Likelihood-Based Inference on Haplotype Effects in Genetic Association Studies

D. Y. Lin and D. Zeng

A haplotype is a specific sequence of nucleotides on a single chromosome. The population associations between haplotypes and disease phenotypes provide critical information about the genetic basis of complex human diseases. Standard genotyping techniques cannot distinguish the two homologous chromosomes of an individual so that only the unphased genotype (i.e., the combination of the two homologous haplotypes) is directly observable. Statistical inference about haplotype-phenotype associations based on unphased genotype data presents an intriguing missing-data problem, especially when the sampling depends on the disease status. The objective of this paper is to provide a systematic and rigorous treatment of this problem. All commonly used study designs, including cross-sectional, case-control and cohort studies, are considered. The phenotype can be a disease indicator, a quantitative trait or a potentially censored time to disease variable. The effects of haplotypes on the phenotype are formulated through flexible regression models, which can accommodate a variety of genetic mechanisms and gene-environment interactions. Appropriate likelihoods are constructed, which may involve high-dimensional parameters. The identifiability of the parameters, and the consistency, asymptotic normality and efficiency of the maximum likelihood estimators are established. Efficient and reliable numerical algorithms are developed. Simulation studies show that the likelihood-based procedures perform well in practical settings. An application to the Finland-United States Investigation of NIDDM Genetics Study is provided. Areas in need of further development are discussed.

KEY WRODS: Case-control studies; Hardy-Weinberg equilibrium; Missing data; Regression models; Single nucleotide polymorphisms; Unphased genotype.

D. Y. Lin is Dennis Gillings Distinguished Professor, and D. Zeng is Assistant Professor, Department of Biostatistics, CB#7420, University of North Carolina, Chapel Hill, NC 27599-7420 (Emails: lin@bios.unc.edu; dzeng@bios.unc.edu). The authors are grateful to the FUSION Study Group for sharing their data and to Michael Boehnke and Laura Scott for transmitting the data. They also thank the Editor, an Associate Editor and two referees for their comments. This work was supported by the National Institutes of Health.

1. INTRODUCTION

In the early 1900's, there was a fierce debate between Gregor Mendel's followers and the biometrical school led by Francis Galton and Karl Pearson as to whether the patterns of inheritance were consistent with Mendel's law of segregation or a "blending"-type theory. Fisher (1918) reconciled the two conflicting schools by recognizing the difference in the genetic basis for the variation in the trait being studied: for the traits Mendelists studied, the observed variation was due to a simple difference at a single gene; for the traits studied by the biometrical school, individual differences were attributed to many different genes, with no particular gene having a singly large effect.

Like the traits studied by Mendel, many genetic disorders, such as Huntington disease and cystic fibrosis, are caused by mutations of single genes. The genes underlying a number of these Mendelian syndromes have been discovered during the last twenty years through linkage analysis and positional cloning (Risch, 2000). The same approach, however, is failing to unravel the genetic basis of complex human diseases (e.g., hypertension, bipolar disorder, diabetes and schizophrenia), which are influenced by a variety of genetic and environmental factors, just like the traits studied by the biometrical school a century ago. It is widely recognized that genetic dissection of complex human disorders requires large-scale association studies, which relate disease phenotypes to genetic variants, especially single nucleotide polymorphisms (Risch, 2000; Botstein and Risch, 2003).

Single nucleotide polymorphisms or SNPs are DNA sequence variations that occur when a single nucleotide in the genome sequence is altered. SNPs make up about 90% of all human genetic variation and are believed to have a major impact on disease susceptibility. Aided by the sequencing of the human genome (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001), geneticists have identified several millions SNPs (International SNP Map Working Group, 2001). With current technology, it is economically feasible to genotype thousands of subjects for thousands of SNPs. These remarkable scientific and technological advances offer unprecedented opportunities to conduct SNPs-based association studies to unravel the genetic basis of complex diseases.

There are three possible genotypes at each SNP site: homozygous with allele A, homozygous with allele a, or heterozygous with one allele A and one allele a. Thus, the assessment of the association between a SNP and a disease phenotype is a trivial three-sample problem. It is, however, desirable to deal with multiple SNPs simultaneously. One appealing approach is to consider the haplotypes for multiple SNPs within candidate genes (Hallman et al., 1999; International SNP Map Working Group, 2001; Patil et al., 2001; Stephens et al., 2001).

The haplotype, which is a specific combination of nucleotides at a series of closely linked SNPs

on the same chromosome of an individual, contains information about the protein products. Because the actual number of haplotypes within a candidate gene is much smaller than the number of all possible haplotypes, haplotyping serves as an effective data-reduction strategy. The use of SNPs-based haplotypes may yield more powerful tests of genetic associations than the use of individual, unorganized SNPs, especially when the causal variants are not measured directly or when there are strong interactions of 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).

The determination of the haplotype requires the parental origin or gametic phase information, which cannot be easily obtained with the current genotyping technology. As a result, only the unphased genotype, i.e., the combination of the two homologous haplotypes, can be determined. Statistically speaking, this is a missing-data problem in which the variable of interest pertains to two ordered sequences of zeros and ones but only the summation of the two sequences is observed. This type of missing-data problem has not been studied in the statistical literature.

Many authors (e.g., Clark, 1990; Excoffier and Slatkin, 1995; Stephens et al., 2001; Zhang et al., 2001; Niu et al., 2002; Qin et al., 2002) proposed methods to infer haplotypes or estimate haplotype frequencies from unphased genotype data. To make inference about haplotype effects, one may then relate the probabilistically inferred haplotypes to the phenotype through a regression model (e.g., Zaykin et al., 2002). This approach does not account for the variation due to haplotype estimation, and does not yield consistent estimators of regression parameters.

A growing number of papers have been published in genetic journals on how to make proper inference about the effects of haplotypes on disease phenotypes. Most of these papers deal with case-control studies. Specifically, Zhao et al. (2003) developed an estimating function which approximates the expectation of the complete-data prospective-likelihood score function given the observable data. This method assumes that the disease is rare and that haplotypes are independent of environmental variables, and is not statistically efficient. Epstein and Satten (2003) derived a retrospective likelihood for the relative risk, which does not accommodate environmental variables. Stram et al. (2003) proposed a conditional likelihood for the odds ratio assuming that cases and controls are chosen randomly with known probabilities from the target population, and did not consider environmental variables either. The properties of the estimator were not investigated. Building on the earlier work of Schaid et al. (2002), Lake et al. (2003) discussed likelihood-based inference for cross-sectional studies under generalized linear models. Seltman, Roeder and Devlin (2003) provided a similar discussion based on the cladistic approach. Recently, 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. All the aforementioned work assumes Hardy-Weinberg equilibrium (Weir, 1996, p. 40). Simulation studies (Epstein and Satten 2003; Lake et al. 2003; Satten and Epstein 2004) revealed that viloation of this assumption can adversely affect the validity of the inference.

The aim of this paper is to address statistical issues in estimating haplotype effects in a systematic and rigorous manner. For case-control studies, we allow environmental variables and derive efficient inference procedures. For cross-sectional and cohort studies, we consider more versatile models than the existing literature. For all study designs, we accommodate Hardy-Weinberg disequilibrium. We construct appropriate likelihoods for a variety of models. Under the case-control sampling, the likelihood pertains to the distribution of genotypes and environmental variables conditional on the case-control status, which involves infinite-dimensional nuisance parameters if environmental variables are continuous. In cohort studies, it is desirable not to parametrize the distribution of time to disease, so that the likelihood also involves infinite-dimensional parameters. The presence of infinite-dimensional parameters entails considerable theoretical and computational challenges. We establish the theoretical properties of the maximum likelihood estimators by appealing to modern asymptotic techniques, and develop efficient and stable algorithms to implement the corresponding inference procedures. We assess the performance of the proposed methods through simulation studies and provide an application to a major genetic study of Type 2 diabetes.

2. INFERENCE PROCEDURES

2.1. Preliminaries

We consider SNPs-based association studies of unrelated individuals. Suppose that each individual is genotyped at M biallelic SNPs within a candidate gene. At each SNP site, we indicate the two possible alleles by the values 0 and 1. Thus, each haplotype h is a unique sequence of M numbers from $\{0, 1\}$. The total number of possible haplotypes is $K \equiv 2^M$. The actual number of haplotypes consistent with the data is usually much smaller. For $k = 1, \ldots, K$, let h_k denote the kth possible haplotype. Figure 1 shows the 8 possible haplotypes for 3 SNPs.

