from Condition 2.4 that $\boldsymbol{\mu}_{\boldsymbol{\beta}}=\mathbf{0}$ and $\phi(0)=0$. Thus, (2.20) reduces to

$$\phi(\widetilde{Y})+\frac{\left[Q''(\Lambda_0(\widetilde{Y})e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}})-\left(Q'(\Lambda_0(\widetilde{Y})e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}})\right)^2\right]\int_0^{\widetilde{Y}}\phi(t)d\Lambda_0(t)e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}}}{Q'(\Lambda_0(\widetilde{Y})e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}})}=0$$

for $H=(h,h)$. Since Q is strictly increasing, we conclude that $\phi(y)=0$ for any y.

THEOREM 2.4 Under the conditions of Theorem 2.3 and Conditions 2.6-2.7, $|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0|+\sup_{\mathbf{x}}|\widehat{F}(\mathbf{x})-F_0(\mathbf{x})|+\sup_{t\in[0,\tau]}|\widehat{\Lambda}(t)-\Lambda_0(t)|\rightarrow 0$ almost surely. In addition, $n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0,\widehat{\Lambda}-\Lambda_0)$ converges weakly to a zero-mean Gaussian process in $R^d\times l^{\infty}([0,\tau])$, where d is the dimension of $\boldsymbol{\theta}_0$, and $l^{\infty}([0,\tau])$ is the space of all bounded functions on $[0,\tau]$ equipped with the supremum norm. Furthermore, the limiting covariance matrix of $\widehat{\boldsymbol{\theta}}$ attains the semiparametric efficiency bound.

Proof: First, we show that $\widehat{\Lambda}$ is uniformly bounded in $[0,\tau]$ as $n \to \infty$. Note that $\widehat{\Lambda}$ maximizes $\widetilde{L}_n(\Lambda) \equiv L_n(\widehat{\boldsymbol{\theta}}, \Lambda, \widehat{F}) / \prod_{i=1}^n \widehat{F}\{\mathbf{X}_i\}$. Clearly,

$$\widetilde{L}_n(\Lambda) \le c_0 \prod_{i=1}^n \sum_{H \in \mathcal{S}(G_i)} \left\{ \Lambda'(\widetilde{Y}_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}_i, H)} Q'\big(-\Lambda(\widetilde{Y}_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}_i, H_i)}\big) \right\}^{\Delta_i} \exp\left\{ -Q\big(\Lambda(\widetilde{Y}_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathbf{X}_i, H_i)}\big) \right\}$$

for some constant c_0. According to the conditions of this theorem and Appendix B of Zeng and Lin (2007), $\widetilde{L}_n(\Lambda) \le c_1 \prod_{i=1}^n \left[\Lambda'(\widetilde{Y}_i)^{\Delta_i} (1 + \Lambda(\widetilde{Y}_i))^{-(\Delta_i + \delta_0)} \right]$ for some positive constants c_1 and δ_0. By the partitioning arguments in the proof of Theorem 1 of Zeng and Lin (2007), we can show that if $\widehat{\Lambda}(\tau)$ is unbounded, then the difference between $L_n(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}, \widehat{F})$ and $L_n(\boldsymbol{\theta}_0, \widetilde{\Lambda}, \widehat{F})$, where $\widetilde{\Lambda}$ is a step function converging to Λ_0, diverges to $-\infty$. Thus, $\widehat{\Lambda}(\tau)$ must be bounded with probability one.

Using the above result and the arguments in the proof of Theorem 2.3, we choose a uniformly convergent subsequence from any subsequence of $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}, \widehat{F})$. By the Glivenko-Cantelli theorem and the property of the Kullback-Leibler information, the limit of the convergent subsequence must be the true parameters $(\boldsymbol{\theta}_0, \Lambda_0, F_0)$. The asymptotic distribution of $\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}$ and $\widehat{F}$ follows from the arguments used in the proof of Theorem 2.3.

2.6.5 More Numerical Results

We conducted simulation studies in the set-up of Chen et al. (2009). Specifically, we generated haplotypes under HWE from the distribution given in Table 1 of Chen et al. (2009) and generated a binary environmental covariate X with $\Pr(X = 1) = 0.3$, $\zeta_{1,3} = 0$ or $-.4$ and $\zeta_{1,j} = 0$ $(j \neq 3)$. Given H and X, the disease status was generated from model (13) of Chen et al. (2009).

For each simulated data set, we calculated the proposed estimator of $\boldsymbol{\beta}$ allowing for gene-environment dependence and the Lin-Zeng estimator assuming gene-environment independence, denoted as $\widehat{\boldsymbol{\beta}}_{\text{dep}}$ and $\widehat{\boldsymbol{\beta}}_{\text{ind}}$, respectively. Given these two estimators, we constructed two empirical Bayes estimators using formula (7) of Chen et al. (2009). Specifically, the multivariate shrinkage estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\text{EB1}} = \widehat{\boldsymbol{\beta}}_{\text{dep}} + \mathbf{K}(\widehat{\boldsymbol{\beta}}_{\text{ind}} - \widehat{\boldsymbol{\beta}}_{\text{dep}}),$$

where $\mathbf{K} = \mathbf{V}\big[\mathbf{V} + (\widehat{\boldsymbol{\beta}}_{\text{ind}} - \widehat{\boldsymbol{\beta}}_{\text{dep}})(\widehat{\boldsymbol{\beta}}_{\text{ind}} - \widehat{\boldsymbol{\beta}}_{\text{dep}})^{\text{T}}\big]^{-\mathbf{1}}$, and $\mathbf{V}$ is the estimated covariance matrix of $(\widehat{\boldsymbol{\beta}}_{\text{ind}} - \widehat{\boldsymbol{\beta}}_{\text{dep}})$; the component-wise shrinkage estimator of the jth component of $\boldsymbol{\beta}$ is

$$\widehat{\beta}_{\text{EB2},j} = \widehat{\beta}_{\text{dep},j} + k_j(\widehat{\beta}_{\text{ind},j} - \widehat{\beta}_{\text{dep},j}),$$

where $\widehat{\beta}_{\text{ind},j}$ and $\widehat{\beta}_{\text{dep},j}$ are the jth components of $\widehat{\boldsymbol{\beta}}_{\text{ind}}$ and $\widehat{\boldsymbol{\beta}}_{\text{dep}}$, $k_j = v_j/\big[v_j + (\widehat{\beta}_{\text{ind},j} - \widehat{\beta}_{\text{dep},j})^2\big]$, and v_j is the jth diagonal element of $\mathbf{V}$.

Write $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\text{T}}, \boldsymbol{\chi}^{\text{T}})^{\text{T}}$, where $\boldsymbol{\chi}$ denotes all nuisance parameters (including finite-dimensional nuisance parameters and jump sizes of nuisance functions). Also, let $\boldsymbol{\theta}^*_{\text{ind}}$ and $\boldsymbol{\theta}^*_{\text{dep}}$ be the probability limits of $\widehat{\boldsymbol{\theta}}_{\text{ind}}$ and $\widehat{\boldsymbol{\theta}}_{\text{dep}}$. We note the following representations

$$\widehat{\boldsymbol{\beta}}_{\text{ind}} - \boldsymbol{\beta}^*_{\text{ind}} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \end{pmatrix} \mathcal{I}^{-1}_{\text{ind}}(\boldsymbol{\theta}^*_{\text{ind}}) \sum_{i=1}^{n} U_{\text{ind},i}(\boldsymbol{\theta}^*_{\text{ind}}) + o_p(n^{-1/2}),$$

and

$$\widehat{\boldsymbol{\beta}}_{\text{dep}} - \boldsymbol{\beta}^*_{\text{dep}} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \end{pmatrix} \mathcal{I}^{-1}_{\text{dep}}(\boldsymbol{\theta}^*_{\text{dep}}) \sum_{i=1}^{n} U_{\text{dep},i}(\boldsymbol{\theta}^*_{\text{dep}}) + o_p(n^{-1/2}),$$

where $U_{\text{ind},i}(\boldsymbol{\theta})$ and $U_{\text{dep},i}(\boldsymbol{\theta})$ are the ith subject's contributions to the score functions of $\boldsymbol{\theta}$ under the Lin-Zeng and proposed methods, respectively, $\mathcal{I}_{\text{ind}}(\boldsymbol{\theta})$ and $\mathcal{I}_{\text{dep}}(\boldsymbol{\theta})$ are the corresponding information matrices, $\mathbf{I}_p$ is the $p \times p$ identity matrix, and $\mathbf{0}$ is the $p \times q$ zero matrix, with p and q being the dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\chi}$, respectively. Thus, we estimate the covariance matrices for $\widehat{\boldsymbol{\beta}}_{\text{ind}}$ and $\widehat{\boldsymbol{\beta}}_{\text{dep}}$ as follows:

$$\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}_{\text{ind}}) \equiv \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \end{pmatrix} \mathcal{I}_{\text{ind}}^{-1}(\widehat{\boldsymbol{\theta}}_{\text{ind}}) \Big\{ \sum_{i=1}^{n} U_{\text{ind},i}(\widehat{\boldsymbol{\theta}}_{\text{ind}}) U_{\text{ind},i}^{\text{T}}(\widehat{\boldsymbol{\theta}}_{\text{ind}}) \Big\} \mathcal{I}_{\text{ind}}^{-1}(\widehat{\boldsymbol{\theta}}_{\text{ind}}) \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \end{pmatrix}^{\text{T}},$$

$$\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}_{\text{dep}}) \equiv \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \end{pmatrix} \mathcal{I}_{\text{dep}}^{-1}(\widehat{\boldsymbol{\theta}}_{\text{dep}}) \Big\{ \sum_{i=1}^{n} U_{\text{dep},i}(\widehat{\boldsymbol{\theta}}_{\text{dep}}) U_{\text{dep},i}^{\text{T}}(\widehat{\boldsymbol{\theta}}_{\text{dep}}) \Big\} \mathcal{I}_{\text{dep}}^{-1}(\widehat{\boldsymbol{\theta}}_{\text{dep}}) \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \end{pmatrix}^{\text{T}},$$

$$\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}}_{\text{ind}}, \widehat{\boldsymbol{\beta}}_{\text{dep}}) \equiv \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \end{pmatrix} \mathcal{I}_{\text{ind}}^{-1}(\widehat{\boldsymbol{\theta}}_{\text{ind}}) \Big\{ \sum_{i=1}^{n} U_{\text{ind},i}(\widehat{\boldsymbol{\theta}}_{\text{ind}}) U_{\text{dep},i}^{\text{T}}(\widehat{\boldsymbol{\theta}}_{\text{dep}}) \Big\} \mathcal{I}_{\text{dep}}^{-1}(\widehat{\boldsymbol{\theta}}_{\text{dep}}) \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \end{pmatrix}^{\text{T}}.$$

The simulation results for the dominant and recessive models are presented in Table 2.6, in the same format as Tables 2 and 3 of Chen et al. (2009). Our results for the Lin-Zeng estimator (i.e., $\widehat{\boldsymbol{\beta}}_{\text{ind}}$) are similar to those of Chen et al.'s (2009) model-based estimator, especially under the recessive model. Under the dominant model, the proposed estimator (i.e., $\widehat{\boldsymbol{\beta}}_{\text{dep}}$) tends to be more efficient than Chen et al. (2009)'s model-free estimator, particularly in estimating gene-environment interactions. The efficiency gain is much more substantial under the recessive model, for both main effects and interactions. Consequently, our empirical Bayes estimators are more efficient than Chen et al.'s, especially under the recessive model.

Table 2.6: Simulation results of the mean square error (bias of the parameter estimator) for the empirical Bayes estimators under dominant and recessive models

		$n_1 = n_0 = 150$		$n_1 = n_0 = 300$		$n_1 = n_0 = 600$	
Dominant Model		H	$H \times X$	H	$H \times X$	H	$H \times X$
$\zeta_{1,3} = 0$	$\widehat{\beta}_{\text{dep}}$	.109(-.016)	.292(.024)	.054(-.008)	.141(.003)	.025(-.006)	.069(-.007)
	$\widehat{\beta}_{\text{ind}}$	.097(-.001)	.203(-.002)	.049(.004)	.095(-.022)	.023(.004)	.047(-.031)
	$\widehat{\beta}_{\text{EB1}}$	.106(-.018)	.274(.024)	.054(-.007)	.137(.001)	.025(-.005)	.067(-.008)
	$\widehat{\beta}_{\text{EB2}}$	.101(-.013)	.234(.016)	.052(-.003)	.118(-.007)	.024(-.002)	.059(-.016)
$\zeta_{1,3} = -.4$	$\widehat{\beta}_{\text{dep}}$	.111(-.008)	.310(.044)	.051(-.005)	.156(.005)	.026(-.005)	.074(-.013)
	$\widehat{\beta}_{\text{ind}}$	.120(.133)	.375(-.398)	.062(.128)	.275(-.416)	.038(.122)	.225(-.418)
	$\widehat{\beta}_{\text{EB1}}$	.107(-.004)	.293(.026)	.051(-.001)	.155(-.008)	.026(-.002)	.074(-.022)
	$\widehat{\beta}_{\text{EB2}}$	.107(.028)	.290(-.072)	.051(.021)	.163(-.079)	.026(.012)	.081(-.065)
Recessive Model							
$\zeta_{1,3} = 0$	$\widehat{\beta}_{\text{dep}}$	.099(-.048)	.261(-.049)	.049(-.027)	.127(-.031)	.029(-.023)	.073(-.023)
	$\widehat{\beta}_{\text{ind}}$	.095(-.057)	.197(-.043)	.047(-.030)	.092(-.034)	.026(-.023)	.052(-.031)
	$\widehat{\beta}_{\text{EB1}}$	.097(-.050)	.239(-.046)	.049(-.028)	.115(-.030)	.028(-.023)	.066(-.024)
	$\widehat{\beta}_{\text{EB2}}$	.096(-.054)	.225(-.047)	.048(-.030)	.108(-.031)	.027(-.024)	.062(-.026)
$\zeta_{1,3} = -.4$	$\widehat{\beta}_{\text{dep}}$	.087(-.050)	.339(-.065)	.044(-.026)	.173(-.031)	.026(-.022)	.088(-.032)
	$\widehat{\beta}_{\text{ind}}$	.095(.117)	.778(-.720)	.065(.147)	.621(-.699)	.047(.149)	.536(-.678)
	$\widehat{\beta}_{\text{EB1}}$	.087(-.039)	.352(-.107)	.044(-.020)	.177(-.053)	.026(-.018)	.090(-.044)
	$\widehat{\beta}_{\text{EB2}}$	.087(-.029)	.370(-.170)	.044(-.015)	.185(-.080)	.026(-.015)	.094(-.069)

NOTE: $\widehat{\beta}_{\text{dep}}$ and $\widehat{\beta}_{\text{ind}}$ pertain to the proposed estimator allowing for gene-environment dependence and the Lin-Zeng estimator assuming gene-environment independence, respectively. H and $H \times X$ stand for main haplotype effect and haplotype-environment interaction. Each entry is based on 1,000 replicates.

Chapter 3

Analysis of Untyped SNPs: Maximum Likelihood and Imputation Methods

3.1 Introduction

This chapter provides comprehensive comparisons of (single) imputation and maximum likelihood methods under cross-sectional and case-control designs. We expand the approach of Lin et al. (2008) to encompass both cross-sectional and case-control studies. In addition, we develop a tagging-based imputation strategy. We establish the theoretical properties of the proposed imputation method and conduct extensive simulation studies to evaluate the performance of the imputation and maximum-likelihood methods in testing/estimating genetic effects and gene-environment interactions. We apply the two methods to the GWAS data from the Wellcome Trust Case-Control Consortium (WTCCC) (Burton et al., 2007).

3.2 Methods

3.2.1 Imputation

Suppose that we are interested in a particular untyped SNP, whose genotype is denoted by G_u. Let Y denote the phenotype of interest, which can be quantitative or qualitative. Also, let $\mathbf{X}$ denote a set of environmental factors. We characterize the effects of genetic and environmental factors on the phenotype through the conditional density function $P_{\alpha,\boldsymbol{\beta},\boldsymbol{\xi}}(Y|G_u, \mathbf{W})$, where $\mathbf{W}$ consists of $\mathbf{X}$ and the genotypes of typed SNPs, and α, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ pertain to the intercept, regression parameters, and nuisance parameters (e.g., error variance), respectively. (If we are interested in the marginal effect of G_u, then $\mathbf{W}$ is an empty set.) We formulate $P_{\alpha,\boldsymbol{\beta},\boldsymbol{\xi}}(Y|G_u, \mathbf{W})$ through a generalized linear regression model with linear predictor $\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(G_u, \mathbf{W})$, where $\mathcal{Z}(G_u, \mathbf{W})$ is a vector-function of G_u and $\mathbf{W}$ under a particular mode of inheritance. We assume the additive mode of inheritance here, although all the formulas can be easily modified to accommodate other modes of inheritance. For a quantitative trait, we specify the linear regression model:

$$Y = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(G_u, \mathbf{W}) + \epsilon,$$

where ϵ is zero-mean normal with variance σ^2. For a binary trait, it is natural to use the logistic regression model:

$$\Pr(Y = 1|G_u, \mathbf{W}) = \frac{e^{\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(G_u,\mathbf{W})}}{1 + e^{\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(G_u,\mathbf{W})}}. \tag{3.1}$$

We use the LD information from a reference panel to select a set of $(M-1)$ typed SNPs that provides the most accurate prediction of the untyped SNP, where M is a small number, which is set to five here. The accuracy of prediction is measured by R^2 of Stram (2004). The M-locus genotype G consists of G_u and G_t, where G_t is the genotype of the $(M-1)$ typed SNPs. Suppose that the M SNPs have a total of K haplotypes. For $k=1,\ldots,K$, let h_k denote the kth haplotype, and π_k denote the frequency of h_k. Assume that the HWE holds. For a reference panel of $\widetilde{n}$ trios, the likelihood for $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_K)^{\mathrm{T}}$ is

$$L_R(\boldsymbol{\pi})=\prod_{j=1}^{\widetilde{n}}\sum_{(h_k,h_l,h_{k'},h_{l'})\sim G_j}\pi_k\pi_l\pi_{k'}\pi_{l'}, \tag{3.2}$$

where $G_j=(GF_j,GM_j,GC_j)$ is the genotype data for the jth trio with the M-locus genotypes GF_j, GM_j and GC_j for the father, mother and child, respectively, and $(h_k,h_l,h_{k'},h_{l'})\sim G_j$ means that (h_k,h_l) is compatible with GF_j, $(h_{k'},h_{l'})$ is compatible with GM_j, and $(h_k,h_{k'})$, $(h_k,h_{l'})$, $(h_l,h_{k'})$, or $(h_l,h_{l'})$ is compatible with GC_j.

By maximizing $L_R(\boldsymbol{\pi})$ given in equation (3.2) via the EM algorithm, we obtain the maximum-likelihood estimator $\widetilde{\boldsymbol{\pi}}=(\widetilde{\pi}_1,\ldots,\widetilde{\pi}_K)^{\mathrm{T}}$. Assuming that the haplotype frequencies are the same between the study population and the external panel, we can estimate the probability distribution of G_u from the observed values of G_t for each study subject according to the formula

$$\Pr(G_u=g|\ G_t;\widetilde{\boldsymbol{\pi}})=\frac{\sum_{(h_k,h_l)\sim(G_t,G_u=g)}\widetilde{\pi}_k\widetilde{\pi}_l}{\sum_{g'=0,1,2}\sum_{(h_k,h_l)\sim(G_t,G_u=g')}\widetilde{\pi}_k\widetilde{\pi}_l},\quad g=0,1,2, \tag{3.3}$$

where $(h_k,h_l)\sim(G_t,G_u=g)$ means that (h_k,h_l) is compatible with $(G_t,G_u=g)$. We use this (estimated) probability distribution to impute the unknown value of G_u, either as the expected count (i.e., dosage) or the most likely value of G_u. We replace the unknown values of G_u by the imputed values for all study subjects to create a "complete" data set, which is then analyzed by standard regression methods.

In the Appendix, we prove that the above imputation method yields a valid test of the null hypothesis $H_0:\beta_G=0$ under the linear predictor $\alpha+\beta_{G_u}G_u+\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}$, where β_{G_u} and $\boldsymbol{\beta}_{\mathbf{W}}$ pertain to the effects of G_u and $\mathbf{W}$, respectively, provided that G_t

is independent of Y conditional on $\mathbf{W}$. This result holds for both cross-sectional and case-control studies, even when the reference panel and the study sample are drawn from different underlying populations. However, the estimator of β_{G_u} is generally biased with underestimated variance when $\beta_{G_u} \neq 0$, and type I error may not be properly controlled for other hypotheses.

3.2.2 Maximum Likelihood

Let H denote the diplotype associated with the M-locus genotype G. We write $H = (h_k, h_l)$ if the diplotype consists of haplotypes h_k and h_l. In the previous subsection, we formulate the effects of G and $\mathbf{X}$ through the conditional density function $P_{\alpha,\boldsymbol{\beta},\boldsymbol{\xi}}(Y|G_u, \mathbf{W})$, where $\mathbf{W}$ consists of G_t and $\mathbf{X}$. In this subsection, we represent the same regression model in the form of $P_{\alpha,\boldsymbol{\beta},\boldsymbol{\xi}}(Y|\mathcal{G}(h_k, h_l), \mathbf{X})$, where $\mathcal{G}(h_k, h_l)$ denote the genotype G induced by the diplotype (h_k, h_l). We assume that H and $\mathbf{X}$ are independent.

Let n denote the total number of study subjects. For $i = 1, \ldots, n$, let Y_i, G_{ti} and $\mathbf{X}_i$ denote the values of Y, G_t and $\mathbf{X}$ on the ith subject. For a cross-sectional study, the likelihood for $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\pi}$ takes the form

$$L_S(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \sum_{(h_k,h_l)\sim G_{ti}} P_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\big(Y_i|\mathcal{G}(h_k, h_l), \mathbf{X}_i\big)\pi_k\pi_l, \tag{3.4}$$

where $(h_k, h_l) \sim G_{ti}$ means that the diplotype (h_k, h_l) is compatible with genotype G_{ti}.