The human chromosomes come in pairs, one inherited from our mother and one from our father. These pairs are called homologous chromosomes. Thus, each individual has a pair of homologous haplotypes, which may or may not be identical. Routine genotyping procedures cannot separate the two homologous chromosomes, so that only the (unphased) genotypes, i.e., the combinations of the two homologous haplotypes, are directly observable. For each individual, the multi-SNP genotype is an ordered sequence of M numbers from $\{0, 1, 2\}$.

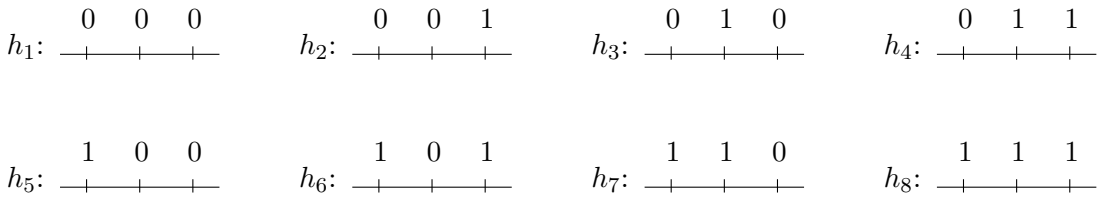

Fig. 1. Possible haplotype configurations with 3 SNPs.

Let H and G denote the pair of haplotypes and the genotype for an individual. We write $H = (h_k, h_l)$ if the individual's pair of haplotypes are h_k and h_l, in which case $G = h_k + h_l$. The ordering of the two homologous haplotypes within an individual is considered arbitrary. By allowing genotypes to include missing SNP information, we may assume that G is known for each individual. Given G, the value of H is unknown if the individual is heterozygous at more than 1 SNP or if any SNP genotype is missing. For the case of $M = 3$ shown in Fig. 1, if $G = (0, 2, 1)$, then $H = (h_3, h_4)$; if $G = (0, 1, 1)$, then $H = (h_1, h_4)$ or $H = (h_2, h_3)$.

The goal of the association studies is to relate the pair of haplotypes to disease phenotypes or traits. The simplest phenotype is the binary indicator for the disease status, which takes the value 1 if the individual is diseased and 0 otherwise. The diseased individuals may be further classified into several categories corresponding to different types of disease or varying degrees of severity. If the age of onset is likely to be genetically mediated, then it is desirable to use the age of onset as the phenotype. One may also be interested in disease-related traits, such as blood pressures.

The data on the disease phenotype may be gathered in a number of manners. The simplest approach is to obtain a random sample from the target population and to measure the phenotype of interest on every individual in the sample. Such studies are referred to as cross-sectional studies, which are feasible if the disease is relatively frequent or if one is only interested in some readily measured traits that are related to the disease. If one is interested in the age at the onset of a disease, then it is necessary to follow a cohort of individuals forward in time, in which case the phenotype, i.e., time to disease occurrence, may be censored. When the disease is relatively rare, it is more cost-effective to employ the case-control design, which collects data retrospectively on a sample of diseased individuals and on a separate sample of disease-free individuals. It is often desirable to collect data on environmental variables or covariates so as to investigate gene-environment interactions.

Let Y be the phenotype of interest, and $\mathbf{X}$ be the covariates. For cross-sectional and case-control studies, the association between Y and $(\mathbf{X}, H)$ is characterized by the conditional density of $Y = y$ given $H = (h_k, h_l)$ and $\mathbf{X} = \mathbf{x}$, denoted by $P_{\alpha,\boldsymbol{\beta},\xi}(y|\mathbf{x}, (h_k, h_l))$, where α per-

tains to the intercept(s), $\boldsymbol{\beta}$ to the regression effects and ξ to the nuisance parameters (e.g., variance and overdispersion parameters). There are considerable flexibilities in specifying the regression relationship. Suppose that h^* is the target haplotype of interest and there are no covariates. Then a linear predictor in the form of $\alpha + \beta I(h_k = h_l = h^*)$ pertains to a recessive model, $\alpha+\beta\{I(h_k = h^*)+I(h_l = h^*)-I(h_k = h_l = h^*)\}$ to a dominant model, $\alpha+\beta\{I(h_k = h^*)+I(h_l = h^*)\}$ to an additive model, and $\alpha + \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 I(h_k = h_l = h^*)$ to a codominant model, where $I(\cdot)$ is the indicator function. Clearly, the codominant model contains the other three models as special cases. A codominant model with gene-environment interactions has the following linear predictor

$$\begin{aligned}&\alpha + \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 I(h_k = h_l = h^*)\\ &+\boldsymbol{\beta}_3^T\mathbf{x} + \boldsymbol{\beta}_4^T\{I(h_k = h^*) + I(h_l = h^*)\}\mathbf{x} + \boldsymbol{\beta}_5^T I(h_k = h_l = h^*)\mathbf{x}. \qquad (1)\end{aligned}$$

Additional terms may be included so as to examine the effects of several haplotype configurations or to investigate the joint effects of multiple candidate genes.

Although we are interested in the effects of H and $\mathbf{X}$ on Y, we observe G instead of H. As mentioned earlier, G is the summation of the paired sequences in H. Thus, we have a regression problem with missing data in which the primary explanatory variable pertains to two ordered sequences of numbers from $\{0, 1\}$ but only the summation of the two sequences is observed. We assume that $\mathbf{X}$ is independent of H conditional on G and that $(1, \mathbf{X}^T)$ is linearly independent with positive probability.

Write $\pi_{kl} = P\{H = (h_k, h_l)\}$ and $\pi_k = P(h = h_k)$, $k, l = 1, \ldots, K$. As will be demonstrated in this paper, it is sometimes possible to make inference about haplotype effects without imposing any structures on $\{\pi_{kl}\}$, although the estimation of $\{\pi_k\}$ and the testing of no haplotype effects require some restrictions on $\{\pi_{kl}\}$. Under Hardy-Weinberg equilibrium,

$$\pi_{kl} = \pi_k\pi_l, \quad k, l = 1, \ldots, K. \qquad (2)$$

We consider two specific forms of departures from Hardy-Weinberg equilibrium:

$$\pi_{kl} = (1 - \rho)\pi_k\pi_l + \delta_{kl}\rho\pi_k, \qquad (3)$$

and

$$\pi_{kl} = \frac{(1 - \rho + \delta_{kl}\rho)\pi_k\pi_l}{1 - \rho + \rho\sum_{j=1}^K \pi_j^2}, \qquad (4)$$

where $0 \le \pi_k \le 1$, $\sum_{k=1}^K \pi_k = 1$, $\delta_{kk} = 1$ and $\delta_{kl} = 0$ $(k \ne l)$. In (3), ρ is called the inbreeding coefficient or fixation index (Weir, 1996, p. 93), and corresponds to Cohen (1960)'s kappa measure of agreement. Equation (4) creates disequilibrium by giving different fitness values to the homozygous

and heterozygous pairs (Niu et al, 2002). The denominator is a normalizing constant. Both (3) and (4) reduce to (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. Recently, Satten and Epstein (2004) considered equation (3) for the control population under the case-control design. We abuse the notation slightly in that $\{\pi_k\}$ in (4) do not pertain to the marginal distribution of H unless $\rho = 0$.

Let $\widetilde{h}$ denote a haplotype that differs from h only at one SNP. Write $\nabla_{\mathbf{x}} f(\mathbf{x}, y) = \partial f(\mathbf{x}, y)/\partial \mathbf{x}$. The lemma below states that, under equation (3) or (4), $\{\pi_k\}$ and ρ are identifiable from the data on G and the data on G provides positive information about these parameters.

Lemma 1. Assume that either equation (3) or (4) holds. The parameters $\{\pi_k\}$ and ρ are uniquely determined by the distribution of G. For non-degenerate distribution $\{\pi_k\}$, if there exist a constant μ and a vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)^T$ such that $\sum_{k=1}^K \nu_k = 0$ and $\mu \nabla_\rho \log P(G = g) + \sum_{k=1}^K \nu_k \nabla_{\pi_k} \log P(G = g) = 0$ for $g = 2h$, then $\mu = 0$ and $\boldsymbol{\nu} = \mathbf{0}$.

In the sequel, $\mathcal{G}$ denotes the set of all possible genotypes, and $\mathcal{S}(G)$ the set of haplotype pairs that are consistent with genotype G. We suppose that $\pi_k > 0$ for all $k = 1, \ldots, K$, where K is now interpreted as the total number of haplotypes that exist in the population. For any parameter $\boldsymbol{\theta}$, we will use $\boldsymbol{\theta}_0$ to denote its true value if the distinction is necessary. We assume that the true value of any Euclidean parameter belongs to the interior of a known compact set within the domain of $\boldsymbol{\theta}$. Lemma 1 and all the theorems are proved in the appendix.

2.2. Cross-Sectional Studies

There is a random sample of n individuals from the underlying population. The observable data consists of $(Y_i, \mathbf{X}_i, G_i)$, $i = 1, \ldots, n$. The trait Y can be discrete or continuous, univariate or multivariate. As stated in §2.1, the conditional density of Y given $\mathbf{X}$ and H is given by $P_{\alpha,\boldsymbol{\beta},\xi}(Y|\mathbf{X}, H)$. For a univariate trait, this regression model may take the form of a generalized linear model (McCullagh and Nelder, 1989) with the linear predictor given in (1). If the trait is measured repeatedly in a longitudinal study, then generalized linear mixed models (Diggle et al., 2002, Ch. 9) may be used. The following conditions are required for the estimation of $(\alpha, \boldsymbol{\beta}, \xi)$:

Condition 1. If $P_{\alpha,\boldsymbol{\beta},\xi}(Y|\mathbf{X}, H) = P_{\widetilde{\alpha},\widetilde{\boldsymbol{\beta}},\widetilde{\xi}}(Y|\mathbf{X}, H)$ for any $H = (h_k, h_k)$ and $H = (h_k, \widetilde{h}_k)$, $k = 1, \ldots, K$, then $\alpha = \widetilde{\alpha}, \boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$ and $\xi = \widetilde{\xi}$.

Condition 2. If there exists a constant vector $\boldsymbol{\nu}$ such that $\boldsymbol{\nu}^T \nabla_{\alpha,\boldsymbol{\beta},\xi} \log P_{\alpha,\boldsymbol{\beta},\xi}(Y|\mathbf{X}, H) = 0$ for $H = (h_k, h_k)$ and $H = (h_k, \widetilde{h}_k)$, then $\boldsymbol{\nu} = \mathbf{0}$.

Remark 1. Condition 1 ensures that the parameters of interest are identifiable from the genotype data. The linear independence of the score function stated in Condition 2 ensures the nonsingularity of the information matrix. The reason for considering $H = (h_k, h_k)$ and $H = (h_k, \widetilde{h}_k)$ is that these haplotype pairs can be inferred with certainty due to the unique decompositions of the corresponding genotypes $g = 2h_k$ and $g = h_k + \widetilde{h}_k$. All the commonly used regression models, particularly generalized linear (mixed) models with linear predictors in the form of (1), satisfy Conditions 1 and 2.

We show in Appendix A.2.1 that it is possible to estimate the regression parameters without imposing any structure on the joint distribution of H. However, the estimation requires the knowledge about whether or not the dominant effects exist. Specifically, if there are no dominant effects, then only $(\alpha, \boldsymbol{\beta}, \xi)$ and $P(G = g)$ are identifiable; otherwise, $(\alpha, \boldsymbol{\beta}, \xi)$, $P(G = g)$ and $P(H = (h^*, g - h^*))$ are identifiable. If either equation (3) or (4) holds, then it follows from Lemma 1 and Condition 1 that all the parameters are identifiable regardless of the genetic mechanism. Denote the joint distribution of H by $P_{\boldsymbol{\gamma}}(H = (h_k, h_l))$, where $\boldsymbol{\gamma}$ consists of the identifiable parameters in the distribution of H. Under equation (3) or (4), $\boldsymbol{\gamma} = (\rho, \pi_1, \ldots, \pi_K)^T$. When the distribution of H is unspecified, $\boldsymbol{\gamma}$ pertains to the aspects of the distribution of H that are identifiable.

Write $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \xi)$. The likelihood for $\boldsymbol{\theta}$ based on the cross-sectional data is proportional to

$$L_n(\boldsymbol{\theta}) \equiv \prod_{i=1}^{n} \prod_{g \in \mathcal{G}} \{m_g(Y_i, \mathbf{X}_i; \boldsymbol{\theta})\}^{I(G_i = g)}, \tag{5}$$

where

$$m_g(y, \mathbf{x}; \boldsymbol{\theta}) = \sum_{(h_k, h_l) \in \mathcal{S}(g)} P_{\alpha, \boldsymbol{\beta}, \xi}(y | \mathbf{x}, (h_k, h_l)) P_{\boldsymbol{\gamma}}(h_k, h_l).$$

The maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\theta}}$ can be obtained by maximizing (5) via the Newton-Raphson algorithm or an optimization algorithm. It is generally more efficient to use the expectation-maximization (EM) algorithm (Dempster et al., 1977), especially when the distribution of H satisfies equation (3) with $\rho \geq 0$; see Appendix A.2.2 for detail.

By the classical likelihood theory, we can show that $\widehat{\boldsymbol{\theta}}$ is consistent, asymptotically normal and asymptotically efficient under Conditions 1 and 2 and the following condition:

Condition 3. If there exists a constant vector $\boldsymbol{\nu}$ such that $\boldsymbol{\nu}^T \nabla_{\boldsymbol{\theta}} \log m_g(Y, \mathbf{X}; \boldsymbol{\theta}_0) = 0$, then $\boldsymbol{\nu} = \mathbf{0}$.

Remark 2. Condition 3 ensures the nonsingularity of the information matrix. This condition can be easily verified when the joint distribution of H is unspecified, and is implied by Lemma 1 and Condition 2 when the distribution satisfies equation (3) or (4).

2.3. Case-Control Studies With Known Population Totals

We consider case-control data supplemented by information on population totals (Scott and Wild, 1997). There is a finite population of N individuals which is regarded as a random sample from the joint distribution of $(Y, \mathbf{X}, H)$, where Y is a categorical response variable. All that is known about the finite population is the total number of individuals in each category of $Y = y$. A sample of size n stratified on the disease status is drawn from the finite population, and the values of $\mathbf{X}$ and G are recorded for each sampled individual. The supplementary information on population totals is often available from hospital records, cancer registries and official statistics. If a case-control sample is drawn from a cohort study, then the cohort serves as the finite population. The observable data consists of $(Y_i, R_i, R_i\mathbf{X}_i, R_iG_i), i = 1, \ldots, N$, where R_i indicates, by the values 1 versus 0, whether or not the ith individual in the finite population is selected into the case-control sample.

The association between Y and $(\mathbf{X}, H)$ is characterized by $P_{\alpha,\boldsymbol{\beta},\xi}(Y|\mathbf{X}, H)$, where α, $\boldsymbol{\beta}$ and ξ pertain to the intercept(s), regression effects and overdispersion parameters (McCullagh and Nelder 1989), respectively. In the case of a binary response variable, important examples of $P_{\alpha,\boldsymbol{\beta},\xi}(Y|\mathbf{X}, H)$ include the logistic, probit and complementary log-log regression models. When there are more than two categories, examples include the proportional odds model, the multivariate probit and multivariate logistic regression models. Since the data associated with $R_i = 1$ yields the same form of likelihood as that of a cross-sectional study and the data associated $R_i = 0$ yields a missing-data likelihood, all the identifiability results stated in §2.2 apply to the current setting. We again write $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \xi, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ consists of the identifiable parameters in the distribution of H.