For case-control studies, we assume the logistic regression model given in (3.1) with the linear predictor $\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathcal{G}(h_k, h_l), \mathbf{X})$. Because the sampling is conditional on the case-control status, the likelihood takes the retrospective form $\prod_{i=1}^{n} P(G_{ti}, \mathbf{X}_i|Y_i)$. If there are no environmental factors and the disease is rare, then this likelihood becomes

$$L_S(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \frac{\sum_{(h_k,h_l)\sim G_{ti}} e^{Y_i \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathcal{G}(h_k,h_l))} \pi_k \pi_l}{\sum_{k,l} e^{Y_i \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathcal{G}(h_k,h_l))} \pi_k \pi_l}, \tag{3.5}$$

where $\boldsymbol{\theta} = \boldsymbol{\beta}$. In the presence of $\mathbf{X}$, the retrospective likelihood involves the unknown distribution of $\mathbf{X}$, which is high-dimensional. We eliminate the distribution of $\mathbf{X}$ by the profile-likelihood approach (Lin and Zeng, 2006) and replace (3.5) with the following profile likelihood:

$$L_S(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \frac{\sum_{(h_k,h_l)\sim G_{ti}} e^{Y_i \left\{\mu + \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathcal{G}(h_k,h_l),\mathbf{X}_i)\right\}} \pi_k \pi_l}{\sum_{k,l,y} e^{y \left\{\mu + \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(\mathcal{G}(h_k,h_l),\mathbf{X}_i)\right\}} \pi_k \pi_l},$$

where $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, μ is an unknown constant, and the summation in the denominator is taken over $k, l = 1, \ldots, K$ and $y = 0, 1$.

The likelihood that combines the study data and the reference panel is $L_C(\boldsymbol{\theta}, \boldsymbol{\pi}) = L_S(\boldsymbol{\theta}, \boldsymbol{\pi}) L_R(\boldsymbol{\pi})$, where $L_R(\boldsymbol{\pi})$ is given in equation (3.2). We maximize this combined likelihood via the Newton-Raphson algorithm. We set the initial value of $\boldsymbol{\pi}$ at $\widetilde{\boldsymbol{\pi}}$, the maximizer of $L_R(\boldsymbol{\pi})$. To improve numerical stabilities, we exclude the haplotypes whose estimated frequencies are 0 or very close to 0, i.e., less than $\max(2/n, 0.001)$. The maximum-likelihood estimator (MLE) of $(\boldsymbol{\theta}, \boldsymbol{\pi})$ is consistent, asymptotically normal and asymptotically efficient.

Note that the likelihood for case-control studies was previously given in Lin et al. (2008) and is reformulated in this section to conform with the notation for the imputation method. The likelihood for cross-sectional studies is new.

3.3 Simulation Studies

We carried out extensive simulation studies to evaluate the performance of the MLE and imputation methods in realistic settings. We generated genotype data for various sets of five SNPs according to the LD patterns observed in the HapMap CEU sample. For each SNP set, we chose one SNP to be untyped in the study data. For some SNP sets, we picked more than one SNP to be untyped, one at a time, each representing a different scenario. Table 3.1 lists the nine scenarios used in the simulation studies, with R^2 (Stram, 2004) ranging from .41 to .98.

We explored three types of association: (1) single-SNP effects, (2) gene-environment interactions, (3) multi-SNP effects. For each type of model, we considered both cross-sectional and case-control designs. Since the case-control design naturally requires a binary trait, we focused on quantitative traits for cross-sectional studies. Thus, there were six series of simulation studies. For each set-up, we simulated 10,000 data sets with 2,000 study subjects and 60 trios. Under the case-control design, we set the overall disease rate to be approximately 1% and selected an equal number of cases and controls. We chose 60 trios for the reference panel so as to approximate the CEU sample in the current version (i.e., phase 3) of the HapMap database, which consists of 44 trios, 8 duos and 17 singletons. For each simulated data set, we applied the MLE and imputation approaches. For the latter approach, we imputed the unknown genotype by both the dosage and the most likely genotype, which are referred to as the IMP-DOS and IMP-MLG methods, respectively. All the analysis was based on the Wald statistic.

Our first series of simulation studies was concerned with the (marginal) effect of an untyped SNP on a quantitative trait in a cross-sectional study. We generated the trait value from the linear regression model

$$Y = \alpha + \beta G_u + \epsilon,$$

where ϵ is standard normal and $\alpha = 0$. Table 3.2 displays the results for various values of β. As expected, the MLE is virtually unbiased in all cases. IMP-DOS also shows negligible bias, which is not surprising because conditional mean imputation is known to yield consistent estimators of regression parameters under the linear model (Little, 1992). The estimator of β produced by IMP-MLG is seriously biased towards zero and the bias can be as much as 25% of the true parameter value. For non-zero β, both IMP-DOS and IMP-MLG tend to underestimate the variances, so their confidence intervals have poor coverage probabilities. Under scenario S8, in which $R^2 = .98$, the coverage probability of the 99% confidence interval of IMP-DOS is only 98% when $\beta = .9$. IMP-MLG is much worse than IMP-DOS because it suffers from both biased estimation of parameter and underestimation of variance; see S1-S3. As predicted by our theory, both IMP-DOS and IMP-MLG have appropriate type I error. In some cases (i.e., S2, S3 and S5), IMP-DOS is slightly more powerful than MLE. This phenomenon is attributed to the underestimation of variance by IMP-DOS. When R^2 is large (e.g., S7-S9), all methods have the same power.

In our second series of studies, we simulated case-control data under the logistic regression model

$$\Pr(Y = 1|G_u) = \frac{e^{\alpha+\beta G_u}}{1 + e^{\alpha+\beta G_u}},$$

where α was set to -4.6 to yield disease rates of approximately 1%. The results are summarized in Table 3.3. Unlike linear regression, IMP-DOS can produce substantial bias under logistic regression; see S1–S3 and S5. MLE is now uniformly more powerful than both IMP-DOS and IMP-MLG; this feature can be seen more clearly in Figure

3.1. The power gain of MLE over imputation persists as R^2 approaches 1 because MLE exploits the HWE assumption whereas imputation does not. When R^2 is low, the bias of imputation (under non-linear models) also affects its power. Again, all three methods have accurate control of type I error. As in cross-sectional studies, both IMP-DOS and IMP-MLG tend to underestimate the variances (for non-zero β) and thus yield poor confidence interval coverage, especially when β is large and R^2 is low.

Our third and fourth series of studies were focused on gene-environment interactions under the cross-sectional and case-control designs, respectively. We generated data from the same models as in the first two series but with the linear predictors $\alpha+\beta_1 G_u+\beta_2 X+\beta_3 G_u X$, where X is Bernoulli with $\Pr(X = 1) = 0.4$. The results for cross-sectional studies are displayed in Table 3.4. For detecting interactions, both IMP-DOS and IMP-MLG produce confidence intervals with poor coverage probabilities, especially when the effects are large and the LD is low; see S1-S6. Both may lose control of type I error and are substantially less powerful than MLE. The power gain of MLE is largely attributed to its incorporation of gene-environment independence. The power difference decreases as R^2 increases. In the extreme case of $R^2 = 1$, the summation in (3.4) disappears and MLE is equivalent to imputation. The results for case-control studies are shown in Table 3.5. Both imputation methods yield biased estimates, poor confidence interval coverage and diminished power. The power difference between MLE and imputation is further illustrated in Figure 3.2. The power gain of MLE is again largely attributed to its use of gene-environment independence. If we analyzed the imputed genotypes (either the dosage or the most likely genotype) by the method of Chatterjee and Carroll (2005), which also exploits gene-environment independence, then the power gain of MLE was reduced considerably (results not shown).

Our last two series of studies dealt with multi-SNP effects. We set the untyped SNP to be causal and included all five SNPs in the joint analysis. For making inference on the effect of the untyped SNP, the performance of IMP-DOS and IMP-MLG is similar to the first two series of studies (results not shown). In particular, type I error is properly controlled. This is not surprising because our theory indicates that imputation yields a valid test of the untyped SNP even when there are environmental factors or typed SNPs in the model. On the other hand, if the untyped SNP is associated with the trait, the bias in the estimation of its effect can cause bias in estimating the null effects of the typed SNPs. Indeed, both IMP-DOS and IMP-MLG can have inflated type I error in testing the effects of the typed SNPs and the inflation of type I error becomes more severe as the effect of the untyped SNP increases. Figure 3.3 and 3.4 display these results for cross-sectional and case-control studies, respectively. As before, MLE has accurate control of type I error.

3.4 WTCCC Data

We considered WTCCC data on type 1 diabetes (T1D). The database contains 1,963 subjects with T1D and 2,938 controls. For the typed SNPs, we applied the standard Armitage trend test. For the untyped SNPs that are cataloged in the HapMap phase 3 database, we applied both MLE and IMP-DOS, with the phase 3 HapMap CEU sample as the reference panel. For each untyped SNP, we first identified the typed SNPs within 50 kb and then found a set of four that yields the largest R^2. If there were fewer than eight SNPs within 50 kb, we enlarged the window until a minimum of eight SNPs were located. If there were more than twenty SNPs within 50 kb, we restricted our attention to the closest twenty SNPs so as to reduce computation time.

As shown in Figure 3.5, MLE and IMP-DOS produce nearly identical quantile-quantile (Q-Q) plots for the untyped SNPs, which are similar to that of the typed SNPs. The deviations of the test statistics from the null distribution are minor except in the extreme tails, which correspond to significant associations. The over-dispersion parameter (i.e., the genomic control λ) was estimated at approximately 1.05 for all three plots. These results illustrate that, for single-SNP analysis, both MLE and imputation have correct type I error.

Figure 3.6 displays the results of the association tests for both typed and untyped SNPs on chromosomes 1, 6 and 12, which have the strongest evidence of association. Both MLE and IMP-DOS were able to identify untyped SNPs that are more strongly associated with the disease than typed SNPs, but MLE picked out those SNPs more clearly. This is not surprising since MLE is expected to be more powerful than imputation.

3.5 Discussion

We have presented two approaches to the analysis of untyped SNPs and investigated their properties both theoretically and numerically. The maximum-likelihood approach yields approximately unbiased parameter estimators, proper confidence intervals and accurate control of type I error. It tends to be more powerful than the imputation approach, especially for case-control studies and in testing gene-environment interactions. The maximum-likelihood method requires the study sample and reference panel be generated from the same underlying population and may be numerically unstable when the haplotype frequencies are low.

We have assumed gene-environment independence in the maximum likelihood approach. This assumption is satisfied in most applications and can substantially improve the efficiency of association analysis, especially in case-control studies (Chatterjee and Carroll, 2005). It is possible to allow gene-environment dependence, but the analysis will be more complicated and less efficient.

The imputation approach has some advantages over the maximum likelihood approach. Numerically, the former is more stable than the latter. For single-SNP tests, imputation has proper control of type I error even if the reference panel does not match the study population. For testing other hypotheses, however, imputation may have inflated type I error. In general, imputation yields biased parameter estimators and incorrect variance estimators. Because the bias can be upward and the variance is underestimated, imputation can sometimes be more powerful than maximum likelihood. Thus, maximum likelihood and imputation are complementary to each other. One possible strategy is to use imputation (with the dosage as the imputed value) in the initial single-SNP tests and to use maximum likelihood for more complex analysis once a region of disease association has been identified.

For cross-sectional studies, Xie and Stram (2005) showed that the score test based on the dosage of the risk haplotype is asymptotically valid. We have shown that imputation is asymptotically valid for single-SNP tests under both cross-sectional and case-control designs whether the untyped SNP is imputed by the dosage or the most likely genotype. Note that haplotype analysis does not involve external data whereas analysis of untyped SNPs does.

Because it ignores the random variation of the reference panel, the imputation approach generally underestimates the variances of the parameter estimators. As the size of the reference panel increases, the underestimation of variance becomes less severe and thus confidence intervals have better coverage probabilities. The size of the reference panel, however, has little influence on the bias of imputation. On the other hand, increasing the size of the reference panel reduces the variance of the MLE. Indeed, the power of maximum likelihood improves at a faster rate than imputation as the reference panel becomes larger, especially under the case-control design (results not shown).

Both MLE and imputation are computationally fast, and the relevant software is available at our website. It took about 8 hours on a 64 bit, 30 GHz Intel Xeon machine to perform the MLE analysis on chromosome 1 of the WTCCC GWAS data. Imputation was slightly faster. The computational savings of imputation will be more substantial if there are multiple traits of interest because the untyped SNPs only need to be imputed once.

For computational expediency, we used the significance level of 0.01 in our simulation studies. The relatively small number of replicates required for obtaining accurate summary statistics at this significance level allowed us to explore a very wide variety of scenarios. It would be formidable to conduct extensive simulation studies at the significance level of 10^{-4} or lower, which would require at least 1 million replicates. We repeated some of our simulation studies using the significance level of 10^{-4}, and the basic conclusions regarding the relative merits of MLE and imputation remained the same.

We have focused on tagging-based imputation. An alternative approach is to use HMM (Browning and Browning, 2007; Marchini et al., 2007; Li et al., 2010). The latter approach, which explores the LD information over a larger region and incorporates population genetics knowledge, can yield more accurate prediction of untyped genotypes in certain situations. We chose tagging over HMM in this chapter for several reasons: (1) using the same amount of information to infer missing genotypes ensures that the maximum-likelihood and imputation methods are compared on equal footing; (2) an investigation by the imputation working group of GAIN (Manolio et al., 2007) revealed that tagging is nearly as accurate as HMM (unpublished data); (3) tagging is much simpler and faster than HMM and can handle much larger studies. We are currently trying to incorporate HMM into the maximum likelihood framework. The conclusions of this chapter regarding the relative merits of the maximum likelihood versus imputation approaches are expected to hold when tagging is replaced by HMM.

Table 3.1: Haplotype frequencies for the scenarios used in simulation studies

	S1: R^2 = .41, MAF=.39		S2: R^2 = .59, MAF=.28		S3: R^2 = .70, MAF=.15	
Haplotype	UTTTT	Frequency	TTTUT	Frequency	TTUTT	Frequency
h_1	00011	.0513	00100	.3171	00000	.0513
h_2	00100	.0260	00101	.0988	00100	.1460
h_3	01011	.0855	00111	.2027	01000	.6958
h_4	01100	.3094	01001	.1518	10011	.1069
h_5	01101	.1377	10100	.1059		
h_6	11011	.0085	10101	.0209		
h_7	11100	.1775	10111	.0793		
h_8	11101	.0247	11001	.0235		
h_9	11111	.1794				

	S4: R^2 = .81, MAF=.33		S5: R^2 = .84, MAF=.24		S6: R^2 = .93, MAF=.15	
Haplotype	UTTTT	Frequency	TTUTT	Frequency	TTUTT	Frequency
h_1	00011	.2846	01101	.2852	00011	.0513
h_2	00101	.0128	10100	.2510	00100	.0260
h_3	00111	.0342	11000	.2393	01011	.0855
h_4	10101	.2374	11100	.0321	01100	.3094
h_5	10111	.1917	11101	.0963	01101	.1377
h_6	11110	.2393	11110	.0961	11011	.0085
h_7					11100	.1775
h_8					11101	.0247
h_9					11111	.1794

	S7: R^2 = .95, MAF=.09		S8: R^2 = .98, MAF=.28		S9: R^2 = .98, MAF=.29	
Haplotype	UTTTT	Frequency	TTUTT	Frequency	TTTTU	Frequency
h_1	00111	.3809	01000	.4231	01000	.4231
h_2	01110	.2350	01010	.1154	01010	.1154
h_3	01111	.2900	01011	.0043	01011	.0043
h_4	11001	.0897	01111	.2821	01111	.2821
h_5	11111	.0044	10010	.1751	10010	.1751

NOTE: "U" and "T" indicate the untyped and typed SNP positions, respectively. R^2 is the squared correlation between the expected and true allele counts (Stram 2004). MAF is the minor allele frequency of the untyped SNP.

Table 3.2: Simulation results for studying the effect of an untyped SNP on a quantitative trait under the cross-sectional design

		MLE					IMP-DOS					IMP-MLG				
	β	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW
S1	.0	.000	.051	.051	.989	.011	.000	.051	.051	.990	.010	.000	.041	.041	.989	.011
	.1	.000	.051	.051	.990	.274	-.001	.051	.051	.990	.267	-.031	.041	.041	.965	.190
	.2	.000	.052	.051	.990	.905	-.002	.051	.051	.990	.903	-.062	.042	.041	.857	.781
	.6	.000	.053	.052	.990	1.00	-.007	.056	.053	.987	1.00	-.185	.047	.043	.063	1.00
	.9	-.002	.054	.052	.988	1.00	-.010	.061	.056	.982	1.00	-.277	.053	.046	.001	1.00
S2	.0	.000	.046	.046	.991	.009	.000	.046	.046	.990	.010	.000	.036	.036	.991	.009
	.1	.000	.046	.047	.991	.325	.000	.046	.046	.990	.336	-.026	.036	.036	.965	.309
	.2	.000	.048	.048	.992	.960	.000	.048	.046	.988	.963	-.051	.038	.036	.843	.937
	.6	.001	.057	.057	.988	1.00	-.001	.061	.047	.954	1.00	-.154	.048	.037	.151	1.00
	.9	.000	.060	.060	.985	1.00	-.001	.077	.049	.903	1.00	-.231	.060	.038	.059	1.00
S3	.0	.000	.055	.054	.992	.008	.000	.055	.054	.991	.009	.000	.041	.041	.990	.010
	.1	.000	.055	.055	.992	.217	.001	.055	.054	.989	.237	-.025	.041	.041	.977	.233
	.2	.001	.057	.057	.991	.856	.001	.058	.054	.986	.871	-.050	.042	.041	.907	.863
	.6	.001	.070	.070	.987	1.00	.004	.078	.055	.936	1.00	-.149	.046	.041	.172	1.00
	.9	.000	.074	.073	.987	1.00	.006	.100	.056	.863	1.00	-.224	.051	.042	.033	1.00
S4	.0	.000	.037	.037	.989	.011	.000	.037	.037	.989	.011	.000	.035	.035	.989	.011
	.1	.000	.037	.037	.989	.544	.000	.037	.037	.989	.544	-.007	.035	.035	.986	.540
	.2	.000	.038	.037	.989	.998	-.001	.037	.037	.988	.998	-.013	.035	.035	.983	.997
	.6	.000	.038	.038	.988	1.00	-.002	.039	.038	.986	1.00	-.039	.036	.036	.929	1.00
	.9	.000	.039	.038	.989	1.00	-.003	.041	.038	.984	1.00	-.059	.036	.036	.829	1.00
S5	.0	.000	.041	.040	.991	.009	.000	.041	.040	.991	.009	.000	.036	.036	.991	.009
	.1	.000	.041	.041	.991	.456	.000	.041	.040	.990	.463	-.012	.036	.036	.985	.463
	.2	.001	.042	.042	.991	.991	.001	.042	.040	.988	.991	-.023	.036	.036	.971	.991
	.6	.001	.048	.048	.988	1.00	.001	.050	.041	.966	1.00	-.071	.036	.036	.722	1.00
	.9	.001	.052	.051	.987	1.00	.002	.060	.041	.929	1.00	-.106	.037	.036	.365	1.00
S6	.0	.000	.050	.050	.990	.010	.000	.050	.050	.990	.010	.000	.048	.047	.990	.010
	.1	.000	.050	.050	.990	.286	.000	.050	.050	.990	.284	-.008	.048	.047	.988	.266
	.2	.000	.050	.050	.990	.919	-.001	.050	.050	.990	.918	-.015	.048	.047	.985	.904
	.6	.000	.050	.050	.990	1.00	-.003	.050	.050	.990	1.00	-.046	.048	.048	.943	1.00
	.9	.000	.050	.049	.989	1.00	-.004	.051	.051	.991	1.00	-.069	.049	.049	.876	1.00
S7	.0	.000	.056	.056	.990	.010	.000	.056	.056	.990	.010	.000	.055	.055	.990	.010
	.1	.000	.056	.056	.990	.215	.000	.056	.056	.990	.214	.000	.055	.055	.990	.215
	.2	.000	.056	.056	.990	.844	.000	.056	.056	.990	.844	-.001	.055	.055	.990	.845
	.6	.000	.056	.056	.990	1.00	.000	.056	.056	.990	1.00	-.003	.055	.055	.990	1.00
	.9	.000	.056	.056	.990	1.00	.000	.056	.056	.990	1.00	-.004	.056	.056	.990	1.00
S8	.0	.000	.036	.036	.990	.010	.000	.036	.036	.990	.010	.000	.035	.035	.990	.010
	.1	.000	.036	.036	.990	.590	.000	.036	.036	.990	.590	-.002	.035	.035	.991	.590
	.2	.000	.036	.036	.991	.999	.000	.036	.036	.990	.999	-.003	.035	.035	.991	.999
	.6	.000	.037	.037	.990	1.00	.000	.037	.036	.987	1.00	-.009	.035	.035	.988	1.00
	.9	.000	.038	.038	.989	1.00	.000	.038	.036	.982	1.00	-.014	.035	.035	.984	1.00
S9	.0	.000	.036	.035	.990	.010	.000	.036	.035	.990	.010	.000	.035	.035	.990	.010
	.1	.000	.036	.035	.990	.599	.000	.036	.035	.990	.599	-.001	.035	.035	.990	.598
	.2	.000	.036	.035	.990	.999	.000	.036	.035	.990	.999	-.001	.035	.035	.990	.999
	.6	.000	.036	.035	.990	1.00	.000	.036	.035	.990	1.00	-.004	.035	.035	.991	1.00
	.9	-.001	.036	.035	.990	1.00	-.001	.036	.035	.989	1.00	-.005	.035	.035	.990	1.00

NOTE: S1-S9 denote the nine scenarios listed in Table 3.1. Bias and SE are the bias and standard error of the parameter estimator. SEE is the mean of the standard error estimator. CP is the coverage probability of the 99% confidence interval. PW is the type I error/power for testing zero parameter value at the .01 nominal significance level. Each entry is based on 10,000 replicates.

Table 3.3: Simulation results for studying the effect of an untyped SNP on the risk of disease under the case-control design

		MLE					IMP-DOS					IMP-MLG				
	β	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW
S1	.0	.001	.100	.099	.989	.011	.001	.101	.101	.990	.010	.001	.082	.082	.990	.010
	.3	.000	.103	.101	.990	.651	-.010	.101	.100	.989	.633	-.093	.082	.081	.922	.483
	.6	.000	.108	.106	.989	.999	-.031	.102	.099	.984	.999	-.190	.082	.081	.580	.993
	.9	-.008	.115	.113	.986	1.00	-.070	.104	.100	.961	1.00	-.297	.083	.082	.156	1.00
S2	.0	-.002	.091	.090	.991	.009	-.001	.093	.092	.989	.011	-.001	.073	.072	.988	.012
	.3	.000	.085	.084	.989	.844	.012	.094	.091	.987	.805	-.067	.075	.072	.934	.753
	.6	-.002	.084	.082	.990	1.00	.046	.103	.092	.971	1.00	-.117	.083	.073	.757	1.00
	.9	-.009	.084	.083	.989	1.00	.093	.115	.094	.919	1.00	-.157	.097	.075	.566	1.00
S3	.0	-.002	.106	.106	.992	.008	-.001	.109	.108	.989	.011	.000	.082	.082	.989	.011
	.3	-.001	.099	.098	.991	.693	.012	.110	.105	.987	.651	-.067	.081	.079	.953	.641
	.6	.000	.096	.096	.990	1.00	.047	.120	.103	.969	1.00	-.116	.082	.078	.841	1.00
	.9	-.003	.096	.097	.989	1.00	.096	.136	.103	.915	1.00	-.153	.087	.078	.682	1.00
S4	.0	.000	.073	.073	.990	.010	.000	.075	.074	.989	.011	.000	.071	.070	.989	.011
	.3	.001	.075	.074	.990	.929	.002	.078	.077	.990	.913	-.016	.074	.073	.986	.909
	.6	.000	.077	.077	.989	1.00	.007	.082	.081	.990	1.00	-.028	.077	.077	.984	1.00
	.9	-.006	.082	.081	.990	1.00	.011	.088	.087	.989	1.00	-.039	.083	.082	.979	1.00
S5	.0	.000	.079	.079	.989	.011	.000	.082	.081	.989	.011	.000	.072	.071	.989	.011
	.3	.001	.086	.087	.990	.837	-.005	.085	.084	.988	.830	-.040	.074	.074	.977	.830
	.6	.002	.103	.102	.989	1.00	-.023	.093	.088	.980	1.00	-.092	.078	.077	.912	1.00
	.9	.009	.130	.127	.988	1.00	-.051	.102	.093	.961	1.00	-.154	.082	.081	.744	1.00
S6	.0	.000	.096	.097	.991	.009	.000	.100	.100	.991	.009	.000	.095	.095	.991	.009
	.3	.001	.102	.103	.990	.634	.000	.105	.106	.989	.603	-.022	.100	.101	.989	.578
	.6	-.002	.112	.112	.989	.999	-.007	.114	.114	.990	.997	-.049	.108	.108	.981	.996
	.9	-.004	.122	.121	.989	1.00	-.014	.125	.123	.987	1.00	-.077	.118	.117	.964	1.00
S7	.0	.000	.109	.108	.991	.009	.000	.112	.111	.992	.008	.000	.112	.111	.992	.008
	.3	-.001	.103	.102	.989	.632	-.001	.106	.106	.991	.601	-.002	.106	.105	.991	.599
	.6	.000	.097	.098	.990	1.00	.001	.102	.102	.990	1.00	-.002	.101	.101	.990	1.00
	.9	-.003	.096	.095	.989	1.00	-.001	.101	.100	.989	1.00	-.005	.100	.099	.989	1.00
S8	.0	.000	.070	.069	.989	.011	.000	.072	.071	.989	.011	.000	.071	.070	.989	.011
	.3	.000	.066	.067	.990	.972	.002	.069	.070	.990	.962	-.003	.068	.069	.989	.962
	.6	-.001	.066	.066	.990	1.00	.004	.071	.070	.990	1.00	-.006	.069	.069	.989	1.00
	.9	-.006	.066	.066	.990	1.00	.005	.072	.071	.990	1.00	-.008	.070	.070	.991	1.00
S9	.0	.000	.069	.069	.988	.012	.000	.071	.071	.989	.011	.000	.071	.070	.989	.011
	.3	.000	.067	.067	.992	.971	.000	.069	.069	.991	.960	-.001	.069	.069	.991	.960
	.6	-.002	.066	.066	.991	1.00	-.001	.069	.069	.992	1.00	-.004	.069	.069	.992	1.00
	.9	-.007	.067	.066	.989	1.00	-.002	.072	.071	.989	1.00	-.007	.071	.070	.988	1.00

NOTE: See the Note to Table 3.2.

Table 3.4: Simulation results for studying gene-environment interactions under the cross-sectional design with a quantitative trait

			MLE					IMP-DOS					IMP-MLG				
	β_1	β_3	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW
S1	.5	.0	.000	.097	.096	.990	.010	-.002	.107	.107	.991	.009	-.001	.087	.087	.990	.010
	.5	.2	.000	.095	.094	.990	.328	-.004	.108	.109	.991	.210	-.063	.088	.089	.970	.150
	.5	.3	.000	.094	.093	.990	.740	-.005	.108	.111	.992	.531	-.094	.089	.090	.943	.385
	.5	.9	.000	.092	.091	.990	1.00	-.012	.115	.121	.993	1.00	-.279	.098	.100	.409	1.00
	1.2	.0	.000	.085	.084	.991	.009	-.002	.116	.123	.994	.006	-.001	.099	.102	.992	.008
S2	.5	.0	-.001	.092	.092	.991	.009	-.002	.097	.096	.990	.010	-.001	.076	.075	.991	.009
	.5	.2	-.001	.092	.091	.990	.352	-.002	.098	.097	.989	.303	-.053	.077	.076	.964	.273
	.5	.3	-.001	.092	.091	.990	.763	-.002	.100	.098	.989	.682	-.078	.078	.076	.926	.633
	.5	.9	-.001	.091	.090	.991	1.00	-.003	.118	.103	.977	1.00	-.232	.092	.081	.377	1.00
	1.2	.0	.000	.087	.087	.989	.011	-.002	.107	.105	.989	.011	-.001	.083	.083	.992	.009
S3	.5	.0	.000	.110	.110	.991	.009	.000	.113	.112	.989	.011	.000	.085	.084	.989	.011
	.5	.2	.001	.110	.110	.990	.224	.001	.115	.112	.987	.223	-.050	.086	.085	.974	.218
	.5	.3	.000	.111	.111	.989	.560	.002	.118	.113	.986	.541	-.075	.087	.085	.951	.533
	.5	.9	.000	.110	.111	.991	1.00	.006	.144	.116	.964	1.00	-.224	.094	.087	.487	1.00
	1.2	.0	.000	.108	.108	.990	.010	-.001	.124	.116	.984	.017	.000	.093	.088	.984	.016
S4	.5	.0	-.001	.077	.076	.988	.012	-.001	.077	.077	.990	.010	-.001	.073	.072	.989	.011
	.5	.2	-.001	.077	.076	.988	.516	-.002	.077	.077	.990	.496	-.014	.073	.073	.988	.488
	.5	.3	-.001	.077	.076	.988	.913	-.002	.078	.077	.991	.901	-.021	.073	.073	.987	.895
	.5	.9	-.001	.076	.075	.988	1.00	-.004	.079	.080	.990	1.00	-.060	.074	.076	.966	1.00
	1.2	.0	-.001	.076	.075	.988	.012	-.001	.080	.080	.990	.010	-.001	.076	.076	.990	.010
S5	.5	.0	-.001	.084	.083	.990	.011	-.001	.084	.083	.989	.011	-.001	.074	.073	.989	.011
	.5	.2	-.001	.084	.083	.989	.432	-.001	.085	.083	.988	.432	-.025	.075	.073	.985	.432
	.5	.3	-.001	.085	.084	.989	.842	-.001	.086	.084	.988	.840	-.036	.075	.074	.978	.840
	.5	.9	-.001	.085	.084	.989	1.00	.000	.097	.085	.978	1.00	-.108	.077	.075	.872	1.00
	1.2	.0	-.001	.084	.083	.988	.012	-.001	.089	.086	.987	.013	-.001	.078	.076	.987	.013
S6	.5	.0	-.001	.102	.101	.989	.011	.000	.102	.102	.990	.010	-.001	.097	.098	.990	.010
	.5	.2	-.001	.102	.101	.989	.277	-.001	.102	.103	.990	.261	-.016	.098	.098	.990	.245
	.5	.3	-.001	.101	.101	.989	.649	-.002	.102	.103	.990	.621	-.024	.098	.098	.989	.591
	.5	.9	-.001	.101	.100	.989	1.00	-.005	.103	.105	.991	1.00	-.070	.099	.100	.973	1.00
	1.2	.0	-.001	.099	.099	.989	.011	.000	.103	.105	.991	.009	-.001	.100	.101	.991	.009
S7	.5	.0	-.002	.114	.114	.991	.009	-.002	.114	.114	.991	.009	-.002	.114	.113	.991	.009
	.5	.2	-.002	.114	.114	.991	.202	-.002	.114	.114	.991	.202	-.003	.114	.113	.991	.201
	.5	.3	-.002	.114	.114	.990	.515	-.002	.114	.114	.991	.512	-.004	.114	.113	.991	.512
	.5	.9	-.004	.114	.114	.990	1.00	-.003	.114	.114	.991	1.00	-.007	.114	.114	.991	1.00
	1.2	.0	-.002	.114	.114	.990	.010	-.002	.115	.114	.991	.009	-.002	.114	.114	.991	.009
S8	.5	.0	.001	.073	.073	.989	.011	.001	.074	.073	.989	.011	.001	.072	.072	.989	.011
	.5	.2	.001	.074	.073	.989	.572	.001	.074	.073	.989	.572	-.002	.072	.072	.989	.572
	.5	.3	.001	.074	.073	.989	.939	.001	.074	.073	.990	.939	-.004	.072	.072	.989	.939
	.5	.9	-.001	.074	.073	.989	1.00	.001	.075	.073	.989	1.00	-.013	.073	.072	.987	1.00
	1.2	.0	.001	.074	.073	.990	.011	.001	.074	.073	.989	.011	.001	.073	.072	.989	.011
S9	.5	.0	.001	.073	.072	.990	.010	.001	.073	.072	.990	.010	.001	.073	.072	.990	.010
	.5	.2	.001	.073	.072	.990	.579	.001	.073	.072	.990	.577	.000	.073	.072	.990	.577
	.5	.3	.000	.073	.072	.990	.943	.001	.073	.072	.990	.942	-.001	.073	.072	.990	.943
	.5	.9	-.001	.073	.072	.990	1.00	.000	.073	.073	.990	1.00	-.005	.073	.072	.989	1.00
	1.2	.0	.001	.073	.072	.990	.010	.001	.073	.073	.990	.011	.001	.073	.072	.990	.010

NOTE: $\beta_2 = .2$. See the Note to Table 3.2.

Table 3.5: Simulation results for studying gene-environment interactions under the case-control design

		MLE					IMP-DOS					IMP-MLG				
	$\beta3$	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW	Bias	SE	SEE	CP	PW
S1	.0	.000	.151	.149	.990	.010	-.002	.210	.209	.990	.010	-.002	.169	.169	.991	.009
	.5	-.008	.165	.163	.988	.679	-.021	.218	.216	.988	.359	-.157	.175	.175	.952	.270
	.9	-.025	.188	.187	.987	.985	-.065	.233	.231	.985	.859	-.295	.187	.187	.840	.751
S2	.0	.000	.136	.135	.992	.008	-.001	.189	.189	.991	.009	-.001	.148	.147	.991	.009
	.5	-.006	.140	.139	.991	.863	.032	.198	.194	.989	.565	-.103	.157	.152	.965	.521
	.9	-.023	.153	.153	.989	1.00	.098	.216	.204	.978	.990	-.153	.174	.160	.927	.980
S3	.0	.002	.162	.160	.992	.008	.001	.226	.223	.991	.009	.000	.170	.168	.990	.010
	.5	-.002	.160	.158	.990	.756	.034	.232	.223	.987	.424	-.100	.171	.168	.973	.421
	.9	-.016	.167	.166	.987	1.00	.099	.247	.227	.977	.968	-.152	.177	.171	.948	.967
S4	.0	-.002	.110	.109	.988	.012	-.002	.154	.153	.990	.010	-.002	.145	.144	.990	.010
	.5	-.011	.130	.129	.989	.882	.004	.169	.169	.990	.660	-.026	.160	.159	.989	.657
	.9	-.036	.157	.158	.990	.996	.007	.193	.194	.990	.977	-.044	.182	.183	.989	.977
S5	.0	-.001	.118	.118	.992	.008	.001	.166	.166	.991	.009	.001	.146	.146	.991	.009
	.5	-.011	.150	.150	.991	.756	-.014	.187	.185	.989	.521	-.072	.163	.163	.982	.521
	.9	-.031	.200	.199	.991	.963	-.056	.221	.216	.987	.904	-.157	.191	.190	.965	.904
S6	.0	-.003	.146	.146	.990	.010	-.003	.205	.206	.992	.008	-.003	.196	.195	.993	.007
	.5	-.017	.186	.185	.991	.521	-.005	.238	.235	.990	.329	-.041	.226	.223	.988	.307
	.9	-.049	.243	.241	.993	.806	-.023	.282	.280	.991	.701	-.087	.268	.266	.989	.682
S7	.0	.003	.165	.163	.991	.009	.000	.230	.230	.990	.010	.000	.229	.229	.990	.010
	.5	-.004	.157	.156	.991	.735	-.001	.227	.226	.988	.356	-.003	.226	.225	.989	.354
	.9	-.015	.157	.156	.990	1.00	.000	.225	.226	.992	.926	-.004	.224	.225	.991	.926
S8	.0	.001	.104	.104	.991	.009	.000	.146	.146	.991	.009	.000	.144	.144	.991	.009
	.5	-.007	.107	.107	.990	.985	.005	.149	.149	.991	.792	-.003	.147	.147	.991	.792
	.9	-.025	.116	.116	.986	1.00	.008	.159	.157	.989	.999	-.006	.156	.155	.989	.999
S9	.0	.001	.104	.103	.991	.009	.000	.146	.145	.990	.010	.000	.145	.145	.990	.010
	.5	-.007	.108	.107	.989	.983	.002	.148	.149	.989	.792	-.001	.147	.148	.990	.792
	.9	-.026	.117	.116	.986	1.00	.000	.158	.157	.991	.999	-.006	.157	.156	.990	.999

NOTE: $\beta_1 = .0$, $\beta_2 = .1$. See the Note to Table 3.2.

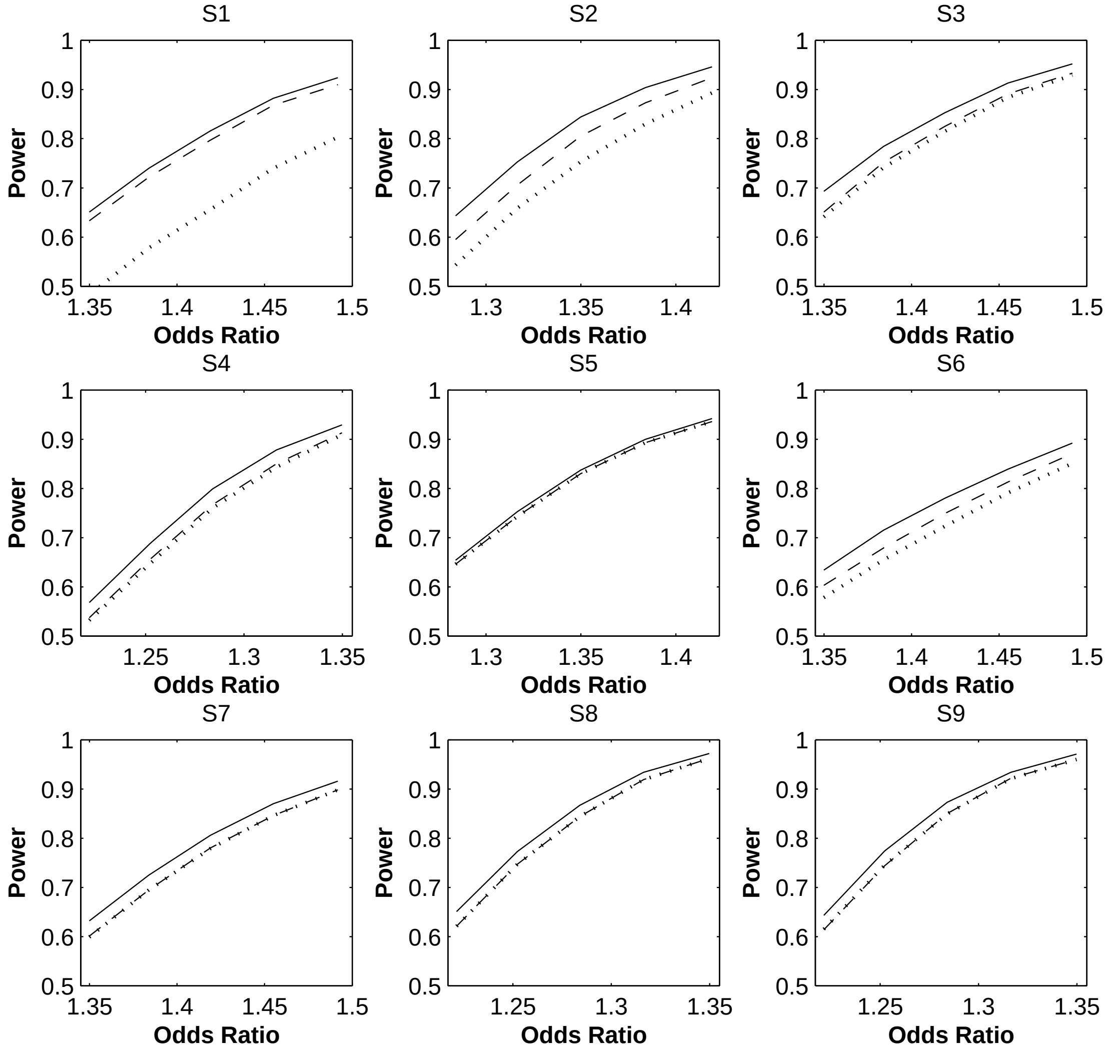

Figure 3.1: Power of testing the effect of an untyped SNP at the 1% nominal significance level under the case-control design. The solid, dashed and dotted curves pertain to MLE, IMP-DOS and IMP-MLG, respectively.

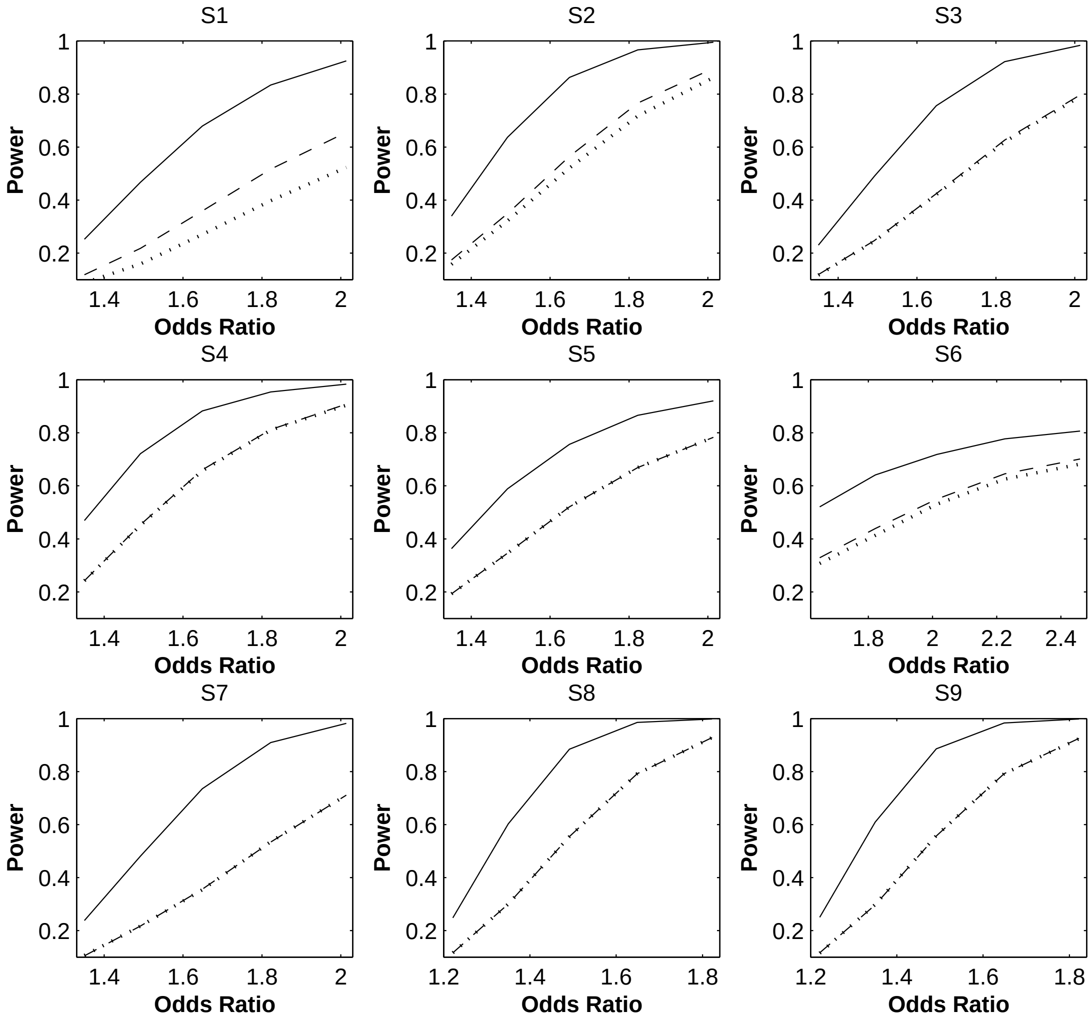

Figure 3.2: Power of testing gene-environment interactions at the 1% nominal significance level under the case-control design. The solid, dashed and dotted curves pertain to MLE, IMP-DOS and IMP-MLG, respectively.

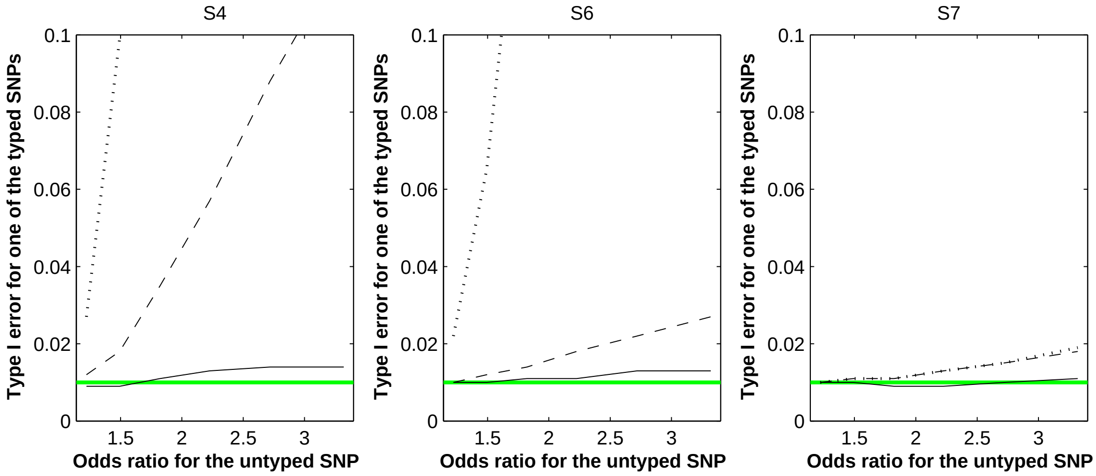

Figure 3.3: Type I error for testing the null effect of a typed SNP on a quantitative trait at the 1% nominal significance level in the joint analysis involving a causal, untyped SNP under the cross-sectional design. The solid, dashed and dotted curves pertain to MLE, IMP-DOS and IMP-MLG, respectively.