Let $F_g(\cdot)$ be the cumulative distribution function of $\mathbf{X}$ given $G = g$, and let $f_g(\mathbf{x})$ be the density of $F_g(\mathbf{x})$ with respect to a dominating measure $\mu(\mathbf{x})$. Note that $F_g(\cdot)$ is infinite-dimensional if $\mathbf{X}$ has continuous components. The joint density of $(Y = y, G = g, \mathbf{X} = \mathbf{x})$ is $m_g(y, \mathbf{x}; \boldsymbol{\theta})f_g(\mathbf{x})$. The likelihood concerning $\boldsymbol{\theta}$ and $\{F_g\}$ takes the form

$$L_n(\boldsymbol{\theta}, \{F_g\}) = \prod_{i=1}^{N} \left[\prod_{g \in \mathcal{G}} \{m_g(Y_i, \mathbf{X}_i; \boldsymbol{\theta}) f_g(\mathbf{X}_i)\}^{I(G_i=g)} \right]^{R_i} \left[\sum_{g \in \mathcal{G}} \int m_g(Y_i, \mathbf{x}; \boldsymbol{\theta}) dF_g(\mathbf{x}) \right]^{1-R_i}. \quad (6)$$

Unlike the likelihood for the cross-sectional design given in (5), the density functions of $\mathbf{X}$ given G cannot be factored out of the likelihood given in (6) and thus cannot be omitted from the likelihood.

We maximize (6) to obtain the MLEs $\widehat{\boldsymbol{\theta}}$ and $\{\widehat{F}_g(\cdot)\}$. The latter is an empirical function with point masses at the observed $\mathbf{X}_i$ such that $G_i = g$ and $R_i = 1$. The maximization can be carried out via the Newton-Raphson, profile-likelihood or large-scale optimization methods. An alternative way of calculating the MLEs is via the EM algorithm described in Appendix A.3.1.

We impose the following regularity condition and state the asymptotic results in Theorem 1.

Condition 4. For any $g \in \mathcal{G}$, $f_g(\mathbf{x})$ is positive in its support and continuously differentiable with respect to a suitable measure.

Theorem 1. Under Conditions 1–4, $\widehat{\boldsymbol{\theta}}$ and $\{\widehat{F}_g(\cdot)\}$ are consistent in that $|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sup_{\mathbf{x},g} |\widehat{F}_g(\mathbf{x}) - F_g(\mathbf{x})| \to 0$ 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.

Let $pl_n(\boldsymbol{\theta})$ be the profile log-likelihood for $\boldsymbol{\theta}$, i.e., $pl_n(\boldsymbol{\theta}) = \max_{\{F_g\}} \log L_n(\boldsymbol{\theta}, \{F_g\})$. Then the (s,t)th element of the covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by $-\epsilon_n^{-2}\{pl_n(\widehat{\boldsymbol{\theta}} + \epsilon_n \mathbf{e}_s + \epsilon_n \mathbf{e}_t) - pl_n(\widehat{\boldsymbol{\theta}} + \epsilon_n \mathbf{e}_s - \epsilon_n \mathbf{e}_t) - pl_n(\widehat{\boldsymbol{\theta}} - \epsilon_n \mathbf{e}_s + \epsilon_n \mathbf{e}_t) + pl_n(\widehat{\boldsymbol{\theta}})\}$, where ϵ_n is a constant of the order $n^{-1/2}$, and $\mathbf{e}_s$ and $\mathbf{e}_t$ are the sth and tth canonical vectors, respectively. The function $pl_n(\boldsymbol{\theta})$ can be calculated via the EM algorithm by holding $\boldsymbol{\theta}$ constant in both the E-step and the M-step.

Remark 3. If N is much larger than n or if the population frequencies rather than the totals are known, then we maximize $\prod_{i=1}^n \prod_{g\in\mathcal{G}} \{m_g(Y_i, \mathbf{X}_i; \boldsymbol{\theta}) f_g(\mathbf{X}_i)\}^{I(G_i=g)}$ subject to the constraints that $\sum_{g\in\mathcal{G}} \int m_g(y, \mathbf{x}; \boldsymbol{\theta}) dF_g(\mathbf{x}) = p_y$, where p_y is the population frequency of $Y = y$. The resultant estimator of $\boldsymbol{\theta}_0$ is consistent, asymptotically normal and asymptotically efficient. The results in this section can be extended straightforwardly to accommodate stratifications on covariates.

2.4. Case-Control Studies With Unknown Population Totals

We consider the classical case-control design, which measures $\mathbf{X}$ and G on n_1 cases ($Y = 1$) and n_0 controls ($Y = 0$) and which requires no knowledge about the finite population. With the notation introduced in the previous section, the likelihood contribution from one individual takes the form

$$RL(\boldsymbol{\theta}, \{F_g\}) = \frac{\prod_{g\in\mathcal{G}} \{m_g(y, \mathbf{X}; \boldsymbol{\theta}) f_g(\mathbf{X})\}^{I(G=g)}}{\sum_{g\in\mathcal{G}} \int m_g(y, \mathbf{x}; \boldsymbol{\theta}) dF_g(\mathbf{x})}, \tag{7}$$

where we use y instead of Y to emphasize that y is not random.

Define

$$f_g^\dagger(\mathbf{x}) = \frac{m_g(0, \mathbf{x}; \boldsymbol{\theta}) f_g(\mathbf{x})}{\int m_g(0, \mathbf{x}; \boldsymbol{\theta}) dF_g(\mathbf{x})}, \quad q_g = \frac{\int m_g(0, \mathbf{x}; \boldsymbol{\theta}) dF_g(\mathbf{x})}{\sum_{\tilde{g}\in\mathcal{G}} \int m_{\tilde{g}}(0, \mathbf{x}; \boldsymbol{\theta}) dF_{\tilde{g}}(\mathbf{x})}.$$

Clearly, $f_g^\dagger(\mathbf{x})$ is the conditional density of $\mathbf{X}$ given $G = g$ and $Y = 0$, and q_g is the conditional probability of $G = g$ given $Y = 0$. Let g_0 and $\mathbf{x}_0$ be some specific values of G and $\mathbf{X}$. Write $F_g^\dagger(\mathbf{x}) = \int_0^{\mathbf{x}} f_g^\dagger(s) d\mu(s)$. We can express (7) as

$$RL(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\}) = \frac{\prod_{g\in\mathcal{G}} \left\{\eta(y, \mathbf{X}, g; \boldsymbol{\theta}) f_g^\dagger(\mathbf{x}) q_g\right\}^{I(G=g)}}{\sum_{g\in\mathcal{G}} q_g \left\{\int \eta(y, \mathbf{x}, g; \boldsymbol{\theta}) dF_g^\dagger(\mathbf{x})\right\}}, \tag{8}$$

where

$$\eta(y, \mathbf{x}, g; \boldsymbol{\theta}) = \frac{m_g(y, \mathbf{x}; \boldsymbol{\theta}) m_{g_0}(0, \mathbf{x}_0; \boldsymbol{\theta})}{m_g(0, \mathbf{x}; \boldsymbol{\theta}) m_{g_0}(y, \mathbf{x}_0; \boldsymbol{\theta})}.$$

We refer to η as the generalized odds ratio (Liang and Qin, 2000), which reduces to the ordinary odds ratio when $\mathcal{S}(g)$ is a singleton.

Remark 4. The parameter q_g is a functional of $f_g^\dagger$ and $\boldsymbol{\theta}$ because $\int m_g(0, \mathbf{x}; \boldsymbol{\theta}) dF_g(\mathbf{x}) = \left\{ \int m_g^{-1}(0, \mathbf{x}; \boldsymbol{\theta}) dF_g^\dagger(\mathbf{x}) \right\}^{-1}$. This constraint makes it very difficult to study the identifiability of the parameters. Thus, we treat q_g as a free parameter in our development.

For traditional case-control data, the odds ratio is identifiable (whereas the intercept is not) and its MLE can be obtained by maximizing the prospective likelihood (Prentice and Pyke, 1979). Similar results hold when the exposure is measured with error (Roeder, Carroll and Lindsay, 1996); however, the distribution of the measurement error needs to be estimated from a validation set or an external source. With unphased genotype data, identifiability is much more delicate. We show in Appendix A.4.1 that the components of $\boldsymbol{\theta}$ that are identifiable from the retrospective likelihood are exactly those that are identifiable from the generalized odds ratio. Thus, we assume that the generalized odds ratio only depends on a set of identifiable parameters, still denoted by $\boldsymbol{\theta}$; otherwise, the inference is not tractable. For the logistic link function with linear predictor (1), we show in Appendix A.4.2 that if there are no dominant effects, then $\boldsymbol{\theta}$ only consists of $\boldsymbol{\beta}$; if there are no covariate effects but there exists a dominant main effect, then $\boldsymbol{\beta}$ is identifiable and $P(H = (h^*, g - h^*))/P(G = g)$ is identifiable up to a scalar constant; if the dominant effect depends on a continuous covariate or if the dominant main effect and the main effect of a continuous covariate are non-zero, then $\boldsymbol{\theta}$ consists of α, $\boldsymbol{\beta}$ and $P(H = (h^*, g - h^*))/P(G = g)$. For the probit and complementary log-log link functions, we show in Appendix A.4.3 that if there are dominant effects and at least one continuous covariate has an effect, then $\boldsymbol{\theta}$ consists of α, $\boldsymbol{\beta}$ and $P(H = (h^*, g - h^*))/P(G = g)$.