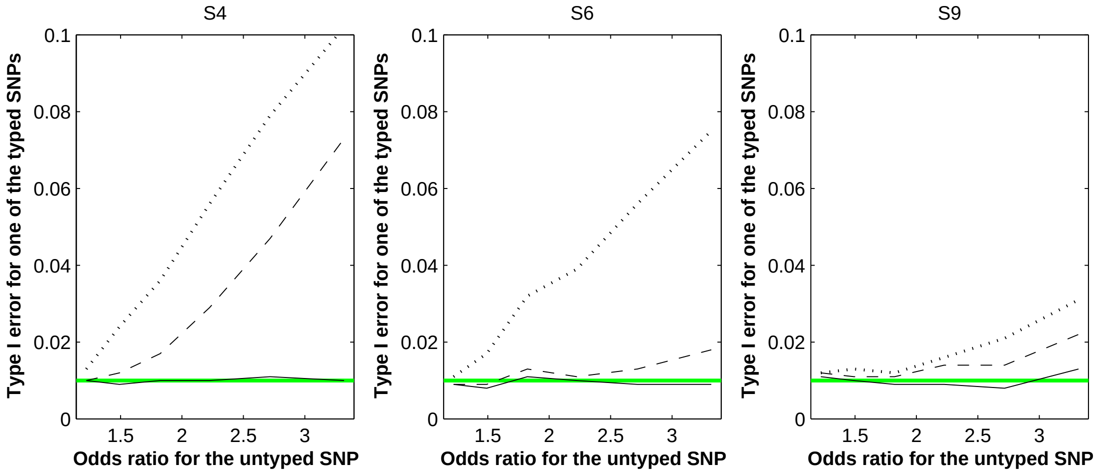

Figure 3.4: Type I error for testing the null effect of a typed SNP at the 1% nominal significance level in the joint analysis involving a causal, untyped SNP under the case-control design. The solid, dashed and dotted curves pertain to MLE, IMP-DOS and IMP-MLG, respectively.

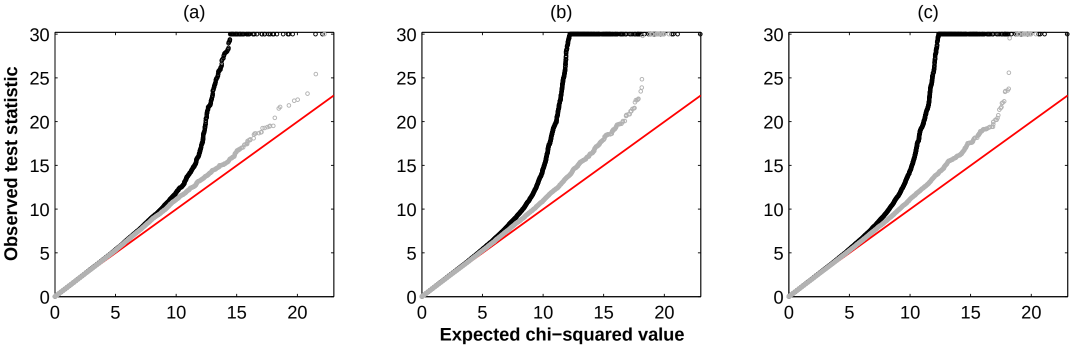

Figure 3.5: Q-Q plots for the single-SNP analysis of the T1D data from the WTCCC study: (a) Armitage trend test for typed SNPs, (b) MLE for untyped SNPs, and (c) IMP-DOS for untyped SNPs. Chi-squared statistics exceeding 30 are truncated. The black curve in (a) pertains to 392,746 typed SNPs that pass the standard project filters, have minor allele frequencies (MAF) $> 1\%$ and missing data rates $< 1\%$, and have good cluster plots. The black curves in (b) and (c) pertain to 819,727 untyped SNPs that are cataloged in Phase 3 of HapMap with MAF $> 1\%$. The Q-Q plots which exclude all SNPs located in the regions of association listed in Table 3 of the WTCCC (Burton et al., 2007) paper are superimposed in grey. The grey curves show that departures in the extreme tails of the distributions of test statistics are due to regions with strong signals for association.

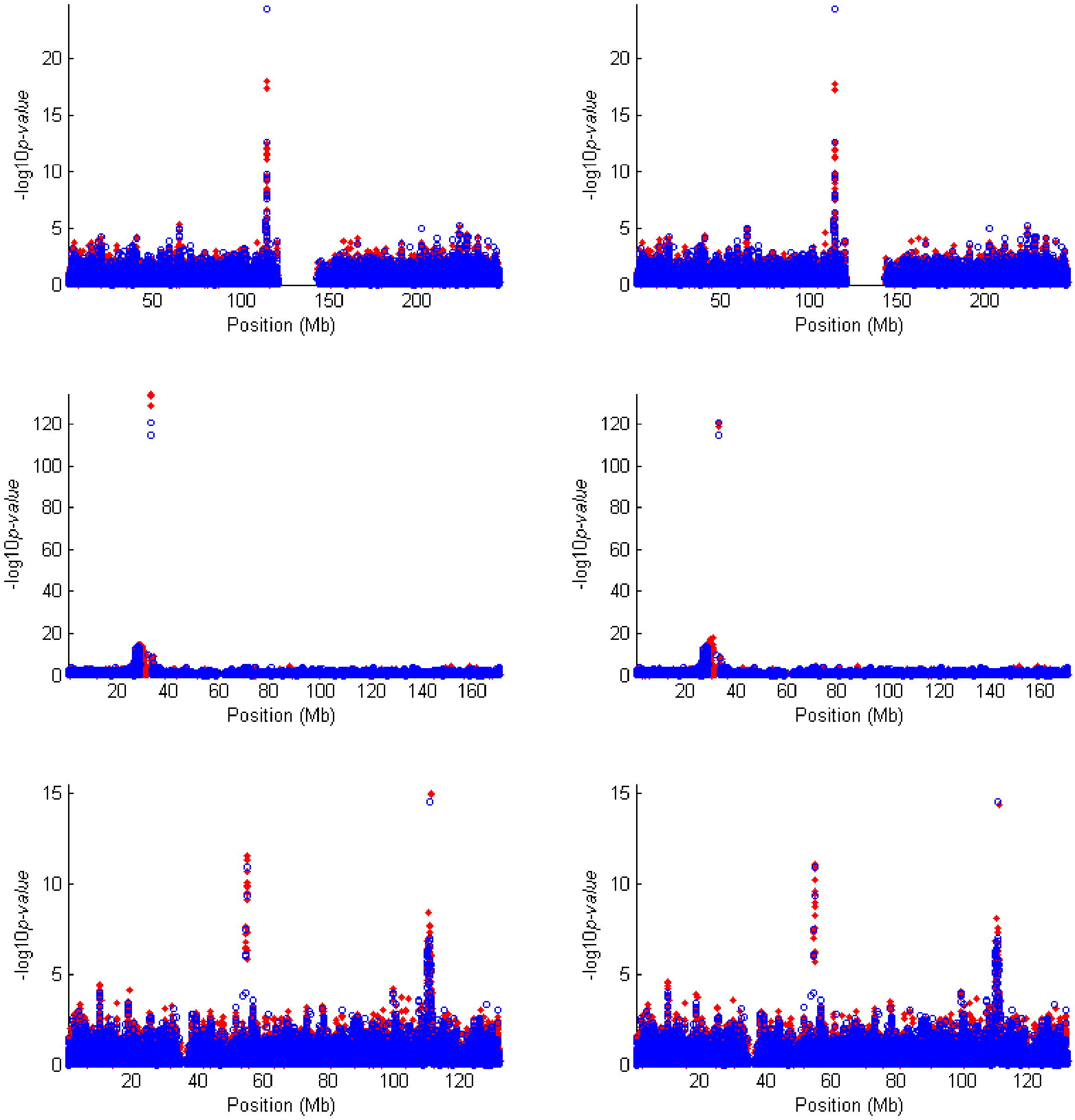

Figure 3.6: Results of single-SNP association tests for the WTCCC study of T1D. The $\log_{10}$ p-values for typed SNPs and untyped SNPs are shown in blue circles and red dots, respectively. The three rows correspond to chromosomes 1, 6 and 12, which have the strongest evidence of association. The left column corresponds to the trend test for the typed SNPs and the MLE method for the untyped SNPs. The right column corresponds to the trend test for the typed SNPs and the IMP-DOS method for the untyped SNPs. All typed SNPs pass the standard project filters, have MAF $> 1\%$ and missing data rate $< 1\%$, and have good cluster plots. All untyped SNPs have MAF $> 1\%$ in HapMap.

3.6 Appendix

We are interested in the effect of the untyped SNP genotype G_u on the phenotype Y adjusted for the effects of covariates $\mathbf{W}$ (if there are any). The covariates, which are required to be fully observed, may include environmental factors and typed SNPs and are allowed to be correlated with the untyped SNP. The linear predictor is assumed to take the form of $\alpha+\beta_{G_u}G_u+\boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}$, where β_{G_u} and $\boldsymbol{\beta}_{\mathbf{W}}$ represent the regression effects of G_u and $\mathbf{W}$, respectively. Write $\boldsymbol{\beta} = (\beta_{G_u}, \boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}})^{\mathrm{T}}$. We are particularly interested in testing the null hypothesis $H_0 : \beta_{G_u} = 0$.

Let n denote the total number of study subjects. For $i = 1, \ldots, n$, let Y_i, G_{ui} and $\mathbf{W}_i$ denote the values of Y, G_u and $\mathbf{W}$ on the ith subject. We replace G_{ui} by $\widehat{G}_{ui}$, where $\widehat{G}_{ui}$ is the imputed value of G_{ui} based on equation (3.3), and then apply standard likelihood methods to the imputed data set $(Y_i, \widehat{G}_{ui}, \mathbf{W}_i)$ $(i = 1, \ldots, n)$. The validity of such analysis does not follow from standard likelihood theory because the n imputed values $\{\widehat{G}_{ui}\}$ $(i = 1, \ldots, n)$ are correlated due to the presence of the estimator $\widetilde{\boldsymbol{\pi}}$ in them.

We first consider cross-sectional studies. The "likelihood" for $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}$ based on the imputed data set takes the form $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} P_{\alpha,\boldsymbol{\beta},\boldsymbol{\xi}}(Y_i|\widehat{G}_{ui}, \mathbf{W}_i)$. Denote the resulting estimator by $\widehat{\boldsymbol{\theta}}$. As mentioned above, standard likelihood theory is not applicable to $\widehat{\boldsymbol{\theta}}$ because the n terms in $L(\boldsymbol{\theta})$ are not independent.

Under $H_0 : \beta_{G_u} = 0$, Y is related to $\mathbf{W}$ only and is independent of G_u given $\mathbf{W}$. Assume that G_t is independent of Y given $\mathbf{W}$. (This assumption holds if G_t is independent of Y or is part of $\mathbf{W}$.) Then $\widehat{G}_u$, which is a function of G_t and $\widetilde{\boldsymbol{\pi}}$, is also independent of Y given $\mathbf{W}$, regardless of the value of $\widetilde{\boldsymbol{\pi}}$. In other words, the regression effects of $\widehat{G}_u$ and $\mathbf{W}$ on Y are the same as those of G_u and $\mathbf{W}$ under H_0. Denote the reference panel by R. Conditional on R, the imputed values are uncorrelated, so that, under H_0, the random vector $\mathbf{I}^{1/2}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to a multivariate normal distribution with mean zero and identity covariance matrix, where $\mathbf{I}(\boldsymbol{\theta}) = -\partial^2 \log L(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^2$. Because the limiting distribution does not depend on R, the convergence also holds unconditionally. Thus, standard likelihood methods can be used to test H_0 (even if the study sample and reference panel are drawn from different populations).

The above result hinges critically on the null hypothesis $H_0 : \beta_{G_u} = 0$ under the linear predictor $\alpha + \beta_{G_u} G_u + \boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}} \mathbf{W}$, which ensures that $\widehat{\boldsymbol{\theta}}$ converges to the true value of $\boldsymbol{\theta}$ conditional on R. If $\beta_{G_u} \neq 0$, then the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ conditional on R depends on R, so that the inverse information matrix $\mathbf{I}^{-1}(\widehat{\boldsymbol{\theta}})$, which ignores the variability in the reference panel, will underestimate the true variation of $\widehat{\boldsymbol{\theta}}$. Thus, the confidence intervals for β_{G_u} will not have proper coverage probabilities unless $\beta_{G_u} = 0$. It should also be pointed out that the validity of association testing is not guaranteed if the linear predictor does not take the form of $\alpha + \beta_{G_u} G_u + \boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}} \mathbf{W}$.

We now consider the analysis of case-control data under the logistic regression model

$$\Pr(Y = 1 | G_u, \mathbf{W}) = \frac{e^{\alpha + \beta_{G_u} G_u + \boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}} \mathbf{W}}}{1 + e^{\alpha + \beta_{G_u} G_u + \boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}} \mathbf{W}}}.$$

Write $\boldsymbol{\theta} = (\alpha, \beta_{G_u}, \boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}})^{\mathrm{T}}$. If G_u were observed on all study subjects, then the maximum likelihood estimator of $\boldsymbol{\theta}$ (based on the prospective likelihood) would converge to $\boldsymbol{\theta}^*$ and its covariance matrix would be consistently estimated by the inverse information matrix, where $\boldsymbol{\theta}^*$ is the same as $\boldsymbol{\theta}$ except that α is replaced by a different constant (Prentice and Pyke, 1979). Let $\widehat{\boldsymbol{\theta}}$ be the maximizer of the (prospective) likelihood based on the imputed data:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{e^{Y_i(\alpha + \beta_{G_u} \widehat{G}_{ui} + \boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}} \mathbf{W}_i)}}{1 + e^{\alpha + \beta_{G_u} \widehat{G}_{ui} + \boldsymbol{\beta}_{\mathbf{W}}^{\mathrm{T}} \mathbf{W}_i}}.$$

It then follows from the above arguments for cross-sectional studies that, under H_0 :

$\beta_{G_u} = 0$, the random vector $\mathbf{I}^{1/2}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ converges to a multivariate normal distribution with mean zero and identity covariance matrix, where $\mathbf{I}(\boldsymbol{\theta}) = -\partial^2 \log L(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^2$. Thus, the association testing is valid. Again, the variance is underestimated by the inverse information matrix if $\beta_{G_u} \neq 0$, and the association testing may not be valid for other types of hypotheses.

Chapter 4

A Likelihood-Based Framework for Association Analysis of Allele-Specific Copy Numbers

4.1 Introduction

In this chapter, we propose a framework for the integrated analysis of CNVs and SNPs in association studies, including analysis of total copy numbers as a special case. We allow for differential errors. We focus on case-control studies, although our methods can easily be modified for quantitative trait association analysis. We unify the ASCN calling and association analysis into a single step, so that the ASCN calling is informed by the phenotype and the association analysis fully accounts for the uncertainty in the calling. We formulate the effects of CNVs and SNPs on the phenotype through flexible regression models, which can accommodate various genetic mechanisms and gene-environment interactions. We construct appropriate likelihoods, which may involve high-dimensional parameters. We establish the consistency, asymptotic normality, and asymptotic efficiency of the maximum likelihood estimators by appealing to modern asymptotic techniques. We develop efficient and reliable numerical algorithms.

We conduct extensive simulation studies to assess the performance of the proposed methods and compare them with existing approaches. We illustrate our methods with a GWAS data of schizophrenia (Shi et al., 2009).

4.2 Methods

4.2.1 Notation and Model Assumptions

Suppose that the SNP has two alleles, A and B. Denote the total copy number and the B allele copy number by K and L, respectively, where $0 \leq L \leq K \leq S$, and S is the maximum copy number we will consider. Let Y be the phenotype of interest and $\mathbf{X}$ be a set of environmental factors. For case-control studies, the conditional density of $Y = y$ given $(K = k, L = l, \mathbf{X} = \mathbf{x})$ is formulated through the logistic regression model

$$P_{\alpha,\boldsymbol{\beta}}(y|k,l,\mathbf{x}) = \frac{\exp\{y(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{x}))\}}{1 + \exp\{\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{x})\}}, \tag{4.1}$$

where $\mathcal{Z}(k,l,\mathbf{x})$ is a design vector excluding the unit component. There is considerable flexibility in specifying the disease model. Suppose that there are no environmental factors. A linear predictor in the form of $\alpha + \beta k$ pertains to an additive effect of the total copy number and $\alpha + \beta_1 I(k = 1) + \ldots + \beta_S I(k = S)$ pertains to a saturated model. Replacing k with l in the above linear predictors leads to additive and saturated models of the B allele copy number. Combining k and l, we may specify $\alpha + \beta_1(k - l) + \beta_2 l$ with β_1 and β_2 corresponding to the A allele and B allele copy numbers, respectively, or $\alpha + \beta_1 k + \beta_2\{(k - l) - l\}$ with β_1 and β_2 corresponding to the total copy number and allelic difference, respectively.

Although we are interested in the effects of $(K, L, \mathbf{X})$ on Y, we only observe allele-specific intensity measurements on the Affymetrix platform and the transformed quantities on the Illumina platform, instead of (K, L). We denote the observed two-dimensional measurements by $\mathbf{R}$. Thus we have a regression problem with measurement errors. We describe below how to model the measurement error distribution $P(\mathbf{R}|Y, K, L, \mathbf{X})$. Note that by modelling the distribution conditional on Y and $\mathbf{X}$, we allow for the distribution to depend on the disease group and environmental factors such as indicators of batches. Specific formula of $P(\mathbf{R}|Y, K, L, \mathbf{X})$ depends on the platforms.

Affymetrix SNP Data

Each SNP on the Affymetrix platform generates a pair of intensity measurements $(R_{\mathrm{A}}, R_{\mathrm{B}})$ for the A and B alleles, respectively. The two measurements are log2-transformed values from the normalized raw intensities $(\widetilde{R}_{\mathrm{A}}, \widetilde{R}_{\mathrm{B}})$. We assume that $(R_{\mathrm{A}}, R_{\mathrm{B}})$ given $(Y, K, L, \mathbf{X})$ follows a bivariate Gaussian distribution:

$$\begin{aligned} & P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R_{\mathrm{A}}, R_{\mathrm{B}}|Y, K, L, \mathbf{X}) \\ & \quad = \phi\left\{ \begin{bmatrix} R_{\mathrm{A}} \\ R_{\mathrm{B}} \end{bmatrix} ; \begin{bmatrix} \boldsymbol{\gamma}_{\mathrm{A}}^{\mathrm{T}}\mathcal{A} \\ \boldsymbol{\gamma}_{\mathrm{B}}^{\mathrm{T}}\mathcal{B} \end{bmatrix} , \begin{bmatrix} g(\boldsymbol{\delta}_{\mathrm{A}}^{\mathrm{T}}\mathcal{C}) & \boldsymbol{\delta}_{\rho}^{\mathrm{T}}\mathcal{W}\sqrt{g(\boldsymbol{\delta}_{\mathrm{A}}^{\mathrm{T}}\mathcal{C})g(\boldsymbol{\delta}_{\mathrm{B}}^{\mathrm{T}}\mathcal{D})} \\ \boldsymbol{\delta}_{\rho}^{\mathrm{T}}\mathcal{W}\sqrt{g(\boldsymbol{\delta}_{\mathrm{A}}^{\mathrm{T}}\mathcal{C})g(\boldsymbol{\delta}_{\mathrm{B}}^{\mathrm{T}}\mathcal{D})} & g(\boldsymbol{\delta}_{\mathrm{B}}^{\mathrm{T}}\mathcal{D}) \end{bmatrix} \right\}, \end{aligned}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{\mathrm{A}}, \boldsymbol{\gamma}_{\mathrm{B}})$, $\boldsymbol{\delta} = (\boldsymbol{\delta}_{\mathrm{A}}, \boldsymbol{\delta}_{\mathrm{B}}, \boldsymbol{\delta}_{\rho})$, $\phi(\mathbf{r}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the bivariate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{W}$ stand for $\mathcal{A}(y, k, l, \mathbf{x})$, $\mathcal{B}(y, k, l, \mathbf{x})$, $\mathcal{C}(y, k, l, \mathbf{x})$, $\mathcal{D}(y, k, l, \mathbf{x})$ and $\mathcal{W}(y, k, l, \mathbf{x})$, respectively, which are the design vectors for the means, variances and the correlation coefficient, and $g(.)$ is a link function, such as the exponential function, that constraints the variances to be non-negative. We may utilize a saturated model for the dependence of $(R_{\mathrm{A}}, R_{\mathrm{B}})$ on $(Y, K, L, \mathbf{X})$, so each bivariate Gaussian distribution of $(R_{\text{A}}, R_{\text{B}})$ given $(Y, K, L, \mathbf{X})$ are completely determined by five parameters consisting of two means, two variances and a correlation coefficient.

Illumina SNP Data

On the Illumina platform, we obtain the measurements on so-called Log R ratio and B allele frequency $(R_{\text{LRR}}, R_{\text{BAF}})$, which are transformed values of the raw allele-specific intensities $(\widetilde{R}_{\text{A}}, \widetilde{R}_{\text{B}})$. Let $\widetilde{R}_{\text{T}} = \widetilde{R}_{\text{A}} + \widetilde{R}_{\text{B}}$ and $\eta = \arctan(\widetilde{R}_{\text{A}}/\widetilde{R}_{\text{B}})/(\pi/2)$ so that $\widetilde{R}_{\text{T}}$ measures the total copy number and η measures the allelic contrast. Then, R_{LRR}, a normalized measure of the total signal intensity for each SNP, is calculated as $R_{\text{LRR}} = \log_2(\widetilde{R}_{\text{T,observed}}/\widetilde{R}_{\text{T,expected}})$, where $\widetilde{R}_{\text{T,expected}}$ is computed from a linear interpolation of the canonical genotype clusters corresponding to AA, AB and BB. A normalized measure of η is calculated as

$$R_{\text{BAF}} = \begin{cases} 0 & \text{if } \eta < \eta_{AA}, \\ 0.5(\eta - \eta_{AA})/(\eta_{AB} - \eta_{AA}) & \text{if } \eta_{AA} \le \eta < \eta_{AB}, \\ 0.5 + 0.5(\eta - \eta_{AB})/(\eta_{BB} - \eta_{AB}) & \text{if } \eta_{AB} \le \eta < \eta_{BB}, \\ 1 & \text{if } \eta \ge \eta_{BB}, \end{cases}$$

where η_{AA}, η_{AB}, and η_{BB} are the η values for the three canonical genotype clusters generated from a large set of reference samples. As a result, R_{BAF} should be around 0, 0.5 and 1 for genotype AA, AB and BB, respectively. If the R_{BAF} value of a SNP deviates from the three values, it may indicate CNV. For instance, a R_{BAF} value 0.33 may indicate a genotype of AAB.

By their definitions, R_{LRR} and R_{BAF} can be treated as independent given $(Y, K, L, \mathbf{X})$.

We model the conditional distribution of $R_{\mathrm{LRR}}|(Y,K,L,\mathbf{X})$ by a normal density

$$P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R_{\mathrm{LRR}}|Y,K,L,\mathbf{X}) = \phi(R_{\mathrm{LRR}}; \boldsymbol{\gamma}_{\mathrm{LRR}}^{\mathrm{T}}\mathcal{A}(Y,K,\mathbf{X}), g(\boldsymbol{\delta}_{\mathrm{LRR}}^{\mathrm{T}}\mathcal{C}(Y,K,\mathbf{X}))),$$

where $\phi(r;\mu,\sigma^2)$ is the univariate normal density function with mean μ and variance σ^2, $\mathcal{A}(y,k,\mathbf{x})$ and $\mathcal{C}(y,k,\mathbf{x})$ are the design vectors for the mean and variance, respectively, and $g(.)$ is the link function for the variance. Note that the design vectors do not depend on L.

We model the distribution of R_{BAF} given $(Y,K,L,\mathbf{X})$ by a (truncated) normal density,

$$\begin{aligned} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R_{\mathrm{BAF}}|Y,K,L,\mathbf{X}) &= \phi(R_{\mathrm{BAF}}; \boldsymbol{\gamma}_{\mathrm{BAF}}^{\mathrm{T}}\mathcal{B}, g(\boldsymbol{\delta}_{\mathrm{BAF}}^{\mathrm{T}}\mathcal{D}))^{I(0<R_{\mathrm{BAF}}<1)} \\ &\times \Phi(0; \boldsymbol{\gamma}_{\mathrm{BAF}}^{\mathrm{T}}\mathcal{B}, g(\boldsymbol{\delta}_{\mathrm{BAF}}^{\mathrm{T}}\mathcal{D}))^{I(R_{\mathrm{BAF}}=0)} \\ &\times \Big(1-\Phi(1; \boldsymbol{\gamma}_{\mathrm{BAF}}^{\mathrm{T}}\mathcal{B}, g(\boldsymbol{\delta}_{\mathrm{BAF}}^{\mathrm{T}}\mathcal{D}))\Big)^{I(R_{\mathrm{BAF}}=1)}, \end{aligned}$$

where $\Phi(r;\mu,\sigma^2)$ is the cumulative distribution function corresponding to $\phi(r;\mu,\sigma^2)$, and $\mathcal{B}$, $\mathcal{D}$ stand for $\mathcal{B}(y,k,l,\mathbf{x})$, $\mathcal{D}(y,k,l,\mathbf{x})$, respectively, which are the design vectors for the mean and variance. Note that $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{\mathrm{LRR}}, \boldsymbol{\gamma}_{\mathrm{BAF}})$ and $\boldsymbol{\delta} = (\boldsymbol{\delta}_{\mathrm{LRR}}, \boldsymbol{\delta}_{\mathrm{BAF}})$. When $K = 0$ indicating deletion of both copies, we assume that the mean of R_{BAF} to be a constant 0.5 and the variance is smaller than 0.15 so that the probability of truncation is smaller than 0.001. Therefore as an approximation, we can estimate the variance as if $P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R_{\mathrm{BAF}}|Y,K=0,L,\mathbf{X})$ is non-truncated normal. When $K>0$, we assume the means to be 0.0 and 1.0 for the two homozygous genotypes, so we only need to estimate the variances, which is straightforward because the truncation points are exactly the mean values. Specifically, we simply use the observed R_{BAF} such that $0 < R_{\mathrm{BAF}} < 1$ to estimate the variance. For all the other normal components (which correspond to heterozygous genotypes), we assume that the mean values are far away from the boundary (0 or 1) so that the truncation effects can be neglected.

Write $\mathbf{R} = (R_{\mathrm{A}}, R_{\mathrm{B}})$ for Affymetrix data and $\mathbf{R} = (R_{\mathrm{LRR}}, R_{\mathrm{BAF}})$ for Illumina data. In addition, write $\xi_{k,l} = P(K = k, L = l)$ and $\boldsymbol{\xi} = (\xi_{0,0}, \xi_{1,0}, \xi_{1,1} \ldots, \xi_{S,S})$. We suppose that $\xi_{k,l} > 0$ for all (k, l). In some applications, $\mathbf{X}$ and (K, L) are correlated. One important example is when $\mathbf{X}$ represents the principal components for ancestry. Also, certain genes may influence both environmental exposure and disease occurrence, as is the case in a lung cancer study involving a gene and a smoking variable (Amos et al., 2008). In such cases, we allow gene-environment dependence by leaving the distributions of $\mathbf{X}$ given (K, L), denoted by $F_{k,l}(.)$, completely unspecified. Because of the case-control sampling, we adopt the retrospective likelihood

$$L_{\mathrm{r}}(\boldsymbol{\theta}, \{F_{k,l}\}) = \prod_{i=1}^{n} \left\{ \sum_{k=0}^{S} \sum_{l=0}^{k} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i | Y_i, k, l, \mathbf{X}_i) \frac{P_{\alpha,\boldsymbol{\beta}}(Y_i | k, l, \mathbf{X}_i) f_{k,l}(\mathbf{X}_i) \xi_{k,l}}{\sum_{k'} \sum_{l'} \int_{\mathbf{x}} P_{\alpha,\boldsymbol{\beta}}(Y_i | k', l', \mathbf{x}) \xi_{k',l'} \mathrm{d}F_{k',l'}(\mathbf{x})} \right\},$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\delta})$, and n is the number of study subjects. We can see that the observed data $(\mathbf{R}_i, Y_i, \mathbf{X}_i)$ for subject i is modeled by a mixture of bivariate-Gaussian clusters. The intensity measurements are used to infer the cluster membership, separately within cases and controls at each level of $\mathbf{X}$. The inferred frequencies of each cluster are compared between cases and controls, and differences of the frequencies are attributed to the disease model as association.

Because the distribution of the covariates $(K, L, \mathbf{X})$ is completely unspecified, the retrospective maximum likelihood estimate of $\boldsymbol{\beta}$ can be obtained by maximizing the prospective likelihood,

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \{F_{k,l}\}) = \prod_{i=1}^{n} \left\{ \sum_{k=0}^{S} \sum_{l=0}^{k} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i | Y_i, k, l, \mathbf{X}_i) P_{\alpha,\boldsymbol{\beta}}(Y_i | k, l, \mathbf{X}_i) f_{k,l}(\mathbf{X}_i) \xi_{k,l} \right\}; \qquad (4.2)$$

see Roeder et al. (1996) for justification for such equivalence. We use the NPMLE approach. In this approach, the distribution functions $\{F_{k,l}\}$ are treated as right-continuous functions with jumps at the observed $\mathbf{X}$. The objective function to be maximized is obtained from $L_{\mathrm{p}}(\boldsymbol{\theta}, \{F_{k,l}\})$ by replacing $f_{k,l}(\mathbf{x})$ with the jump size of $F_{k,l}$ at $\mathbf{x}$. The maximization can be carried out by the EM algorithm described in the Appendix.

In many applications, it is appropriate to assume gene-environment independence, so that $\{F_{k,l}\}$ reduces to a single distribution function F. In addition, by the HWE assumption, $L = l$ given $K = k$ follows a binomial distribution with parameters k and p_{B}, where p_{B} is the population frequency of the B allele. We denote the binomial distribution by $P_{k,p_{\mathrm{B}}}(l)$ and denote $\pi_k = P(K = k)$, so that $\xi_{k,l} = P_{k,p_{\mathrm{B}}}(l)\pi_k$. Let $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_S)$. When we impose such structures in the covariate distribution, the equivalence between the retrospective and prospective likelihoods no longer holds and the retrospective likelihood should be used.

There is very little information about α in the retrospective likelihood, so the problem is virtually nonidentifiable. One possible solution is to assume that the disease is rare, which is generally true in case-control studies. Then we have the following approximation to (4.1):

$$P_{\alpha,\boldsymbol{\beta}}(y|k,l,\mathbf{x}) \approx \exp\{y(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{x}))\}.$$

Write $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\pi}, p_{\mathrm{B}}, \boldsymbol{\gamma}, \boldsymbol{\delta})$. The retrospective likelihood can then be approximated by

$$\begin{aligned}
&\widetilde{L}_{\mathrm{r}}(\boldsymbol{\theta}, F) \\
&= \prod_{i=1}^{n} \left\{ \sum_{k=0}^{S} \sum_{l=0}^{k} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i|Y_i,k,l,\mathbf{X}_i) \frac{\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{X}_i)\}P_{k,p_{\mathrm{B}}}(l)\pi_k f(X_i)}{\int_{\mathbf{x}} \sum_{k'} \sum_{l'} \exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k',l',\mathbf{x})\}P_{k',p_{\mathrm{B}}}(l')\pi_{k'}\mathrm{d}F(\mathbf{x})} \right\}^{I(Y_i=1)} \\
&\qquad \times \left\{ \sum_{k=0}^{S} \sum_{l=0}^{k} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i|Y_i,k,l,\mathbf{X}_i)P_{k,p_{\mathrm{B}}}(l)\pi_k f(X_i) \right\}^{I(Y_i=0)}. \quad (4.3)
\end{aligned}$$

We see that α is dropped from the likelihood. We again adopt the NPMLE approach, which is implemented via the EM algorithm described in the Appendix.

4.2.2 Total Copy Number Measurement

In some cases, such as the copy number probes of Affymetrix 6.0 array and the array comparative genomic hybridization (CGH), the observed data only contain measurements of the total copy number. Let R_i be the one-dimensional measurement and $P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R|Y,k,\mathbf{X})$ be a univariate normal density, which is the same as $P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R_{\mathrm{LRR}}|Y,k,l,\mathbf{X})$. We can easily accommodate such cases by reducing (4.2) and (4.3) to

$$L_{\mathrm{p}}(\boldsymbol{\theta},\{F_k\}) = \prod_{i=1}^{n}\left\{\sum_{k=0}^{S} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R_i|Y_i,k,\mathbf{X}_i)P_{\alpha,\boldsymbol{\beta}}(Y_i|k,\mathbf{X}_i)f_k(\mathbf{X}_i)\pi_k\right\}, \tag{4.4}$$

where $\boldsymbol{\theta} = (\alpha,\boldsymbol{\beta},\boldsymbol{\pi},\boldsymbol{\gamma},\boldsymbol{\delta})$, and

$$\begin{aligned}\widetilde{L}_{\mathrm{r}}(\boldsymbol{\theta},F) = \prod_{i=1}^{n}&\left\{\sum_{k=0}^{S} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R_i|Y_i,k,\mathbf{X}_i)\frac{\exp\{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k,\mathbf{X}_i)\}\pi_k f(X_i)}{\int_{\mathbf{x}}\sum_{k'}\exp\{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k',\mathbf{X}_i)\}\pi_{k'}\mathrm{d}F(\mathbf{x})}\right\}^{I(Y_i=1)}\\ &\times\left\{\sum_{k=0}^{S} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R_i|Y_i,k,\mathbf{X}_i)\pi_k f(X_i)\right\}^{I(Y_i=0)},\end{aligned} \tag{4.5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\pi},\boldsymbol{\gamma},\boldsymbol{\delta})$, respectively. Barnes et al. (2008) dealt with this problem by adopting a simpler prospective likelihood

$$\prod_{i=1}^{n}\left\{\sum_{k=0}^{S} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R_i|Y_i,k,\mathbf{X}_i)P_{\alpha,\boldsymbol{\beta}}(Y_i|k,\mathbf{X}_i)P_{\boldsymbol{\zeta}}(k|\mathbf{X}_i)\right\}, \tag{4.6}$$

where $P_{\boldsymbol{\zeta}}(k|\mathbf{x})$ is the multinomial regression model of $K = k$ given $\mathbf{X} = \mathbf{x}$. When there are no environmental factors, (4.6) is exactly the same as (4.4). Thus, Barnes' method is justified. In the presence of environmental factors, Barnes et al. (2008) decomposed the joint density function $P(K,\mathbf{X})$ as $P(K|\mathbf{X})P(\mathbf{X})$ and imposed a parametric structure on $P(K|\mathbf{X})$. With the parametric restriction, the use of prospective likelihood is no longer appropriate.

4.3 Simulation Studies

We conducted extensive simulation studies to evaluate the performance of the proposed and existing methods in realistic settings. We generated data with the pattern observed at one SNP site in the GWAS data of schizophrenia (Shi et al., 2009); see Figure 4.1(a). Specifically, we assumed that in the general population the total copy number state K takes values 0, 1, and 2 with probabilities 0.06, 0.33 and 0.61. Given K, the B allele copy number L follows the binomial distribution with probability 0.65. We simulated the disease status from the logistic regression model $\text{logit}P(Y = 1|K, L) = \alpha + \boldsymbol{\beta}^{\text{T}}\mathcal{Z}(K, L)$, where $\alpha = -4.6$ yielding disease rate about 1% and $\mathcal{Z}(K, L)$ is a specific function of K and L. We repeated the above processes until we obtained 1,000 controls and 1,000 cases to form the case-control samples. For controls, the intensity measurements given each (K, L) cluster were normal with the observed means and variances. The distribution of intensity measurements for cases were allowed to have different means or variances as compared to controls. We obtained 5,000 replicates of the dataset.

Our first set of simulation studies was designed to explore the sensitivity of the type I error to the differences in the cluster mean and variance between cases and controls. We simulated the disease status from the logistic regression model with an additive effect of the B allele copy number:

$$\text{logit}P(Y = 1|K, L) = \alpha + \beta_0 L. \tag{4.7}$$

We applied our method as well as two alternative imputation methods that mimic the existing calling algorithms. Both imputation methods use a two-dimensional GMM to assign each individual to the most likely ASCN cluster. While one fits the GMM with cases and controls combined (imputation-C), the other one with cases and controls separated (imputation-S). As shown in Figure 4.2, imputation-C is robust to differential variances between cases and controls in that it does not generate differential misclassification. However, this approach breaks down in the presence of location shifts. Imputation-S is not affected by differential errors in either means or variances as cases and controls are modeled separately. However, inflation of type I error remains in all cases. The inflation results from ignoring the uncertainty in cluster assignment and from over-estimating the differences in cluster frequencies between cases and controls, as nuisance parameters are allowed to vary between cases and controls. As a result, the true variance of the estimator of β_0 is greater than the naive variance estimator. Figure 4.2(b) shows that the inflation of type I error for imputation-S grows as the overall variation increases and is independent of differential variances. By contrast, the proposed method provides the most robust test by modelling cases and controls separately and accounting for all uncertainties.

In the second set of simulation studies, we investigated the pitfalls of the imputation approach more closely. We assumed no differences in cluster means and variances between cases and controls in order to separate the influence of imputation itself from that of differential errors; the rest of the simulation set-up was the same as the first set of simulation studies. Table 4.1 shows that the coefficient estimator of imputation-C is biased towards the null, which is due to ignoring the phenotype when inferring the ASCN states so that the imputed ASCN states are more homogeneous between cases and controls than they actually are. As a result, the power of imputation-C is diminished compared to the proposed method. Imputation-C yields correct variance estimator under the null and thus has proper control of type I error; see Hu and Lin

(2010) for a proof in the context of SNP association analysis. As expected, imputation-S is not subject to bias. However, Table 4.1 exhibits large discrepancies between the true variances of $\widehat{\beta}_0$ and the naive estimators. The discrepancy remains under the null, leading to inflated type I error. As a result, imputation-S can sometimes be more powerful than the proposed method.

When only the effect of the total copy number is of interest, Barnes' method can also be used to account for differential errors. In this case, our method still relies on the two-dimensional measurements for ASCN while Barnes' method relies on the summed measurements for the total copy number. We compared the two methods in the third set of simulation studies. We generated the disease status from the logistic regression model

$$\text{logit}P(Y = 1|K) = \alpha + \beta_0 K. \tag{4.8}$$

The results are summarized in Table 4.2. Both methods yield unbiased estimators for β_0 and correct variance estimators and consequently correct type I error. However, the proposed method gains power by exploiting more information; see Figure 4.3.

Our previous simulation studies are based on ASCN intensity data at SNP sites. It is also of interest to compare the proposed and Barnes' methods when only total copy number measurements are obtained. Since we theoretically proved that the two methods are equivalent in the absence of environmental factors, we focused on testing the gene-environment interaction on the disease in the forth set of studies. We adopted the disease model

$$\text{logit}P(Y = 1|K, X) = \alpha + \beta_1 K + \beta_2 X + \beta_3 KX,$$

where X follows the standard normal distribution and is independent of K. We generated cluster means and variances of the one-dimensional measurements mimicking those of CN_615718 in the schizophrenia data; see Figure 4.1(b). As shown in Figure 4.4, the proposed method based on the retrospective likelihood (4.3) is substantially more powerful than Barnes' method. The power gain of the propose method is largely attributed to its incorporation of gene-environment independence.

4.4 Schizophrenia Data

Schizophrenia is a severe psychiatric disorder marked by hallucinations, delusions, cognitive deficits and apathy, with a lifetime prevalence of 0.4-1%. Schizophrenia has high heritability ($\sim$ 80%) and genetically heterogeneous. Recent studies implicated that common SNPs and rare, large CNVs are associated with schizophrenia, but the joint effects of CNVs and SNPs have not been investigated. We applied our methods for integrated analysis of CNVs and SNPs using the data from the Molecular Genetics of Schizophrenia (MGS) GWAS (Shi et al., 2009). MGS GWAS collected self-reported European ancestry (EA) and African American (AA) unrelated adult cases with DSM-IV schizophrenia from ten sites in the United States and Australia, and recruited EA and AA unrelated adult controls through Knowledge Networks by phone calls. Part of the MGS GWAS was genotyped with the Affymetrix 6.0 platform at the Broad Institute, under the auspices of the Genetic Association Information Network (GAIN) and was referred to as the GAIN samples. The GAIN samples consist of both EA and AA subjects. Our data is the EA portion, including 1172 cases and 1378 controls. The different collection processes of cases and controls imply the possibility of differential errors. Indeed, when we treated all controls as if they were from the 11th site, the principal components (PCs) calculated from the raw intensity data were correlated with many of these eleven site indicators.

For intensity data at SNP sites, our method regresses the disease status on both the total copy number and the difference of the allelic copy number

$$\text{logit}P(Y=1|K,L)=\alpha+\beta_1 K+\beta_2\{(K-L)-L\}.$$

A two-degrees-of-freedom test of the null hypothesis $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$ provides a combined test of CNV and allelic effects; the null hypothesis $\mathrm{H}_0 : \beta_1 = 0$ gives a test of the CNV effect controlling for allelic variation; the null hypothesis $\mathrm{H}_0 : \beta_2 = 0$ gives a test of the allelic effect controlling for copy number variation.

The testing results of the intensity data displayed in Figure 1.1 are shown in Table 4.3. As Figure 1.1 suggests serious differential errors but no appreciable differences of cluster frequencies between cases and controls, the proposed method yielded non-significant p-values for all three tests. As expected, imputation-C is sensitive to differential means. Imputation-S is sensitive to cluster variances, especially when the variances are large so that the clusters are not well-separated.

On the other hand, the intensity data at the SNP site displayed in Figure 4.5 show little differential errors but a sign of true association. Consistent with the second set of simulation results, the proposed method tends to generate a smaller p-value than both imputation methods (Table 4.4).

For intensity data at copy number probes, we fitted the saturated model

$$\text{logit}P(Y=1|K)=\alpha+\beta_1 I(K=0)+\beta_2 I(K=1).$$

A two-degrees-of-freedom test of the null hypothesis $\mathrm{H}_0 : \beta_1 = \beta_2 = 0$ provides an overall test of CNV effects. As a comparison, we also included an imputation approach with CNVs called by PennCNV. Note that PennCNV assumes the intensity means and variances given a CNV state to be constant, regardless of the disease status. The intensity data and testing results are presented in Figure 4.6 and Table 4.5, respectively.

Again, the proposed method is robust to differential errors and more powerful in the presence of true association as compared to the imputation method.

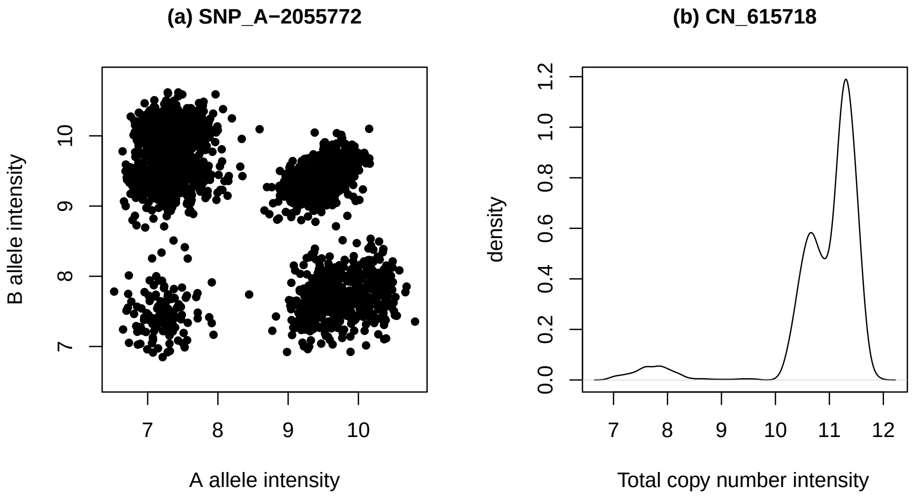

Figure 4.1: Observed intensity measurements in the schizophrenia data. (a) Intensity measurements at the SNP site "SNP_A-2055772". (b) Intensity measurements at the copy number probe "CN_615718".

Table 4.1: Simulation results for studying the effect of the B allele copy number when there are no differential errors

	Proposed					Imputation-C					Imputation-S				
β_0	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
.00	.000	.064	.064	.993	.007	.001	.062	.063	.992	.008	.000	.072	.063	.975	.025
.14	.000	.063	.064	.992	.346	-.004	.062	.063	.992	.335	.000	.072	.063	.976	.384
.18	.001	.065	.064	.989	.599	-.005	.064	.063	.989	.590	.001	.073	.063	.973	.610
.22	.001	.064	.064	.991	.806	-.005	.062	.063	.991	.797	.002	.073	.063	.975	.793
.26	.001	.065	.064	.988	.929	-.007	.063	.063	.988	.925	.000	.073	.063	.976	.912
.30	.001	.065	.064	.990	.981	-.008	.063	.063	.988	.980	.002	.073	.063	.974	.975
.40	.000	.065	.065	.990	.999	-.013	.065	.063	.985	1.00	.000	.075	.064	.972	.999
.80	-.002	.066	.066	.991	1.00	-.027	.067	.067	.983	1.00	.002	.079	.067	.973	1.00

NOTE: Bias and SE are the bias and standard error of the parameter estimator. SEE is the mean of the standard error estimator. CP is the coverage probability of the 99% confidence interval. Power is the type I error/power for testing $H_0 : \beta_0 = 0$ at the .01 nominal significance level.