We maximize the product of (8) over the $n \equiv n_1 + n_0$ individuals in the case-control sample to produce the MLEs $\widehat{\boldsymbol{\theta}}$, $\{\widehat{F}_g^\dagger(\cdot)\}$ and $\{\widehat{q}_g\}$. Although $\{F_g^\dagger(\cdot)\}$ are high-dimensional, we show in Appendix A.4.4 that $\widehat{\boldsymbol{\theta}}$ can be obtained by profiling a likelihood function over a scalar nuisance parameter.

To state the asymptotic properties of the MLEs, we impose the following conditions:

Condition 5. If there exists a vector $\mathbf{v}$ such that $\mathbf{v}^T \nabla_{\boldsymbol{\theta}} \log \eta(1, \mathbf{x}, g; \boldsymbol{\theta})$ is a constant with probability 1, then $\mathbf{v} = \mathbf{0}$.

Condition 6. The function $f_g^\dagger$ is positive in its support and continuously differentiable.

Condition 7. The fraction $n_1/n \to \varrho \in (0, 1)$.

Remark 5. Condition 5 implies the nonsingularity of the information matrix for $\boldsymbol{\theta}_0$, and can be shown to hold for the logistic, probit and complementary log-log link functions. Condition 7 ensures that there are both cases and controls in the sample.

Theorem 2. Under Conditions 5–7, $|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sup_g |\widehat{q}_g - q_g| + \sup_{\mathbf{x},g} |\widehat{F}_g^{\dagger}(\mathbf{x}) - F_g^{\dagger}(\mathbf{x})| \to 0$ almost surely. In addition, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a normal random vector whose covariance matrix attains the semiparametric efficiency bound.

In most case-control studies, the disease is (relatively) rare. When the disease is rare, considerable simplicity arises because of the following approximation for the logistic regression model:

$$P_{\alpha,\boldsymbol{\beta}}(Y|\mathbf{X}, H) \approx \exp\{Y(\alpha + \boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}, H))\},$$

where $\mathcal{Z}(\mathbf{X}, H)$ is a specific function of $\mathbf{X}$ and H. We assume that either equation (3) or (4) holds. The likelihood based on $(\mathbf{X}_i, G_i, y_i)$, $i = 1, \ldots, n$, can be approximated by

$$\widetilde{L}_n(\boldsymbol{\theta}, \{F_g\}) = \prod_{i=1}^{n} \left(\frac{\prod_{g\in\mathcal{G}} [f_g(\mathbf{X}_i) \sum_{(h_k,h_l)\in\mathcal{S}(g)} \exp\{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i, h_k, h_l)\} P_{\boldsymbol{\gamma}}(h_k, h_l)]^{I(G_i=g)}}{\sum_{g\in\mathcal{G}} \int_{\mathbf{x}} \sum_{(h_k,h_l)\in\mathcal{S}(g)} \exp\{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{x}, h_k, h_l)\} P_{\boldsymbol{\gamma}}(h_k, h_l) dF_g(\mathbf{x})} \right)^{y_i}$$

$$\times \left[\prod_{g\in\mathcal{G}} \left\{ f_g(\mathbf{X}_i) \sum_{(h_k,h_l)\in\mathcal{S}(g)} P_{\boldsymbol{\gamma}}(h_k, h_l) \right\}^{I(G_i=g)} \right]^{1-y_i}. \tag{9}$$

We impose the following condition:

Condition 8. If $\alpha + \boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}, H) = \widetilde{\alpha} + \widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X}, H)$ for $H = (h_k, h_k)$ and $H = (h_k, \widetilde{h}_k)$, then $\alpha = \widetilde{\alpha}$ and $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$.

This condition is similar to Condition 1 stated in §2.2, and it holds for the codominant model. Under this condition, it follows from Lemma 1 that no two sets of parameters can give the same likelihood with probability 1. Thus, the maximizer of (9), denoted by $(\widehat{\boldsymbol{\theta}}, \{\widehat{F}_g\})$, is locally unique. We show in Appendix A.4.5 that $\widehat{\boldsymbol{\theta}}$ can be easily obtained by profiling over a small number of parameters.

To derive the asymptotic properties, we provide a mathematical definition of rare disease:

Condition 9. For $i = 1, \ldots, n$, the conditional distribution of Y_i given $(\mathbf{X}_i, H_i)$ satisfies that $P(Y_i = 1|\mathbf{X}_i, H_i) = a_n \exp\{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X}_i, H_i)\}/[1 + a_n \exp\{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X}_i, H_i)\}]$, where $a_n = o(n^{-1/2})$.

Theorem 3. Under Conditions 6–9, $|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sup_{\mathbf{x},g} |\widehat{F}_g(\mathbf{x}) - F_g(\mathbf{x})| \to^{P_n} 0$, where P_n is the probability measure given by Condition 9. Furthermore, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a normal random vector whose covariance matrix achieves the semiparametric efficiency bound.

2.5. Cohort Studies

In a cohort study, Y represents the time to disease occurrence, which is subject to right censorship by C. The data consists 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 relate Y_i to $(\mathbf{X}_i, H_i)$ through a class of semiparametric linear transformation models

$$\Gamma(Y_i) = -\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i, H_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{10}$$

where Γ is an unknown increasing function, $\mathcal{Z}(\mathbf{X}, H)$ is a known function of $\mathbf{X}$ and H, and the ϵ_i are independent errors with a known distribution function F. We may rewrite (10) as

$$P(Y_i \leq t | \mathbf{X}_i, H_i) = Q(\Lambda(t) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i, H_i)}),$$

where $\Lambda(t) = e^{\Gamma(t)}$, and $Q(x) = F(\log x)$ $(x > 0)$. The choices of the extreme-value and standard logistic distributions for F, or equivalently $Q(x) = 1 - e^{-x}$ and $Q(x) = 1 - (1 + x)^{-1}$, yield the proportional hazards model and the proportional odds model (Pettitt, 1984), respectively.

We impose Condition 8. Under this condition, $\boldsymbol{\beta}$ and $\Lambda(\cdot)$ are identifiable from the observable data. The identifiability of the distribution of H is the same as in the case of cross-sectional studies. Under equation (3) or (4) and Condition 8, all the parameters including $\boldsymbol{\beta}$, $\Lambda(\cdot)$ and $\boldsymbol{\gamma}$ are identifiable. This is shown in Appendix A.5.1.

The following assumption on censoring is required in the construction of the likelihood:

Condition 10. Conditional on $\mathbf{X}$ and G, the censoring time C is independent of Y and H.

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

$$L_n(\boldsymbol{\theta}, \Lambda) = \prod_{i=1}^{n} \Bigg[\sum_{(h_k, h_l) \in \mathcal{S}(G_i)} \left\{ \dot{\Lambda}(\widetilde{Y}_i) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i, (h_k, h_l))} \dot{Q}(\Lambda(\widetilde{Y}_i) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i, (h_k, h_l))}) \right\}^{\Delta_i}$$

$$\times \left\{ 1 - Q(\Lambda(\widetilde{Y}_i) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X}_i, (h_k, h_l))}) \right\}^{1 - \Delta_i} P_{\boldsymbol{\gamma}}(h_k, h_l) \Bigg]. \tag{11}$$

Here and in the sequel, $\dot{f}(x) = df(x)/dx$ and $\ddot{f}(x) = d^2 f(x)/dx^2$. Like (6), (8) and (9), this likelihood involves infinite-dimensional parameters. If Λ is restricted to be absolutely continuous, then as in the case of density estimation, there is no maximizer of this likelihood. Thus, we relax Λ to be right-continuous and replace $\dot{\Lambda}(\widetilde{Y}_i)$ in (11) by the jump size of Λ at $\widetilde{Y}_i$. By the arguments of Zeng et al. (2004), the resultant MLE, denoted by $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$, exists, and $\widehat{\Lambda}$ is a step function with jumps only at the observed $\widetilde{Y}_i$ for which $\Delta_i = 1$. The maximization can be carried out through an optimization algorithm. Furthermore, the covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by the profile likelihood method, as discussed in Zeng et al. (2004).

Lin (2004) considered the special case of the proportional hazards model under condition (2), and provided an EM algorithm for obtaining the MLEs. We can modify that algorithm to accommodate Hardy-Weinberg disequilibrium along the lines of Appendix A.2.2. In addition, the EM-algorithm can be used to evaluate the profile likelihood.

We assume the following regularity conditions for the asymptotic results:

Condition 11. There exists some positive constant δ_0 such that $P(C_i \geq \tau|\mathbf{X}_i, G_i) = P(C_i = \tau|\mathbf{X}_i, G_i) \geq \delta_0$ almost surely, where τ corresponds to the end of the study.

Condition 12. The true value $\Lambda_0(t)$ of $\Lambda(t)$ is a strictly increasing function in $[0, \tau]$ and is continuously differentiable. In addition, $\Lambda_0(0) = 0$, $\Lambda_0(\tau) < \infty$ and $\dot{\Lambda}_0(0) > 0$.

Theorem 4. Under Conditions 8 and 10–12, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{\Lambda} - \Lambda_0)$ converges weakly to a 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, $\widehat{\boldsymbol{\theta}}$ is asymptotically efficient.

3. SIMULATION STUDIES

We used Monte Carlo simulation to evaluate the proposed methods in realistic settings. We considered the 5 SNPs on chromosome 22 from the Finland-United States Investigation of NIDDM Genetics (FUSION) Study described in the next section. We obtained the π_k from the frequencies shown in Table 1 by assuming 7% disease rate, and generated haplotypes under equation (3) with $\rho = 0.05$. The R_h^2 in Table 1 is Stram et al. (2003)'s measures of haplotype certainty. We focused on $h^* = (0, 1, 1, 0, 0)$, and considered case-control and cohort studies.

Table 1. Observed Haplotype Frequencies in the FUSION Study

	Frequencies				Frequencies		
Haplotype	Controls	Cases	R_h^2	Haplotype	Controls	Cases	R_h^2
00011	.0042	.0066	.388	10010	$< 10^{-4}$	.0012	.500
00100	.0035	.0034	.336	10011	.3573	.2883	.727
00110	.0018	.0007	.377	10100	.0521	.0597	.402
01011	.1292	.1344	.592	10110	.0317	.0318	.554
01100	.2514	.3183	.738	11011	.1392	.1290	.560
01101	.0012	$< 10^{-4}$	.450	11100	.0109	.0092	.266
01110	$< 10^{-4}$	.0045	.499	11110	$< 10^{-4}$	.0014	$< 10^{-4}$
01111	.0019	$< 10^{-4}$	.325	11111	.0020	$< 10^{-4}$	.338
10000	.0136	.0114	.456				

For the cohort studies, we generated ages of onset from the proportional hazards model

$$\lambda\{t|x, (h_k, h_l)\} = 2t \exp\left[\beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 x + \beta_3\{I(h_k = h^*) + I(h_l = h^*)\}x\right],$$

where X is a Bernoulli variable with $P(X = 1) = 0.2$ that is independent of H. The censoring times were generated from the uniform $(0, \tau)$ distribution, where τ was chosen to yield approximately 250, 500 and 1000 cases under $n = 5000$. We let $\beta_1 = \beta_2 = 0.25$ and varied β_3 from -0.5 to 0.5.

As shown in Table 2, the maximum likelihood estimator is virtually unbiased, the likelihood ratio test has proper type I error, and the confidence interval has reasonable coverage. Additional simulation studies revealed that the proposed methods also perform well for making inference about other parameters and under other genetic models.

Table 2. Simulation Results for the Haplotype-Environment Interactions in Cohort Studies

β_3	cases	Bias	SE	CP	Power
0	250	–.010	.232	.949	.051
	500	–.005	.157	.953	.047
	1000	–.003	.114	.954	.045
–.25	250	–.014	.256	.950	.190
	500	–.008	.172	.949	.334
	1000	–.004	.122	.952	.554
–.5	250	–.014	.287	.949	.500
	500	–.011	.192	.949	.763
	1000	–.004	.133	.950	.883
.25	250	–.007	.216	.947	.207
	500	–.002	.146	.953	.395
	1000	–.001	.109	.954	.614
.5	250	–.003	.204	.943	.693
	500	–.001	.140	.951	.940
	1000	–.001	.105	.952	.998

NOTE: Bias and SE are the bias and standard error of $\widehat{\beta}_3$. CP is the coverage probability of the 95% confidence interval for β_3. Power pertains to the 0.05-level likelihood ratio test of $H_0 : \beta_3 = 0$. Each entry is based on 5,000 replicates.

For the case-control studies, we used the same distributions of H and X and considered the same h^* as the cohort studies. We generated disease incidence from the logistic regression model:

$$\text{logit}P\{Y = 1|x, (h_k, h_l)\} = \alpha + \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 x + \beta_3\{I(h_k = h^*) + I(h_l = h^*)\}x. \tag{12}$$

For making inference on β_1, we set $\beta_2 = \beta_3 = 0.25$ and varied β_1 from -0.5 to 0.5; for making inference on β_3, we set $\beta_1 = \beta_2 = 0.25$ and varied β_3 from -0.5 to 0.5. We chose $\alpha = -3$ or

-4, yielding disease rates between 1.6% and 7%. We let $n_1 = n_0 = 500$ or 1000. We considered both the situations of known and unknown population totals, N being 15 and 30 times of n under $\alpha = -3$ and -4, respectively. For known population totals, we used the EM algorithm described in Appendix A.3.1 and evaluated the inference procedures based on the likelihood ratio statistic. For unknown population totals, we used the profile-likelihood method for rare diseases described in Appendix A.4.5 and set the $\widehat{\pi}_k$ less than $2/n$ to 0 to improve numerical stability. The results for β_1 and β_3 are displayed in Tables 3 and 4, respectively.

Table 3. Simulation Results for the Main Effects of the Haplotype in Case-Control Studies

			Known Totals				Unknown Totals				
$n_1 = n_0$	α	β_1	Bias	SE	CP	Power	Bias	SE	SEE	CP	Power
500	–3	–.5	–.003	.117	.952	.987	.019	.121	.124	.951	.979
		–.25	–.002	.109	.954	.587	.014	.112	.117	.960	.525
		0	–.001	.104	.951	.049	.009	.109	.112	.955	.045
		.25	–.001	.102	.950	.641	.002	.105	.108	.961	.646
		.5	.000	.099	.948	.996	-.005	.103	.106	.958	.998
	–4	–.5	.001	.112	.954	.987	.022	.119	.124	.951	.977
		–.25	–.002	.104	.955	.574	.013	.114	.117	.952	.529
		0	.002	.101	.950	.050	.004	.109	.112	.953	.047
		.25	.003	.094	.956	.661	-.003	.103	.108	.959	.640
		.5	–.000	.094	.950	.999	-.009	.102	.105	.956	.997
1000	–3	–.5	.000	.081	.956	1.00	.005	.087	.087	.949	1.00
		–.25	–.000	.074	.960	.871	.005	.081	.082	.948	.853
		0	–.001	.073	.951	.049	.005	.077	.077	.954	.046
		.25	–.001	.071	.953	.898	.004	.075	.076	.948	.920
		.5	–.001	.070	.953	1.00	.003	.075	.075	.946	1.00
	–4	–.5	–.000	.079	.948	1.00	.005	.087	.088	.947	1.00
		–.25	–.001	.072	.960	.873	.005	.081	.083	.954	.847
		0	.000	.070	.951	.043	.002	.079	.079	.949	.051
		.25	–.001	.066	.960	.914	.000	.074	.076	.955	.909
		.5	–.001	.066	.958	1.00	-.002	.073	.074	.954	1.00

NOTE: Bias and SE are the bias and standard error of $\widehat{\beta}_1$. SEE is the mean of the standard error estimator for $\widehat{\beta}_1$. CP is the coverage probability of the 95% confidence interval for β_1. Power pertains to the 0.05-level test of $H_0 : \beta_1 = 0$. Each entry is based on 5,000 replicates.