Table 4.2: Simulation results for studying the effect of the total copy number

	Proposed					Barnes et al.				
β_0	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
0.0	-0.002	0.087	0.087	0.991	0.009	-0.001	0.089	0.090	0.989	0.011
0.1	0.000	0.090	0.089	0.990	0.069	0.000	0.091	0.092	0.990	0.065
0.2	0.000	0.091	0.091	0.991	0.351	0.000	0.093	0.094	0.991	0.324
0.3	-0.001	0.095	0.094	0.991	0.738	0.000	0.097	0.096	0.990	0.719
0.4	-0.001	0.096	0.097	0.992	0.945	0.000	0.098	0.099	0.992	0.934

NOTE: see the Note to Table 4.1.

Table 4.3: P-values of hypothesis tests at the SNP site showing differential errors

	$\beta_1 = \beta_2 = 0$	$\beta_1 = 0$	$\beta_2 = 0$
Proposed	9.47e-01	7.51e-01	9.41e-01
Imputation-C	9.35e-12	2.96e-10	1.08e-02
Imputation-S	2.39e-41	4.04e-34	2.88e-07

Table 4.4: P-values of hypothesis tests at the SNP site showing true association

	$\beta_1 = \beta_2 = 0$	$\beta_1 = 0$	$\beta_2 = 0$
Proposed	1.65e-02	9.00e-01	4.65e-03
Imputation-C	2.53e-02	1.51e-01	1.47e-02
Imputation-S	9.91e-02	8.24e-01	4.02e-02

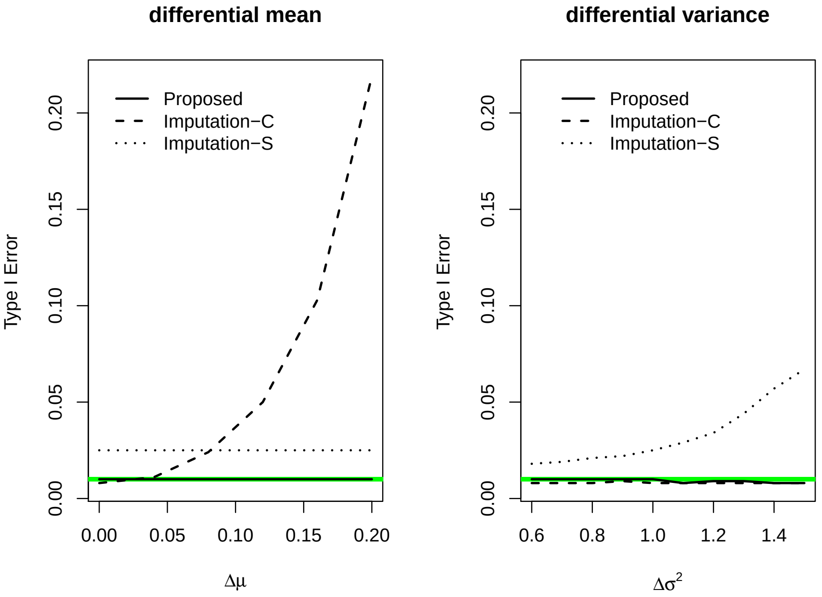

Figure 4.2: Type I error for testing the null effect of the B allele copy number at the 1% nominal significance level. (a) Type I error is estimated for association methods at different values of differential shift of means. We let $\Delta\mu$ to be the uniform difference of all cluster means between cases and controls. (b) Type I error is shown for association methods at different values of differential shift in variances. The cluster variances of cases are inflated against those of controls by a factor of $\Delta\sigma^2$. Note that when $\Delta\sigma^2 < 1.0$, cases have deflated cluster variances compared to controls. The green line indicates the nominal level of 1%.

Table 4.5: P-values of hypothesis tests at the copy number probes

	CN_710839	CN_1197999
Proposed	9.80e-01	1.67e-02
Imputation	8.70e-03	3.74e-01

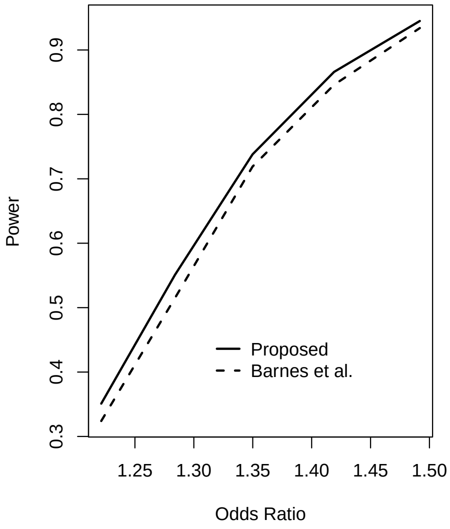

Figure 4.3: Power for testing the effect of the total copy number.

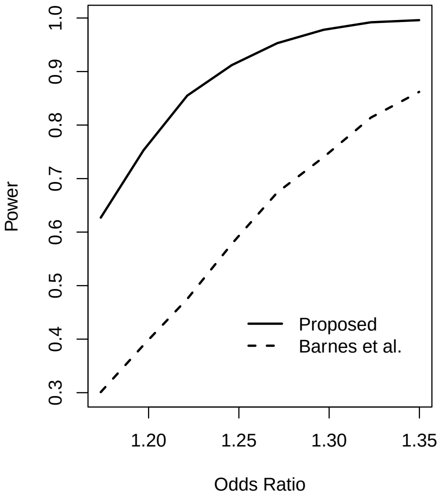

Figure 4.4: Power of testing the gene-environment interaction at the 1% nominal significant level.

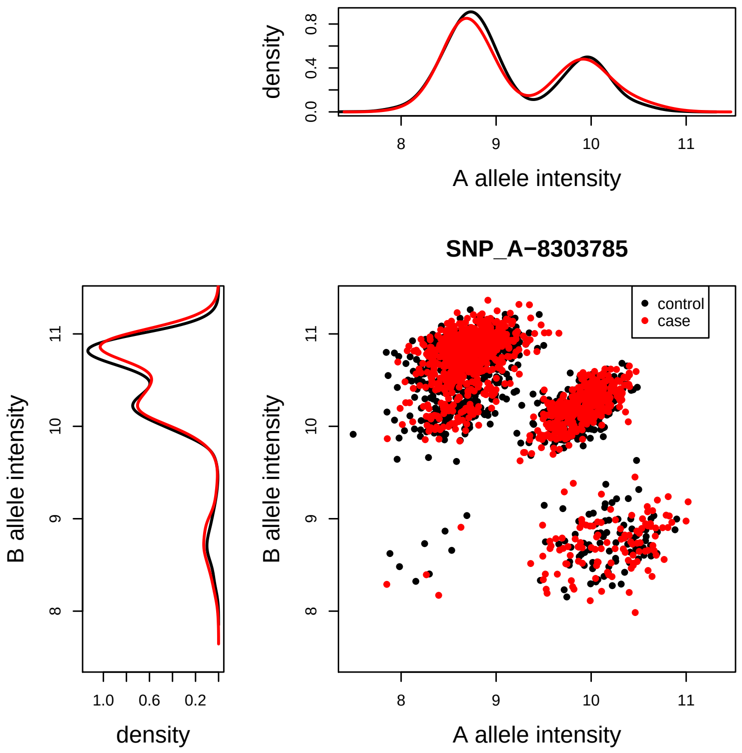

Figure 4.5: Observed intensity measurements at the SNP site “SNP_A-8303785”.

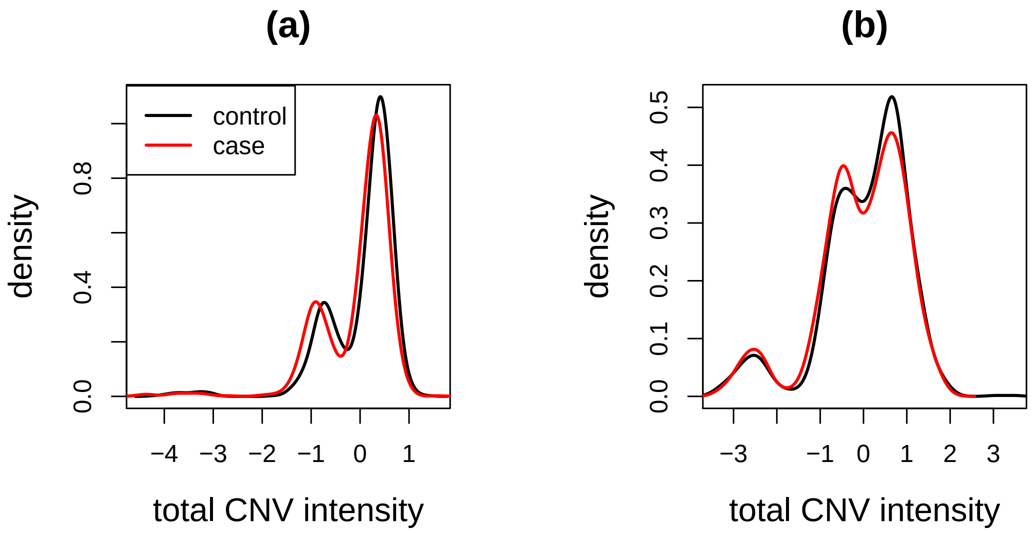

Figure 4.6: PC-summarized intensity data at copy number probes “CN_710839” (left panel) and “CN_1197999” (right panel).

4.5 Appendix

4.5.1 Theoretical Properties

NPMLE of Likelihood (4.2)

We first impose the following identifiability conditions and verity the identifiability of the parameters in Lemma 4.1.

CONDITION 4.1. If $\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(K, L, \mathbf{X}) = \widetilde{\alpha} + \widetilde{\boldsymbol{\beta}}^{\mathrm{T}} \mathcal{Z}(K, L, \mathbf{X})$ for any $(K, L, \mathbf{X})$, then $\alpha = \widetilde{\alpha}$ and $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$.

CONDITION 4.2. If $P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}|Y, K, L, \mathbf{X}) = P_{\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\delta}}}(\mathbf{R}|Y, K, L, \mathbf{X})$ for any $(\mathbf{R}, Y, K, L, \mathbf{X})$, then $\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}}$ and $\boldsymbol{\delta} = \widetilde{\boldsymbol{\delta}}$.

LEMMA 4.1. If two sets of parameters $(\boldsymbol{\theta}, \{F_{k,l}\})$ and $(\widetilde{\boldsymbol{\theta}}, \{\widetilde{F}_{k,l}\})$ yield the same likelihood, then $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$ and $F_{k,l} = \widetilde{F}_{k,l}$ for all (k, l).

Proof: Suppose that

$$\sum_{k=0}^{S} \sum_{l=0}^{k} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}|Y, k, l, \mathbf{X}) P_{\alpha,\boldsymbol{\beta}}(Y|k, l, \mathbf{X}) f_{k,l}(\mathbf{X}) \xi_{k,l}$$
$$= \sum_{k=0}^{S} \sum_{l=0}^{k} P_{\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\delta}}}(\mathbf{R}|Y, k, l, \mathbf{X}) P_{\widetilde{\alpha},\widetilde{\boldsymbol{\beta}}}(Y|k, l, \mathbf{X}) \widetilde{f}_{k,l}(\mathbf{X}) \widetilde{\xi}_{k,l}.$$

By Proposition 1 of Teicher (1963) that all finite mixtures of normal distributions is identifiable and Condition 4.2, $\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}}$, $\boldsymbol{\delta} = \widetilde{\boldsymbol{\delta}}$, and for all (k, l),

$$P_{\alpha,\boldsymbol{\beta}}(Y|k, l, \mathbf{X}) f_{k,l}(\mathbf{X}) \xi_{k,l} = P_{\widetilde{\alpha},\widetilde{\boldsymbol{\beta}}}(Y|k, l, \mathbf{X}) \widetilde{f}_{k,l}(\mathbf{X}) \widetilde{\xi}_{k,l}.$$

Summarizing over $Y = 0, 1$ yields $f_{k,l}(\mathbf{X}) \xi_{k,l} = \widetilde{f}_{k,l}(\mathbf{X}) \widetilde{\xi}_{k,l}$. In addition, integrating over $\mathbf{X}$ gives $\xi_{k,l} = \widetilde{\xi}_{k,l}$ and then $f_{k,l}(.) = \widetilde{f}_{k,l}(.)$ for any (k, l). Thus, $P_{\alpha,\boldsymbol{\beta}}(Y|k, l, \mathbf{X}) =$

$P_{\widetilde{\alpha},\widetilde{\boldsymbol{\beta}}}(Y|k,l,\mathbf{X})$. It then follows from that $P_{\alpha,\boldsymbol{\beta}}(Y|k,l,\mathbf{X})$ is logistic regression model and from Condition 4.1, $\alpha = \widetilde{\alpha}$ and $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$.

We then verify that the information matrices along all non-trivial parametric submodels are non-singular in LEMMA 4.2.

CONDITION 4.3. If there exists a constant $\boldsymbol{\mu}$ such that $\boldsymbol{\mu}^{\mathrm{T}}\nabla_{\alpha,\boldsymbol{\beta}} \log P_{\alpha,\boldsymbol{\beta}}(Y|k,l,\mathbf{X}) = 0$ for any (k,l), then $\boldsymbol{\mu} = 0$.

CONDITION 4.4. For any (k,l), the function $f_{k,l}$ is positive its support and continuously differentiable.

LEMMA 4.2. If there exist a vector $\boldsymbol{\mu}_{\boldsymbol{\theta}} = (\mu_\alpha, \boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\mu}_{\boldsymbol{\xi}}, \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\mu}_{\boldsymbol{\delta}})$ and functions $\psi_{k,l}(\mathbf{x})$ with $E[\psi_{k,l}(\mathbf{X})] = 0$ such that

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, F_0) + \sum_{k,l} l_{F_{k,l}}(\boldsymbol{\theta}_0, F_{k,l,0})[\int \psi_{k,l}\ dF_{k,l,0}] = 0, \tag{4.9}$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$, and $l_{F_{k,l}}[\int \psi_{k,l}\ dF_{k,l,0}]$ is the score function for $F_{k,l}$ along the submodel $F_{k,l,0} + \epsilon \int \psi_{k,l}\ dF_{k,l,0}$, then $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{0}$ and $\psi_{k,l} = 0$ for any (k,l).

Proof: For ease of exposition, we derive the proof based on the likelihood (4.4) with univariate normal densities. A similar equation to (4.9) is expanded as

$$\sum_k P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(R|Y,k,\mathbf{X}) \Bigg\{ P_{\alpha,\boldsymbol{\beta}}(Y|k,\mathbf{X}) f_k(\mathbf{X}) \pi_k \left[\frac{(R - \boldsymbol{\gamma}^{\mathrm{T}}\mathcal{A})}{\exp \boldsymbol{\delta}^{\mathrm{T}}\mathcal{C}} \boldsymbol{\mu}_{\gamma}^{\mathrm{T}} \mathcal{A} - \frac{1}{2}(1 - \frac{(R - \boldsymbol{\gamma}^{\mathrm{T}}\mathcal{A})^2}{\exp \boldsymbol{\delta}^{\mathrm{T}}\mathcal{C}}) \boldsymbol{\mu}_{\boldsymbol{\delta}}^{\mathrm{T}} \mathcal{C} \right]$$
$$+ \boldsymbol{\mu}_{\alpha,\boldsymbol{\beta}}^{\mathrm{T}} \nabla_{\alpha,\boldsymbol{\beta}} P_{\alpha,\boldsymbol{\beta}}(Y|k,\mathbf{X}) f_k(\mathbf{X}) \pi_k + P_{\alpha,\boldsymbol{\beta}}(Y|k,\mathbf{X}) f_k(\mathbf{X}) \boldsymbol{\mu}_{\boldsymbol{\pi}}^{\mathrm{T}} \nabla_{\boldsymbol{\pi}} \pi_k$$
$$+ P_{\alpha,\boldsymbol{\beta}}(Y|k,\mathbf{X}) f_k(\mathbf{X}) \psi_k(\mathbf{X}) \pi_k \Bigg\} = 0,$$

which is essentially

$$\sum_{k=0}^{S} \exp\Big(-\frac{(R-\mu_k)^2}{2\sigma_k^2}\Big)(a_k R^2 + b_k R + c_k) = 0, \tag{4.10}$$

where $\mu_k = \boldsymbol{\gamma}^{\mathrm{T}}\mathcal{A}(Y,k,\mathbf{X})$ and $\sigma_k^2 = \exp \boldsymbol{\delta}^{\mathrm{T}}\mathcal{C}(Y,k,\mathbf{X})$. Let $\phi_k = \exp\big\{-(R-\mu_k)^2/2\sigma_k^2\big\}$ and reorder the component $\phi_0, \ldots, \phi_S$ lexicographically by: $\phi_k \prec \phi_{k'}$ if $\sigma_k > \sigma_{k'}$ or if $\sigma_k = \sigma_{k'}$ but $\mu_k > \mu_{k'}$. Denote the ordered ϕ_k as $(\phi_{\tilde{0}}, \ldots, \phi_{\tilde{S}})$. Dividing (4.10) by $\phi_{\tilde{0}}$, we have

$$a_{\tilde{0}}R^2 + b_{\tilde{0}}R + c_{\tilde{0}} = -\sum_{k=\tilde{1}}^{\tilde{S}} \exp\Big(-\frac{(R-\mu_k)^2}{2\sigma_k^2} + \frac{(R-\mu_{\tilde{0}})^2}{2\sigma_{\tilde{0}}^2}\Big)(a_k R^2 + b_k R + c_k).$$

Each exponential component on the right hand side will be of the order either $\exp(-R^2)$ or $\exp(-R)$, so the right hand side will go to zero as $R \to \infty$. On the contrary, the left hand side will go away from 0 if $a_{\tilde{0}}$ and $b_{\tilde{0}}$ are not 0. Thus $a_{\tilde{0}} = 0$, $b_{\tilde{0}} = 0$ and $c_{\tilde{0}} = 0$. Next, divide the remains of (4.10) by $\phi_{\tilde{1}}$. By the same argument, $a_{\tilde{1}} = 0$, $b_{\tilde{1}} = 0$ and $c_{\tilde{1}} = 0$. Iterate until a_k, b_k and c_k are 0 for all k, which implies $\boldsymbol{\mu}_{\boldsymbol{\gamma},\boldsymbol{\delta}} = 0$ and for any k,

$$\boldsymbol{\mu}_{\alpha,\boldsymbol{\beta}}^{\mathrm{T}} \nabla_{\alpha,\boldsymbol{\beta}} P_{\alpha,\boldsymbol{\beta}}(Y|k,\mathbf{X}) f_k(\mathbf{X})\pi_k + P_{\alpha,\boldsymbol{\beta}}(Y|k,\mathbf{X}) f_k(\mathbf{X}) \boldsymbol{\mu}_{\boldsymbol{\pi}}^{\mathrm{T}} \nabla_{\boldsymbol{\pi}}\pi_k \\ + P_{\alpha,\boldsymbol{\beta}}(Y|k,\mathbf{X}) f_k(\mathbf{X}) \psi_k(\mathbf{X})\pi_k = 0.$$

Summarizing over $Y = 0, 1$ and applying Condition 4.4 yields $\boldsymbol{\mu}_{\boldsymbol{\pi}}^{\mathrm{T}}\nabla_{\boldsymbol{\pi}}\pi_k + \psi_k(\mathbf{X})\pi_k = 0$. Further, taking expectation over $\mathbf{X}$ gives $\boldsymbol{\mu}_{\boldsymbol{\pi}} = 0$ and hence $\psi_k = 0$. Thus, $\boldsymbol{\mu}_{\alpha,\boldsymbol{\beta}}^{\mathrm{T}} \nabla_{\alpha,\boldsymbol{\beta}} P_{\alpha,\boldsymbol{\beta}}(Y|k,\mathbf{X}) f_k(\mathbf{X})\pi_k = 0$, for any k. By Condition 4.3 and 4.4, $\boldsymbol{\mu}_{\alpha,\boldsymbol{\beta}} = 0$.

We state the asymptotic results in Theorem 4.1.

THEOREM 4.1. Assume that Conditions 4.1-4.4 hold. Then $|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sup_{\mathbf{x},k,l} |\widehat{F_{k,l}}(\mathbf{x}) - F_{k,l,0}(\mathbf{x})| \to 0$ almost surely, and $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

NPMLE of Likelihood (4.3)

We first verity the identifiability of the parameters in Lemma 4.3.

LEMMA 4.3. If two sets of parameters $(\boldsymbol{\theta}, F)$ and $(\widetilde{\boldsymbol{\theta}}, \widetilde{F})$ yield the same likelihood, then $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$ and $F = \widetilde{F}$.