Table 4. Simulation Results for the Haplotype-Environment Interactions in Case-Control Studies

			Known Totals				Unknown Totals				
$n_1 = n_0$	α	β_3	Bias	SE	CP	Power	Bias	SE	SEE	CP	Power
500	–3	–.5	–.008	.205	.949	.729	.030	.187	.195	.953	.692
		–.25	–.002	.186	.949	.271	.016	.169	.176	.961	.244
		0	–.001	.173	.946	.054	-.006	.155	.162	.963	.037
		.25	.002	.165	.949	.334	-.038	.144	.151	.958	.287
		.5	.006	.161	.947	.885	-.088	.138	.143	.915	.831
	–4	–.5	–.009	.198	.950	.763	.012	.194	.195	.950	.720
		–.25	–.005	.181	.949	.309	.006	.172	.176	.953	.264
		0	–.002	.168	.945	.055	-.007	.156	.161	.956	.044
		.25	–.001	.157	.944	.370	-.022	.146	.149	.948	.333
		.5	.001	.148	.945	.926	-.047	.136	.141	.945	.904
1000	–3	–.5	–.004	.147	.943	.953	.027	.134	.136	.950	.953
		–.25	–.003	.133	.946	.493	.013	.122	.123	.949	.477
		0	–.001	.123	.951	.049	-.005	.114	.113	.948	.052
		.25	.002	.115	.957	.590	-.034	.107	.106	.934	.535
		.5	.002	.117	.947	.994	-.080	.102	.101	.870	.986
	–4	–.5	–.005	.140	.945	.969	.010	.137	.136	.949	.965
		–.25	–.002	.126	.947	.535	.005	.124	.123	.951	.505
		0	–.003	.115	.956	.044	-.004	.113	.113	.947	.053
		.25	–.000	.108	.949	.626	-.016	.104	.105	.952	.601
		.5	.002	.105	.949	.998	-.037	.099	.099	.937	.995

NOTE: Bias and SE are the bias and standard error of $\widehat{\beta}_3$. SEE is the mean of the standard error estimator for $\widehat{\beta}_3$. CP is the coverage probability of the 95% confidence interval for β_3. Power pertains to the 0.05-level test of $H_0 : \beta_3 = 0$. Each entry is based on 5,000 replicates.

For known population totals, the proposed estimators are virtually unbiased, and the likelihood ratio statistics yield proper tests and confidence intervals. For unknown population totals, $\widehat{\beta}_1$ has little bias, especially for large n, while $\widehat{\beta}_3$ tends to be slightly biased downward; the variance estimators are fairly accurate, and the corresponding confidence intervals have reasonable coverage probabilities except for $\{\alpha = -3, \beta_3 = 0.5\}$. The method with known population totals yields slightly higher power than the method with unknown population totals.

All the aforementioned results pertain to haplotype 01100, which has a relatively high frequency and a large value of R_h^2; the covariate is binary, and ρ is 0.05, which is relatively large. Additional simulation studies showed that the above conclusions continue to hold for other haplotypes, other values of ρ and continuous covariates. Table 5 reports some results for haplotype 10100, which has

a frequency of about 5% and R_h^2 of 0.4. We generated disease incidence from the logistic regression model:

$$\text{logit}P\{Y=1|X_1,X_2,(h_k,h_l)\}=\alpha+\beta_h\{I(h_k=h^*)+I(h_l=h^*)\}$$

$$+\beta_{x1}X_1+\beta_{x_2}X_2+\beta_{hx_2}\{I(h_k=h^*)+I(h_l=h^*)\}X_2,$$

where $h^*=(10100)$, X_1 is Bernoulli with 0.2 success probability, and X_2 is uniform(0,1). We set $\rho=0.01$, $\alpha=-3.7$, $\beta_h=0$, and $\beta_{x_1}=\beta_{x_2}=-\beta_{x_2h}=0.5$, yielding an overall disease rate of 7%. We assumed unknown population totals and used the profile-likelihood method for rare diseases described in Appendix A.4.5. The method performs remarkly well.

Table 5. Simulation Results for Haplotype 10100 in Case-Control Studies

$n_1=n_0$	Parameter	True value	Bias	SE	SEE	CP	Power
500	β_h	0	-.030	.400	.401	.957	.043
	β_{x_1}	0.5	.002	.151	.152	.951	.917
	β_{x_2}	0.5	.001	.228	.230	.953	.584
	β_{x_2h}	-0.5	.015	.641	.644	.956	.118
1000	β_h	0	-.017	.275	.277	.953	.047
	β_{x_1}	0.5	.002	.107	.107	.954	.997
	β_{x_2}	0.5	.000	.162	.161	.950	.871
	β_{x_2h}	-0.5	.012	.441	.443	.950	.198

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 95% confidence interval. Power pertains to the 0.05-level test of zero parameter value. Each entry is based on 5,000 replicates.

4. APPLICATION TO THE FUSION STUDY

Type 2 diabetes or NIDDM is a complex disease characterized by resistance of peripheral tissues to insulin and a deficiency of insulin secretion. Approximately 7% of adults in developed countries suffer from the disease. The FUSION study is a major effort to map and clone genetic variants that predispose to Type 2 diabetes (Valle et al., 1998). We consider a subset of data from this study.

A total of 796 cases and 415 controls were genotyped at five SNPs in a putative susceptibility region on chromosome 22, 131 cases and 82 controls having missing genotype information for at least one SNP. If G_i is missing, the set $\mathcal{S}(G_i)$ is enlarged accordingly in the analysis. Table 1 displays the estimated haplotype frequencies under equation (3) separated by the cases and controls, along with Stram et al. (2003)'s R_h^2 for the controls. We estimated ρ at 0.0002 for controls and 0.03 for cases.

We use the method based on (9) to estimate the effects of the haplotypes whose observed frequencies in the controls are greater than 2%. As shown in Table 6, the results are significant for the two most common haplotypes: haplotype 01100 increases the risk of disease whereas haplotpye 10011 is protective against diabetes. Epstein and Satten (2003) also reported the estimates for these two haplotypes, which agree with our numbers. Although they did not report standard error estimates, their confidence intervals are similar to those based on Table 6. The results under the codominant model as well as the calculations of the Akaike (1985) information criterion (AIC) suggest that the additive model fits the data the best for both haplotypes 01100 and 10011.

Table 6. Estimates of Haplotype Effects Under Various Genetic Models for the FUSION Study

	Recessive	Dominant	Additive	Codominant model	
Haplotype	model	model	model	Additive	Recessive
01011	.319 (.266)	–.017 (.133)	.056 (.128)	.005 (.135)	.315 (.283)
01100	.318 (.143)	.268 (.107)	.352 (.097)	.328 (.111)	.070 (.163)
10011	–.225 (.150)	–.330 (.105)	–.332 (.092)	–.352 (.105)	.063 (.173)
10100	–1.030 (1.014)	.224 (.194)	.141 (.189)	.196 (.192)	–1.156 (1.021)
10110	.856 (.743)	–.014 (.241)	.052 (.241)	.008 (.245)	.851 (.762)
11011	–.221 (.317)	–.088 (.132)	–.120 (.126)	–.101 (.132)	–.148 (.332)

NOTE: Standard error estimates are shown in parentheses.

The FUSION investigators are currently exploring gene-environment interactions on chromosome 22, so that the covariate information is confidential at this stage. To illustrate our method for detecting gene-environment interactions, we artificially created a binary covariate X by setting $X = 1$ for the first 600 individuals in the data set. Under the additive genetic model for haplotype 01100, the estimate of the interaction is 0.047 with an estimated standard error of 0.110. For further illustration, we generated a binary covariate from the conditional distribution of X given Y and G under model (12) with $\alpha = -3.7$, $\beta_1 = 0.32$ and $\beta_2 = 0.25$. Based on 5000 replicates, the power for testing $H_0 : \beta_3 = 0$ is estimated at 0.053, 0.479 or 0.974 under $\beta_3 = 0$, 0.25 or 0.5.

5. DISCUSSION

Inferring haplotype-disease associations is an interesting and difficult statistical problem. The presence of infinite-dimensional nuisance parameters in the likelihoods for case-control and cohort studies entails considerable theoretical and computational challenges. Although we have conducted a systematic and rigorous investigation, providing powerful new methods, there remain substantial open problems. We list below some directions for future research.