Proof: Suppose that

$$\begin{aligned}
&\left\{\sum_{k=0}^{S}\sum_{l=0}^{k} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}|Y,k,l,\mathbf{X}) \frac{\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{X})\}P_{k,p_{\mathrm{B}}}(l)\pi_k f(X)}{\int_{\mathbf{x}}\sum_{k'}\sum_{l'}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k',l',\mathbf{x})\}P_{k',p_{\mathrm{B}}}(l')\pi_{k'}\mathrm{d}F(\mathbf{x})}\right\}^{I(Y=1)} \\
&\qquad \times \left\{\sum_{k=0}^{S}\sum_{l=0}^{k} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}|Y,k,l,\mathbf{X})P_{k,p_{\mathrm{B}}}(l)\pi_k f(\mathbf{X})\right\}^{I(Y=0)} \\
&= \left\{\sum_{k=0}^{S}\sum_{l=0}^{k} P_{\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\delta}}}(\mathbf{R}|Y,k,l,\mathbf{X}) \frac{\exp\{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{X})\}P_{k,\widetilde{p}_{\mathrm{B}}}(l)\widetilde{\pi}_k \widetilde{f}(X)}{\int_{\mathbf{x}}\sum_{k'}\sum_{l'}\exp\{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(k',l',\mathbf{x})\}P_{k',\widetilde{p}_{\mathrm{B}}}(l')\widetilde{\pi}_{k'}\mathrm{d}\widetilde{F}(\mathbf{x})}\right\}^{I(Y=1)} \\
&\qquad \times \left\{\sum_{k=0}^{S}\sum_{l=0}^{k} P_{\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\delta}}}(\mathbf{R}|Y,k,l,\mathbf{X})P_{k,\widetilde{p}_{\mathrm{B}}}(l)\widetilde{\pi}_k \widetilde{f}(\mathbf{X})\right\}^{I(Y=0)}. \qquad (4.11)
\end{aligned}$$

Letting $Y = 0$ in (4.11), we obtain

$$\sum_{k=0}^{S}\sum_{l=0}^{k} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}|Y,k,l,\mathbf{X})P_{k,p_{\mathrm{B}}}(l)\pi_k f(\mathbf{X}) = \sum_{k=0}^{S}\sum_{l=0}^{k} P_{\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\delta}}}(\mathbf{R}|Y,k,l,\mathbf{X})P_{k,\widetilde{p}_{\mathrm{B}}}(l)\widetilde{\pi}_k \widetilde{f}(\mathbf{X}).$$

By Proposition 1 of Teicher (1963) and Condition 4.2, $\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}}$, $\boldsymbol{\delta} = \widetilde{\boldsymbol{\delta}}$, and

$$P_{k,p_{\mathrm{B}}}(l)\pi_k f(\mathbf{X}) = P_{k,\widetilde{p}_{\mathrm{B}}}(l)\widetilde{\pi}_k \widetilde{f}(\mathbf{X}),$$

for any (k, l). Summarizing over l gives $\pi_k f(\mathbf{X}) = \widetilde{\pi}_k \widetilde{f}(\mathbf{X})$. Further, summarizing over k yields $f(.) = \widetilde{f}(.)$ and then $\pi_k = \widetilde{\pi}_k$ for any k. Thus, $P_{k,p_{\mathrm{B}}}(l) = P_{k,\widetilde{p}_{\mathrm{B}}}(l)$, implying

$p_{\mathrm{B}} = \widetilde{p}_{\mathrm{B}}$. Letting $Y = 1$ in (4.11) and applying Proposition 1 of Teicher (1963) again, we see that for any (k, l),

$$\frac{\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{X})\}}{\int_{\mathbf{x}}\sum_{k'}\sum_{l'}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k',l',\mathbf{x})\}P_{k',p_{\mathrm{B}}}(l')\pi_{k'}\mathrm{d}F(\mathbf{x})}$$
$$= \frac{\exp\{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{X})\}}{\int_{\mathbf{x}}\sum_{k'}\sum_{l'}\exp\{\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(k',l',\mathbf{x})\}P_{k',\widetilde{p}_{\mathrm{B}}}(l')\widetilde{\pi}_{k'}\mathrm{d}\widetilde{F}(\mathbf{x})}.$$

It then follows from Condition 4.1 that $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$.

We then verify that the information matrices along all non-trivial parametric submodels are non-singular in LEMMA 4.4.

LEMMA 4.4. If there exist a vector $\boldsymbol{\mu}_{\boldsymbol{\theta}} = (\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\mu}_{\boldsymbol{\pi}}, \mu_{p_{\mathrm{B}}}, \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\mu}_{\boldsymbol{\delta}})$ and functions $\psi(\mathbf{x})$ with $E[\psi(\mathbf{X})] = 0$ such that

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathrm{T}} l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, F_0) + l_F(\boldsymbol{\theta}_0, F_0)[\int \psi \; dF_0] = 0, \tag{4.12}$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$, and $l_F[\int \psi \; dF_0]$ is the score function for F along the submodel $F_0 + \epsilon \int \psi \; dF_0$, then $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbf{0}$ and $\psi = 0$.

Proof: We first set $Y = 0$ in (4.12), which then becomes

$$\sum_{k,l} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}|Y,k,l,\mathbf{X})\Bigg\{\boldsymbol{\mu}_{\boldsymbol{\gamma},\boldsymbol{\delta}}^{\mathrm{T}}\nabla_{\boldsymbol{\gamma},\boldsymbol{\delta}}\log P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}|Y,k,l,\mathbf{X})P_{k,p_{\mathrm{B}}}(l)\pi_k$$
$$+ \boldsymbol{\mu}_{\boldsymbol{\pi},p_{\mathrm{B}}}^{\mathrm{T}}\nabla_{\boldsymbol{\pi},p_{\mathrm{B}}}P_{k,p_{\mathrm{B}}}(l)\pi_k + P_{k,p_{\mathrm{B}}}(l)\pi_k\psi(\mathbf{X})\Bigg\} = 0.$$

By the same argument in the proof of Lemma 4.2, $\boldsymbol{\mu}_{\boldsymbol{\gamma},\boldsymbol{\delta}} = 0$ and for any (k, l),

$$\boldsymbol{\mu}_{\boldsymbol{\pi},p_{\mathrm{B}}}^{\mathrm{T}}\nabla_{\boldsymbol{\pi},p_{\mathrm{B}}}P_{k,p_{\mathrm{B}}}(l)\pi_k + P_{k,p_{\mathrm{B}}}(l)\pi_k\psi(\mathbf{X}) = 0.$$

Summarizing over l yields $\psi = 0$ and then $\boldsymbol{\mu}_{\boldsymbol{\pi},p_{\mathrm{B}}} = 0$.

Letting $Y = 1$ in (4.12), we have

$$\sum_{k,l} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}|Y,k,l,\mathbf{X}) e^{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{X})} P_{k,p_{\mathrm{B}}}(l)\pi_k$$
$$\times \left\{ \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{X}) - \frac{\int_{\mathbf{x}} \sum_{k'} \sum_{l'} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k',l',\mathbf{x})} \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(k',l',\mathbf{X}) P_{k',p_{\mathrm{B}}}(l')\pi_{k'} \mathrm{d}F(\mathbf{x})}{\int_{\mathbf{x}} \sum_{k'} \sum_{l'} e^{\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(k',l',\mathbf{x})} P_{k',p_{\mathrm{B}}}(l')\pi_{k'} \mathrm{d}F(\mathbf{x})} \right\} = 0.$$

Applying Proposition 1 of Teicher (1963) yields $\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathcal{Z}(k,l,\mathbf{X})$ being a constant for all (k,l). By Condition 4.1, $\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}} = 0$.

We impose the following regularity conditions, and then state the asymptotic results in Theorem 4.2.

CONDITION 4.5. The fraction $n1/n \to \varrho \in (0,1)$.

CONDITION 4.6. The function f is positive in its support and continuously differentiable.

CONDITION 4.7. For $i = 1, \ldots, n$, the conditional distribution of Y_i given $(K_i, L_i, \mathbf{X}_i)$ satisfies that

$$P(Y_i|K_i, L_i, \mathbf{X}_i) = a_n \exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}(K_i, L_i, \mathbf{X}_i)\}/[1 + a_n \exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathcal{Z}(K_i, L_i, \mathbf{X}_i)\}],$$

where $a_n = o(n^{-1/2})$.

THEOREM 4.2. Assume that Conditions 4.1-4.2,4.5-4.7 hold. Then $|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sup_{\mathbf{x}} |\widehat{F}(\mathbf{x}) - F_0(\mathbf{x})| \to 0$ almost surely, and $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

Proof: The proof follows the arguments in the proofs of Theorem S.1 and S.2 in Hu et al. (2010).

4.5.2 Numerical Algorithms

EM Algorithm to maximize (4.2)

Suppose that there are $J_{k,l}$ distinct values of $\mathbf{X}$ given $(K = k, L = l)$, denoted by $\mathbf{x}_{k,l,1}, \ldots, \mathbf{x}_{k,l,J_{k,l}}$. Let $\eta_{k,l,j}$ be the jump size of $F_{k,l}$ at $\mathbf{x}_{k,l,j}$. Note that $\sum_{j=1}^{J_{k,l}} \eta_{k,l,j} = 1$. The complete-data score function is

$$\sum_{i=1}^{n}\sum_{k=0}^{S}\sum_{l=0}^{k}\sum_{j=1}^{J_{k,l}} I(K_i = k, L_i = l, \mathbf{X}_i = \mathbf{x}_{k,l,j}) \Bigg\{ \log P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i | Y_i, k, l, \mathbf{X}_i) \\ + \log P_{\alpha,\boldsymbol{\beta}}(Y_i | k, l, \mathbf{X}_i) + \log \eta_{k,l,j} + \log \xi_{k,l} \Bigg\}.$$

In the E-step, we evaluate $E\{I(K_i = k, L_i = l, \mathbf{X}_i = \mathbf{x}_{k,l,j}) | \mathbf{R}_i, Y_i, \mathbf{X}_i\}$, which can be shown to be

$$\omega_{iklj} \equiv \frac{I(\mathbf{X}_i = \mathbf{x}_{k,l,j}) P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i | Y_i, k, l, \mathbf{X}_i) P_{\alpha,\boldsymbol{\beta}}(Y_i | k, l, \mathbf{X}_i) \eta_{k,l,j} \xi_{k,l}}{\sum_{k',l'} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i | Y_i, k', l', \mathbf{X}_i) P_{\alpha,\boldsymbol{\beta}}(Y_i | k', l', \mathbf{X}_i) \eta_{k',l',j} \xi_{k',l'}}.$$

In the M-step we use the one-step Newton-Raphson iteration to update the parameter estimates based on first and second derivatives derived from the complete-data score function with $I(K_i = k, L_i = l, \mathbf{X}_i = \mathbf{x}_{k,l,j})$ replaced by ω_{iklj}. Note that the update of $\eta_{k,l,j}$ is subject to $\sum_{j=1}^{J_{k,l}} \eta_{k,l,j} = 1$ for any (k, l) and all $\eta_{k,l,j}$s are nonnegative. Similarly, $\xi_{k,l}$ is subject to $\sum_{k,l} \xi_{k,l} = 1$ and all $\xi_{k,l}$s are nonnegative. These constraints can be incorporated by transforming the parameters to $\eta_{k,l,j}^{\dagger} = \log(\eta_{k,l,j}/\eta_{k,l,1})$, where $j = 2, \ldots, J_{k,l}$ and $\xi_{k,l}^{\dagger} = \log(\xi_{k,l}/\xi_{0,0})$, where $(k, l) \neq (0, 0)$. The initial values of the parameters are set as follows. We set α, $\boldsymbol{\beta}$, $\eta_{k,l,j}^{\dagger}$ and $\xi_{k,l}^{\dagger}$ all to 0. We let $(\boldsymbol{\gamma}, \boldsymbol{\delta})$

take the empirical means and variances of the clusters classified by any CNV calling method. For example, Birdsuite can directly call allele-specific copy numbers. Starting with such initial values, we iterate between the E-step and M-step until the change of $\log L_{\mathrm{p}}(\boldsymbol{\theta}, \{F_{k,l}\})$ is negligible.

We can estimate the limiting covariance matrix of $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}_{k,l}$ by inverting the (observed-data) information matrix for all the parameters including the jump sizes of $\widehat{F}_{k,l}$. The information matrix is obtained via the Louis (1982) formula. We can also estimate the limiting covariance matrix of $\widehat{\boldsymbol{\theta}}$ by using the profile likelihood function $pl_n(\boldsymbol{\theta}) \equiv \max_{\{F_{k,l}\}} \log L_{\mathrm{p}}(\boldsymbol{\theta}, \{F_{k,l}\})$. Particularly, the (s,t)th element of the inverse covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by $-\epsilon_n^{-2}\Big\{pl_n(\widehat{\boldsymbol{\theta}} + \epsilon_n \mathbf{e}_s + \epsilon_n \mathbf{e}_t) - pl_n(\widehat{\boldsymbol{\theta}} + \epsilon_n \mathbf{e}_s) - pl_n(\widehat{\boldsymbol{\theta}} + \epsilon_n \mathbf{e}_t) + pl_n(\widehat{\boldsymbol{\theta}})\Big\}$, where ϵ_n is a constant of order $n^{-1/2}$, and $\mathbf{e}_s$, and $\mathbf{e}_t$ are the sth and tth canonical vectors. We calculate $pl_n(\boldsymbol{\theta})$ via the EM algorithm by holding $\boldsymbol{\theta}$ constant in both the E-step and M-step.

Profile Likelihood of (4.3)

Suppose that there are J distinct values of $\mathbf{X}$, denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_J$. Let n_{+j} be the number of times that $\mathbf{x}_j$ is observed in the data and let η_j be the jump size of F at $\mathbf{x}_j$. Note that $\sum_{j=1}^J \eta_j = 1$. Before we use EM algorithm, we first show that $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_J)^{\mathrm{T}}$ can be profiled out by introducing one free parameter. The log-likelihood is

$$\begin{aligned}\widetilde{l}_n(\boldsymbol{\theta}, \boldsymbol{\eta}) = & \sum_{i=1}^n \log\left\{\sum_{k,l} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i | Y_i, k, l, \mathbf{X}_i) \exp\{Y_i \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(k, l, \mathbf{X}_i)\} P_{k,p_{\mathrm{B}}}(l) \pi_k\right\} \\ & + \sum_{j=1}^J n_{+j} \log \eta_j - n_1 \log\Big\{\sum_{j'} \sum_{k',l'} \exp\{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(k', l', \mathbf{x}_{j'})\} P_{k',p_{\mathrm{B}}}(l') \pi_{k'} \eta_{j'}\Big\}.\end{aligned}$$

We introduce a Lagrange multiplier λ for the constraint $\sum_{j=1}^J \eta_j = 1$ and set the derivative with respect to η_j to 0. We then obtain

$$\frac{n_{+j}}{\eta_j} - \frac{n_1 \sum_{k,l} \exp\{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(k,l,\mathbf{x}_j)\} P_{k,p_{\mathrm{B}}}(l)\pi_k}{\sum_j \sum_{k,l} \exp\{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(k,l,\mathbf{x}_j)\} P_{k,p_{\mathrm{B}}}(l)\pi_k \eta_j} + \lambda = 0.$$

Multiplying both sides by η_j and summing over $j = 1, \ldots, J$, we see that $\lambda = n_1 - n$. Thus

$$\eta_j = \frac{n_{+j}}{n_0 + n_1 \sum_{k,l} \exp\{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(k,l,\mathbf{x}_j)\} P_{k,p_{\mathrm{B}}}(l)\pi_k / \nu},$$

where $\nu = \sum_j \sum_{k,l} \exp\{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(k,l,\mathbf{x}_j)\} P_{k,p_{\mathrm{B}}}(l)\pi_k \eta_j$. Plugging η_j back into $\widetilde{l}(\boldsymbol{\theta}, \boldsymbol{\eta})$, we see the objective function to be maximized is,

$$\begin{aligned} l_n^*(\boldsymbol{\theta}, \nu) = & \sum_{i=1}^n \log\left\{\sum_{k,l} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i | Y_i, k, l, \mathbf{X}_i) \exp\{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(k,l,\mathbf{X}_i)\} P_{k,p_{\mathrm{B}}}(l)\pi_k\right\} \\ & - n_{+j} \log\left\{1 + \frac{n_1}{n_0 \nu} \sum_{k,l} \exp\{\boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(k,l,\mathbf{x}_j)\} P_{k,p_{\mathrm{B}}}(l)\pi_k\right\} - n_1 \log \nu. \end{aligned}$$

Suppose that the conditional distribution of $(\mathbf{R}, Y)$ given $\mathbf{X}$ is characterized by

$$P(\mathbf{R}, Y | \mathbf{X}) = P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R} | Y, k, l, \mathbf{X}) \frac{\exp\{\boldsymbol{\beta}^{*\mathrm{T}} \mathcal{Z}^*(k,l,\mathbf{X},Y)\} P_{k,p_{\mathrm{B}}}(l)\pi_k}{\sum_{y=0,1} \sum_{k',l'} \exp\{\boldsymbol{\beta}^{*\mathrm{T}} \mathcal{Z}^*(k',l',\mathbf{X},y)\} P_{k',p_{\mathrm{B}}}(l')\pi_{k'}},$$

where

$$\boldsymbol{\beta}^* = \begin{bmatrix} \log(n_1/(n_0\nu)) \\ \boldsymbol{\beta} \end{bmatrix}, \mathcal{Z}^*(k,l,\mathbf{x},y) = \begin{bmatrix} y \\ y\mathcal{A}(k,l,\mathbf{x}) \end{bmatrix}.$$

We can show that $l_n^*(\boldsymbol{\theta}, \nu)$ is equivalent to the log-likelihood

$$l_n^*(\boldsymbol{\vartheta}) = \sum_{i=1}^n \log\left\{\sum_{k,l} P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i | Y_i, k, l, \mathbf{X}_i) \frac{\exp\{\boldsymbol{\beta}^{*\mathrm{T}} \mathcal{Z}^*(k,l,\mathbf{X}_i,Y_i)\} P_{k,p_{\mathrm{B}}}(l)\pi_k}{\sum_{y=0,1} \sum_{k',l'} \exp\{\boldsymbol{\beta}^{*\mathrm{T}} \mathcal{Z}^*(k',l',\mathbf{X}_i,y)\} P_{k',p_{\mathrm{B}}}(l')\pi_{k'}}\right\},$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\beta}^*, \boldsymbol{\pi}, p_{\mathrm{B}}, \boldsymbol{\gamma}, \boldsymbol{\delta})$. We maximize $l_n^*(\boldsymbol{\vartheta})$ through the EM algorithm, in which (K, L) is treated as missing. The estimation of the covariance matrix of $\widehat{\boldsymbol{\vartheta}}$ is based on the information matrix of $l_n^*(\boldsymbol{\vartheta})$.

EM Algorithm to maximize (4.3)

The complete-data score function is

$$\sum_{i=1}^{n}\sum_{k=0}^{S}\sum_{l=0}^{k} I(K_i=k,L_i=l)\left\{\log P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i|Y_i,k,l,\mathbf{X}_i)\right.$$
$$\left.+\log\frac{\exp\{\boldsymbol{\beta}^{*\mathrm{T}}\mathcal{Z}^*(k,l,\mathbf{X}_i,Y_i)\}P_{k,p_{\mathrm{B}}}(l)\pi_k}{\sum_{y=0,1}\sum_{k',l'}\exp\{\boldsymbol{\beta}^{*\mathrm{T}}\mathcal{Z}^*(k',l',\mathbf{X}_i,y)\}P_{k',p_{\mathrm{B}}}(l')\pi_{k'}}\right\}.$$

In the E-step, we evaluate $E\{I(K=k,L=l)|\mathbf{R},Y,\mathbf{X}\}$, which can be shown to be

$$\omega_{ikl}\equiv\frac{P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i|Y_i,k,l,\mathbf{X}_i)\exp\{\boldsymbol{\beta}^{*\mathrm{T}}\mathcal{Z}^*(k,l,\mathbf{X}_i,Y_i)\}P_{k,p_{\mathrm{B}}}(l)\pi_k}{\sum_{k',l'}P_{\boldsymbol{\gamma},\boldsymbol{\delta}}(\mathbf{R}_i|Y_i,k',l',\mathbf{X}_i)\exp\{\boldsymbol{\beta}^{*\mathrm{T}}\mathcal{Z}^*(k',l',\mathbf{X}_i,Y_i)\}P_{k',p_{\mathrm{B}}}(l')\pi_{k'}}.$$

In the M-step we use the one-step Newton-Raphson iteration to update the parameter estimates based on first and second derivatives derived from the complete-data score function with $I(K_i=k,L_i=l)$ replaced by ω_{ikl}. The initial values of the parameters are set as follows. We set $\boldsymbol{\beta}^*=0$ and $\boldsymbol{\pi}=(1/(1+S),\ldots,1/(1+S))$. We let p_B take the pfb (population frequency of B allele) values in the annotation file of the platform. We let $(\boldsymbol{\gamma},\boldsymbol{\delta})$ take the empirical means and variances of the clusters classified by any CNV calling method, for example, Birdsuite, which directly calls allele-specific copy numbers. Starting with such initial values, we iterate between the E-step and M-step until the change in $l_n^*(\boldsymbol{\vartheta})$ is negligible. Finally, the information matrix of $l_n^*(\boldsymbol{\vartheta})$ is obtained via the Louis (1982) formula.

Chapter 5

Ongoing and Future Research

5.1 Analysis of Untyped SNPs: Tagging-Based and HMM-Based Methods

5.1.1 Introduction

In Chapel 3, we compared maximum likelihood and imputation methods in analysis of untyped SNPs, both based on tag SNPs. In this chapter, we compare tagging-based and HMM-based methods in the analysis of untyped SNPs. Specifically, we consider four methods: 1) the expected genotype count (dosage) imputed by tag SNPs (e.g., tagIMPUTE of Hu and Lin, 2010); 2) the expected genotype count imputed by all SNPs in LD with the untyped SNP (e.g., beagle of Browning and Browning, 2007, 2009); 3) maximum likelihood methods based on tag SNPs (SNPMStat of Lin et al., 2008, Hu and Lin, 2010); 4) a new quasi-maximum likelihood approach to incorporate posterior probabilities of genotypes at untyped SNPs generated by any imputation methods (in particular, HMM-based methods) and to account for the uncertainty in inferring the posterior probabilities. We establish the theoretical properties of the proposed quasi-maximum likelihood method and conduct extensive simulation studies, based on a whole-genome simulation program that mimics the LD patterns in human populations, to evaluate the performance of the four methods in testing/estimating genetic effects and gene-environment interactions. We apply the four methods to the GWAS data from the WTCCC (Burton et al., 2007).

5.1.2 Methods

We propose a new method to analyze the imputed posterior probabilities of genotypes at untyped SNPs while properly accounting for the imputation uncertainty. We let G_t denote all genotyped SNPs on the chromosome and let G_u denote the untyped SNP of interest. Write $G = (G_t, G_u)$. As usual, let Y denote the phenotype of interest, which can be quantitative or qualitative, and $\mathbf{X}$ denote a set of environmental factors. The joint density of the observed data $(Y, \mathbf{X}, G_t)$ can be decomposed as

$$P(Y, \mathbf{X}, G_t) = \sum_{g=0,1,2} P(Y|\mathbf{X}, G_t, G_u = g)P(\mathbf{X}|G_t, G_u = g)P(G_u = g|G_t)P(G_t).$$

For analyzing the marginal effect of the untyped SNP, $P(Y|\mathbf{X}, G_t, G_u)$ reduces to $P(Y|\mathbf{X}, G_u)$. In addition, we assume that the genetic factors are independent of the environmental factors, so that $P(\mathbf{X}|G_t, G_u) = P(\mathbf{X})$, which is modelled nonparametricly with distribution function $F(.)$. We can obtain an estimate of $P(G_u|G_t)$ from any imputation method, such as beagle or tagIMPUTE. In the presence of a large number of SNPs in G_t, we can neither estimate $P(G_t)$ nor obtain it from any imputation method, as it is close to zero for every possible value. Fortunately, we can avoid estimating $P(G_t)$ as shown below.

Parameter Estimation

We propose to estimate the parameters by maximizing the objective functions derived below. These objective functions are based on likelihoods for different study designs, but not themselves proper likelihoods because of the use of the reference sample. Thus we adopt the term "quasi-maximum likelihood estimator (qMLE)" for the method.

In a cross-sectional study, we measure $\mathbf{X}$ and G on n study subjects. We characterize the association between Y and $(\mathbf{X}, G_u)$ by the conditional density $P_{\alpha,\boldsymbol{\beta},\xi}(Y|\mathbf{X}, G_u)$, where α,$\boldsymbol{\beta}$, and ξ denote the intercept(s), regression effects and the nuisance parameters (variance and overdispersion parameters), respectively. The prospective likelihood takes the form

$$\begin{aligned} L_n(\alpha, \boldsymbol{\beta}, \xi) &= \prod_{i=1}^{n} \sum_{g=0,1,2} P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, G_u = g)P(G_u = g|G_{Oi})P(G_{Oi}) \\ &\propto \prod_{i=1}^{n} \sum_{g=0,1,2} P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, G_u = g)P(G_u = g|G_{Oi}). \end{aligned}$$

Denoting $c_{i,g}$ as the estimate of $P(G_u = g|G_{Oi})$, $g = 0, 1, 2$, for the ith subject, generated from the previous stage of imputation, the objective function to be maximized is, after plugging in c_{i,G_u},

$$\tilde{L}_n(\alpha, \boldsymbol{\beta}, \xi) = \prod_{i=1}^{n} \sum_{g=0,1,2} P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, G_u = g)c_{i,g}.$$

In a case-control study, we measure $\mathbf{X}$ and G on n_1 cases $(Y = 1)$ and n_0 controls $(Y = 0)$. It is natural to formulate the effects of $\mathbf{X}$ and G_u on Y through the logistic regression model

$$P_{\alpha,\boldsymbol{\beta}}(Y|\mathbf{X}, G_u) = \frac{e^{Y(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X},G_u))}}{1 + e^{\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X},G_u)}},$$

where α is an intercept, $\boldsymbol{\beta}$ is a set of log odds ratios, and $\mathcal{Z}(\mathbf{X}, G_u)$ is a vector-function of $\mathbf{X}$ and G_u under a particular mode of inheritance. We make the rare disease assumption, so $P_{\alpha,\boldsymbol{\beta}}(Y|\mathbf{X}, G_u) \approx e^{Y(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X},G_u))}$. To reflect the case-control sampling, we adopt the retrospective likelihood

$$L_n(\boldsymbol{\beta},F) = \prod_{i=1}^{n} \frac{\sum_{g=0,1,2}\exp\big\{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X}_i,g)\big\}F(\mathbf{X}_i)P(G_u=g|G_{Oi})P(G_{Oi})}{\int_{\mathbf{x}}\sum_{G_t,g}\exp\big\{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{x},g)\big\}P(G_t,G_u=g)dF(\mathbf{x})}$$
$$= \prod_{i=1}^{n} \frac{\sum_{g}\exp\big\{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X}_i,g)\big\}F(\mathbf{X}_i)P(G_u=g|G_{Oi})P(G_{Oi})}{\int_{\mathbf{x}}\sum_{g}\exp\big\{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{x},g)\big\}P(G_u=g)dF(\mathbf{x})}$$

We further let m_g, $g=0,1,2$, denote the estimates of the marginal distributions of the untyped SNP, i.e., $m_g = P(G_u = g)$. Assuming that the study and reference sample are from the same population, we obtain m_g as

$$\frac{n_0}{n_0+\widetilde{n}}\widehat{p}_C(G_u=g) + \frac{\widetilde{n}}{n_0+\widetilde{n}}\widehat{p}_R(G_u=g),$$

where n_0 is the number of controls, $\widetilde{n}$ is the number of founders in the reference panel (if it consists of trios), $\widehat{p}_R(G_u=g)$ is simply the empirical frequency of genotype g at the untyped SNP in the reference sample, and $\widehat{p}_C(G_u=g)=\sum_{i=1}^{n_0} c_{i,g}$. The estimator $\widehat{p}_C(G_u=g)$ of the marginal distribution of G_u in the control sample results from the fact that $P(G_u=g)=E_{G_t}P(G_u=g|G_t)$, which can be approximated by $\sum_{i=1}^{n_0} c_{i,g}$. Plugging in $c_{i,g}$ and m_g, the objective function to be maximized is

$$\tilde{L}_n(\boldsymbol{\beta},F) = \prod_{i=1}^{n} \frac{\sum_{g=0,1,2}\exp\big\{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X}_i,g)\big\}F(\mathbf{X}_i)c_{i,g}}{\int_{\mathbf{x}}\sum_{g}\exp\big\{Y_i\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{x},g)\big\}m_g dF(\mathbf{x})}.$$

Note that $\tilde{L}_n(\boldsymbol{\beta},F)$ involves infinite-dimensional parameters if $\mathbf{X}$ have continuous components. By profiling out $F(\mathbf{X})$, the MLE of $\boldsymbol{\beta}$ can be equivalently obtained by maximizing

$$L_n^*(\boldsymbol{\beta},\tilde{\mu}) = \prod_{i=1}^{n} \frac{\sum_{g=0,1,2} e^{Y_i\{\tilde{\mu}+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X}_i,g)\}}c_{i,g}}{\sum_{y=0,1}\sum_{g} e^{y\{\tilde{\mu}+\boldsymbol{\beta}^{\mathrm{T}}\mathcal{Z}(\mathbf{X}_i,g)\}}m_g},$$

where $\tilde{\mu}$ is a free parameter. When there are no environmental factors, the objective function is based on the retrospective likelihood

$$\tilde{L}_n(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\sum_{g=0,1,2} e^{Y_i \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(G_u)} c_{i,g}}{\sum_g e^{Y_i \boldsymbol{\beta}^{\mathrm{T}} \mathcal{Z}(g)} m_g}.$$

Variance Estimation Accounting for Imputation Uncertainty

We denote all relevant parameters in the objective functions by $\boldsymbol{\theta}$. The usual variance estimatior for $\boldsymbol{\theta}$ based on the inverse the negative second derivatives of the corresponding objective functions does not account for the imputation variation induced by $c_{i,g}$ and m_g. We can incorporate the uncertainty by bootstrapping the samples from which $c_{i,g}$ and m_g are derived. While $c_{i,g}$ solely relies on the reference samples, m_g depends on both the study and reference samples. Thus we bootstrap all samples B times, where B is set to be 20 here. For each bootstrap, we obtain a new set of $c_{i,g}$ and m_g, plug into the objective functions which is always constructed from the original study samples, and obtain the estimator $\widehat{\boldsymbol{\theta}}_b$, $b = 1, \ldots, B$. Finally, denoting $\bar{\theta} = B^{-1} \sum_{b=1}^{B} \widehat{\boldsymbol{\theta}}_b$ as the parameter estimation averaging over the bootstraps, the variance for $\bar{\theta}$ is estimated to be

$$\frac{1}{B} \sum_{b=1}^{B} \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}_b) + \frac{B+1}{B} \cdot \frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\boldsymbol{\theta}}_b - \bar{\boldsymbol{\theta}})^2,$$

where $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}_b)$ is obtained by inverting the negative second derivatives of the objective functions at the bth bootstrap.

5.1.3 Future Work

We will establish the theoretical properties of the proposed quasi-maximum likelihood method. We will conduct extensive simulation studies, based on a whole-genome simulation program that mimics the LD patterns in human populations (e.g., GWAsimulator of Li and Li, 2008), to evaluate the performance of the four methods in testing/estimating genetic effects and gene-environment interactions. Specifically, we will explore the extent of the LD information loss due to the restricted number of tag SNPs. To this end, we will compare tagIMPUTE and beagle in terms of the imputation accuracy and the power of association testing using their imputed dosages. We will also compare the power of beagle-based imputation, MLE and qMLE. We will apply all of these methods to the GWAS data from the WTCCC.

5.2 Association Analysis of Allele-Specific Copy Numbers Using Sequence Data

Figure 5.1 shows that the next-generation sequencing data and SNP array data are highly similar in term of measuring the total copy number. In Chapter 4, we modeled the intensity data from the SNP array by a mixture of normal distributions. We plan to model the number of sequence reads from one window by a Poisson or negative binomial distribution. There are existing score tests to decide whether there is over-dispersion in the Poisson model. If there is over-dispersion, we choose a negative binomial distribution.

One aspect of information missing from Figure 5.1 is the measurement of the ASCN. The measurements of ASCNs using sequencing data do not readily fit into the model for SNP array data, so we are currently working on appropriate modeling of ASCN measurements by sequencing data.

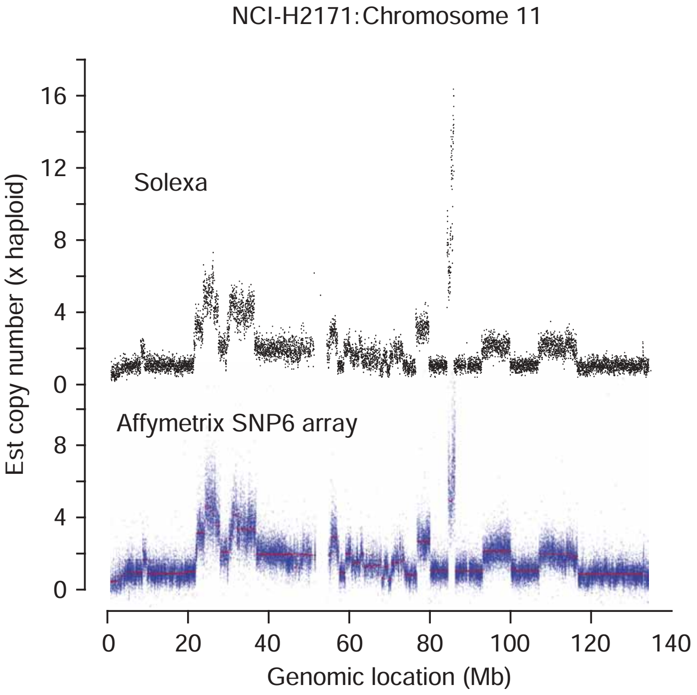

Figure 5.1: Comparison of copy number plots for chromosome 11 of NCI-H2171 between massively parallel paired-end sequencing and Affymetrix SNP6 genomic array data (Campbell et al., 2008). In the upper panel of the sequencing data, each point represents the number of mappable sequence reads in a sliding window of 15kb, which is transformed to the scale of copy number. In the lower panel of the array data, each point represents the fluorescent intensity measurement at a SNP site or a copy number probe.

Bibliography

Aitman, T., Dong, R., Vyse, T., Norsworthy, P., Johnson, M., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A., Petretto, E., et al. (2006), "Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans," *Nature*, 439, 851–855.

Akey, J., Jin, L., and Xiong, M. (2001), "Haplotypes vs single marker linkage disequilibrium tests: what do we gain?" *European Journal of Human Genetics*, 9, 291–300.

Altshuler, D. (2005), "A haplotype map of the human genome," *Nature*, 437, 1299–1320.

Amos, C., Wu, X., Broderick, P., Gorlov, I., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., et al. (2008), "Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25. 1," *Nature Genetics*, 40, 616–622.

Balding, D. (2006), "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, 7, 781–791.

Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D., and Hurles, M. (2008), "A robust statistical method for case-control association testing with copy number variation," *Nature Genetics*, 40, 1245–1252.

Bennett, S. (1983), "Analysis of survival data by the proportional odds model," *Statistics in Medicine*, 2, 273–277.

Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1993), *Efficient and adaptive estimation for semiparametric models*, Baltimore: Johns Hopkins University Press.

Browning, B. and Browning, S. (2009), "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals," *The American Journal of Human Genetics*, 84, 210–223.

Browning, S. and Browning, B. (2007), "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering," *The American Journal of Human Genetics*, 81, 1084–1097.

Burton, P., Clayton, D., Cardon, L., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D., McCarthy, M., Ouwehand, W., Samani, N., et al. (2007), "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, 447, 661–678.

Campbell, P., Stephens, P., Pleasance, E., O'Meara, S., Li, H., Santarius, T., Stebbings, L., Leroy, C., Edkins, S., Hardy, C., et al. (2008), "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing," *Nature Genetics*, 40, 722–729.

Carlson, C., Eberle, M., Rieder, M., Yi, Q., Kruglyak, L., and Nickerson, D. (2004), "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *The American Journal of Human Genetics*, 74, 106–120.

Chatterjee, N. and Carroll, R. (2005), "Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies," *Biometrika*, 92, 399–418.

Chen, H. (2004), "Nonparametric and semiparametric models for missing covariates in parametric regression," *Journal of the American Statistical Association*, 99, 1176–1189.

Chen, Y., Chatterjee, N., and Carroll, R. (2008), "Retrospective analysis of haplotype-based case–control studies under a flexible model for gene–environment association," *Biostatistics*, 9, 81–99.

— (2009), "Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies," *Journal of the American Statistical Association*, 104, 220–233.

Colella, S., Yau, C., Taylor, J., Mirza, G., Butler, H., Clouston, P., Bassett, A., Seller, A., Holmes, C., and Ragoussis, J. (2007), "QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data," *Nucleic Acids Research*, 35, 2013–2025.

Cordell, H. (2006), "Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures," *Genetic Epidemiology*, 30, 259–275.

Cox, D. (1972), "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187–220.

de Bakker, P., Yelensky, R., Pe'er, I., Gabriel, S., Daly, M., and Altshuler, D. (2005), "Efficiency and power in genetic association studies," *Nature Genetics*, 37, 1217–1223.

Epstein, M. and Satten, G. (2003), "Inference on haplotype effects in case-control studies using unphased genotype data," *The American Journal of Human Genetics*, 73, 1316–1329.

Excoffier, L. and Slatkin, M. (1995), "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population." *Molecular Biology and Evolution*, 12, 921–927.

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D., and Schork, N. (2001), "Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease," *Genome Research*, 11, 143–151.

Fanciulli, M., Norsworthy, P., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J., Gough, S., de Smith, A., Blakemore, A., et al. (2007), "FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity," *Nature Genetics*, 39, 721–723.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R., Freedman, B., Quinones, M., Bamshad, M., et al. (2005), "The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility," *Science*, 307, 1434–1440.

Hu, Y. and Lin, D. (2010), "Analysis of untyped SNPs: maximum likelihood and imputation methods," *Genetic Epidemiology*, 34, 803–815.

Hu, Y., Lin, D., and Zeng, D. (2010), "A general framework for studying genetic effects and gene–environment interactions with missing data," *Biostatistics*, 11, 583–598.

Iafrate, A., Feuk, L., Rivera, M., Listewnik, M., Donahoe, P., Qi, Y., Scherer, S., and Lee, C. (2004), "Detection of large-scale variation in the human genome," *Nature Genetics*, 36, 949–951.

Kidd, J., Cooper, G., Donahue, W., Hayden, H., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. (2008), "Mapping and sequencing of structural variation from eight human genomes," *Nature*, 453, 56–64.

Korn, J., Kuruvilla, F., McCarroll, S., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P., Darvishi, K., et al. (2008), "Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs," *Nature Genetics*, 40, 1253–1260.

Kraft, P., Cox, D., Paynter, R., Hunter, D., and De Vivo, I. (2005), "Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques," *Genetic Epidemiology*, 28, 261–272.

Lake, S., Lyon, H., Tantisira, K., Silverman, E., Weiss, S., Laird, N., and Schaid, D. (2003), "Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous," *Human Heredity*, 55, 56–65.

Li, C. and Li, M. (2008), "GWAsimulator: a rapid whole-genome simulation program," *Bioinformatics*, 24, 140–142.

Li, H. (2001), "A permutation procedure for the haplotype method for identification of disease-predisposing variants," *Annals of Human Genetics*, 65, 189–196.

Li, N. and Stephens, M. (2003), “Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data,” *Genetics*, 165, 2213–2233.

Li, Y., Willer, C., Ding, J., Scheet, P., and Abecasis, G. (2010), “MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes,” *Genetic Epidemiology*, 34, 816–834.

Lin, D. (2004), “Haplotype-based association analysis in cohort studies of unrelated individuals,” *Genetic Epidemiology*, 26, 255–264.

Lin, D., Hu, Y., and Huang, B. (2008), “Simple and efficient analysis of disease association with missing genotype data,” *The American Journal of Human Genetics*, 82, 444–452.

Lin, D. and Huang, B. (2007), “The use of inferred haplotypes in downstream analyses,” *The American Journal of Human Genetics*, 80, 577–579.

Lin, D. and Zeng, D. (2006), “Likelihood-based inference on haplotype effects in genetic association studies,” *Journal of the American Statistical Association*, 101, 89–118.

Little, R. (1992), “Regression with missing X’s: a review,” *Journal of the American Statistical Association*, 1227–1237.

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

Manolio, T., Rodriguez, L., Brooks, L., Abecasis, G., Ballinger, D., Daly, M., Donnelly, P., Faraone, S., Frazer, K., Gabriel, S., et al. (2007), “New models of collaboration in genome-wide association studies: the Genetic Association Information Network,” *Nature genetics*, 39, 1045–1051.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007), “A new multipoint method for genome-wide association studies by imputation of genotypes,” *Nature Genetics*, 39, 906–913.

McCarroll, S., Kuruvilla, F., Korn, J., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M., de Bakker, P., Maller, J., Kirby, A., et al. (2008), “Integrated detection and population-genetic analysis of SNPs and copy number variation,” *Nature Genetics*, 40, 1166–1174.

Morris, R. and Kaplan, N. (2002), “On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles,” *Genetic Epidemiology*, 23, 221–233.

Murphy, S. and van der Vaart, A. (2000), “On Profile Likelihood,” *Journal of the American Statistical Association*, 95.

Nicolae, D. (2006), "Testing Untyped Alleles (TUNA)applications to genome-wide association studies," *Genetic Epidemiology*, 30, 718–727.

Prentice, R. and Pyke, R. (1979), "Logistic disease incidence models and case-control studies," *Biometrika*, 66, 403–441.

Redon, R., Ishikawa, S., Fitch, K., Feuk, L., Perry, G., Andrews, T., Fiegler, H., Shapero, M., Carson, A., Chen, W., et al. (2006), "Global variation in copy number in the human genome," *Nature*, 444, 444–454.

Roeder, K., Carroll, R., and Lindsay, B. (1996), "A semiparametric mixture approach to case-control studies with errors in covariables," *Journal of the American Statistical Association*, 91, 722–732.

Satten, G. and Epstein, M. (2004), "Comparison of prospective and retrospective methods for haplotype inference in case-control studies," *Genetic Epidemiology*, 27, 192–201.

Schaid, D. (2004), "Evaluating associations of haplotypes with traits," *Genetic Epidemiology*, 27, 348–364.

Schaid, D., Rowland, C., Tines, D., Jacobson, R., and Poland, G. (2002), "Score tests for association between traits and haplotypes when linkage phase is ambiguous," *The American Journal of Human Genetics*, 70, 425–434.

Scheet, P. and Stephens, M. (2006), "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase," *The American Journal of Human Genetics*, 78, 629–644.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al. (2004), "Large-scale copy number polymorphism in the human genome," *Science*, 305, 525–528.

Shi, J., Levinson, D., Duan, J., Sanders, A., Zheng, Y., Peâ, I., et al. (2009), "Common variants on chromosome 6p22. 1 are associated with schizophrenia," *Nature*, 460, 753–757.

Spinka, C., Carroll, R., and Chatterjee, N. (2005), "Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity," *Genetic Epidemiology*, 29, 108–127.

Stephens, M., Smith, N., and Donnelly, P. (2001), "A new statistical method for haplotype reconstruction from population data," *The American Journal of Human Genetics*, 68, 978–989.

Stram, D. (2004), "Tag SNP selection for association studies," *Genetic Epidemiology*, 27, 365–374.

Stram, D., Pearcea, C., Bretskya, P., Freedman, M., Hirschhornb, J., Altshulerb, D., Kolonel, L., Hendersona, B., and Thomasa, D. (2003), "Modeling and EM estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals," *Human Heredity*, 55, 179–190.

Sun, W., Wright, F., Tang, Z., Nordgard, S., Loo, P., Yu, T., Kristensen, V., and Perou, C. (2009), "Integrated study of copy number states and genotype calls using high-density SNP arrays," *Nucleic Acids Research*, 37, 5365–5377.

Teicher, H. (1963), "Identifiability of finite mixtures," *The Annals of Mathematical Statistics*, 34, 1265–1269.

Tuzun, E., Sharp, A., Bailey, J., Kaul, R., Morrison, V., Pertz, L., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. (2005), "Fine-scale structural variation of the human genome," *Nature Genetics*, 37, 727–732.

van der Vaart, A. and Wellner, J. (1996), *Weak convergence and empirical processes*, New York: Springer-Verlag.

Van Loo, P., Nordgard, S., Lingjærde, O., Russnes, H., Rye, I., Sun, W., Weigman, V., Marynen, P., Zetterberg, A., Naume, B., et al. (2010), "Allele-specific copy number analysis of tumors," *Proceedings of the National Academy of Sciences*, 107, 16910–16915.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S., Hakonarson, H., and Bucan, M. (2007), "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data," *Genome Research*, 17, 1665–1674.

Xie, R. and Stram, D. (2005), "Asymptotic equivalence between two score tests for haplotype-specific risk in general linear models," *Genetic Epidemiology*, 29, 166–170.

Yang, Y., Chung, E., Wu, Y., Savelli, S., Nagaraja, H., Zhou, B., and Hebert, M. (2007), "Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans," *The American Journal of Human Genetics*, 80, 1037–1054.

Zaitlen, N., Kang, H., Eskin, E., and Halperin, E. (2007), "Leveraging the HapMap correlation structure in association studies," *The American Journal of Human Genetics*, 80, 683–691.

Zaykin, D., Westfall, P., Young, S., Karnouba, M., Wagner, M., and Ehm, M. (2002), "Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals," *Human Heredity*, 53, 79–91.

Zeggini, E., Scott, L., Saxena, R., Voight, B., Marchini, J., Hu, T., de Bakker, P., Abecasis, G., Almgren, P., Andersen, G., et al. (2008), "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes," *Nature Genetics*, 40, 638–645.

Zeng, D. and Lin, D. (2007), "Maximum likelihood estimation in semiparametric regression models with censored data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 507–564.

Zeng, D., Lin, D., Avery, C., North, K., and Bray, M. (2006), "Efficient semiparametric estimation of haplotype-disease associations in case–cohort and nested case–control studies," *Biostatistics*, 7, 486–502.

Zhao, L., Li, S., and Khalid, N. (2003), "A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies," *The American Journal of Human Genetics*, 72, 1231–1250.