Case-control studies. It is numerically difficult to maximize (6) when N is much larger than n, while algorithms for implementing the constrained maximization mentioned in Remark 3 have yet to be developed. For case-control studies with unknown population totals, identifiability is a thorny issue. We have provided a simple and efficient method under the rare disease assumption, which appears to work well even when the disease is not rare. But can one do better?

Model selection and model assessment. Since our approach is built on likelihood, we can apply likelihood-based model selection criteria, such as the AIC used in §4. Lin (2004) showed that the AIC performs well for the proportional hazards model. It is unclear how to apply the traditional residual-based methods for assessing model adequacy since the haplotypes are not directly observable.

Other genetic variants. We have focused on SNPs-based haplotypes. The proposed inference procedures are potentially applicable to microsatellite loci and other genotype data, although the identifiability of parameters needs to be verified for each kind of genotype data.

Other study designs. It is sometimes desirable to employ the matched case-control design in which one or more controls are individually matched to each case. In large cohort studies with rare diseases, it is cost-effective to adopt the case-cohort design or nested case-control design, so that only a subset of the cohort members needs to be genotyped. We are currently developing efficient inference procedures for such designs.

Population substructure. The presence of latent population substructure can cause bias in association studies of unrelated individuals. There exist several statistical methods to adjust for the effects of population substructure with the aid of genomic markers. It should be possible to extend the proposed methods so as to accommodate potential population substructure.

Studies of related individuals. This paper is concerned with studies of unrelated individuals. Many genetic studies involve multiple family members or relatives. Haplotype ambiguity can potentially be reduced by using the genotype information from related individuals. Inference on haplotype effects needs to account for the intra-class correlation.

Genotyping error and DNA pooling. Laboratory genotyping is prone to error. It is sometimes necessary to pool DNA samples rather genotyping individual samples (Wang, Kidd and Zhao, 2003). Such data creates additional complexity in haplotype analysis.

Many SNPs. The traditional EM algorithm works well for a small number of SNPs. When the number of SNPs is large, the partition-ligation method of Niu et al. (2002) and Qin et al. (2002) and other modifications can potentially be adapted to reduce computing burden. However, the haplotype analysis may not be very useful if the SNPs are weakly linked.

Many haplotypes and rare haplotypes. The approach taken in this paper assumes that one is interested in a small number of haplotype configurations that are relatively frequent. If there are many haplotypes, then we are confronted with the problem of multiple comparisons and sparse data. Schaid (2004) discussed some possible solutions.

Large-scale studies. There is an increasing interest in genome-wide association studies. With a large number of SNPs, one possible approach is to use the sliding windows of 5-10 SNPs and test for the haplotype-disease association in each window. Since most of the SNPs are common between adjacent windows, the test statistics tend to be highly correlated, so that the Bonferroni-type correction for multiple comparisons would be extremely conservative. To properly adjust for multiple comparisons, one needs to ascertain the joint distribution of the test statistics. This can be done by permuting the data or by evaluating the asymptotic multivariate normal distribution of the test statistics; see Lin (2004).

We hope that other statisticians will join us in tackling the above problems and other challenges in genetic association studies.

APPENDIX. TECHNICAL AND COMPUTATIONAL DETAILS

A.1. Proof of Lemma 1

We shall provide a proof under equation (3). The proof under equation (4) is simpler and omitted. To prove the first part of the lemma, we suppose that two sets of parameters $(\{\pi_k\}, \rho)$ and $(\{\widetilde{\pi}_k\}, \widetilde{\rho})$ yield the same distribution of G. We wish to show that these two sets are identical. Consider $g = 2h_k$. For such a choice of g, the set $\mathcal{S}(g)$ is a singleton. Clearly, $(1-\rho)\pi_k^2 + \rho\pi_k = (1-\widetilde{\rho})\widetilde{\pi}_k^2 + \widetilde{\rho}\widetilde{\pi}_k$. We denote this constant by c_k. Then $0 \le c_k \le 1$ for all k, and $0 < c_k < 1$ for at least one k. Since $\pi_k \ge 0$, we have $\pi_k = [-\rho + \{\rho^2 + 4c_k(1-\rho)\}^{1/2}]/2(1-\rho)$. Thus, $(1-\rho)^{-1}$ satisfies the equation $\sum_k[(1-x) + \{(x-1)^2 + 4c_k x\}^{1/2}] = 2$, and $(1-\widetilde{\rho})^{-1}$ satisfies the same equation. It can be shown that the first derivative of $(1-x) + \{(x-1)^2 + 4c_k x\}^{1/2}$ is non-positive, and is strictly negative for at least one k. Thus, the foregoing equation has a unique solution for $x > 1$, which implies that $\rho = \widetilde{\rho}$. It follows immediately that $\pi_k = \widetilde{\pi}_k$ for all k. To prove the second part of the lemma, we choose $g = 2h_k$ to obtain $\nu_k\{2\pi_k(1-\rho) + \rho\} + \mu\pi_k(1-\pi_k) = 0$. Since $\sum_k \nu_k = 0$, we have $\sum_k\{\mu\pi_k(1-\pi_k)\}/\{2\pi_k(1-\rho) + \rho\} = 0$. Therefore, $\mu = 0$ and $\boldsymbol{\nu} = \mathbf{0}$.

A.2. Cross-Sectional Studies

A.2.1. Identifiability Under Arbitrary Distributions of H

Under Condition 1, $(\alpha, \boldsymbol{\beta}, \xi)$ is identifiable. The identifiability of the distribution of H depends on the structure of $P_{\alpha,\boldsymbol{\beta},\xi}$. For concreteness, we consider the codominant logistic regression model for a binary trait. We divide $\mathcal{G}$ into three categories: $\mathcal{G}_1 = \{g \in \mathcal{G} : g = h + h \text{ or } g = h + \widetilde{h}\}$, $\mathcal{G}_2 = \{g \in \mathcal{G} - \mathcal{G}_1 : g \text{ is not } \geq h^*\}$, and $\mathcal{G}_3 = \mathcal{G} - \mathcal{G}_1 - \mathcal{G}_2$. We shall derive the expression for $m_g(y, \mathbf{x}; \boldsymbol{\theta})$ when g belongs to each of the three categories.

For $g \in \mathcal{G}_1$, $\mathcal{S}(g) = \{(h,h)\}$ or $\{(h,\widetilde{h})\}$, so that $m_g(y, \mathbf{x}; \boldsymbol{\theta}) = P_{\alpha,\boldsymbol{\beta},\xi}(Y = y|\mathbf{X} = \mathbf{x}, H = (h,h))P(H = (h,h))$ or $m_g(y, \mathbf{x}; \boldsymbol{\theta}) = P_{\alpha,\boldsymbol{\beta},\xi}(Y = y|\mathbf{X} = \mathbf{x}, H = (h,\widetilde{h}))P(H = (h,\widetilde{h}))$. For $g \in \mathcal{G}_2$, $P_{\alpha,\boldsymbol{\beta},\xi}(Y = y|\mathbf{X} = \mathbf{x}, H = (h_k, h_l))$ does not depend on $(h_k, h_l) \in \mathcal{S}(g)$, so that $m_g(y, \mathbf{x}; \boldsymbol{\theta}) = P_{\alpha,\boldsymbol{\beta},\xi}(Y = y|\mathbf{X} = \mathbf{x}, H = (h_k, h_l))P(G = g)$, where $(h_k, h_l) \in \mathcal{S}(g)$. For $g \in \mathcal{G}_3$,

$$m_g(y, \mathbf{x}; \boldsymbol{\theta}) = \frac{\exp\{y(\alpha + \beta_1 + \boldsymbol{\beta}_3^T\mathbf{x} + \boldsymbol{\beta}_4^T\mathbf{x})\}}{1 + \exp(\alpha + \beta_1 + \boldsymbol{\beta}_3^T\mathbf{x} + \boldsymbol{\beta}_4^T\mathbf{x})}\pi_1(g) + \frac{\exp\{y(\alpha + \boldsymbol{\beta}_3^T\mathbf{x})\}}{1 + \exp(\alpha + \boldsymbol{\beta}_3^T\mathbf{x})}\pi_2(g),$$

where $\pi_1(g) = 2P(H = (h^*, g - h^*)$, and $\pi_2(g) = P(H = (h_k, h_l) : h_k + h_l = g, h_k \neq h^*, h_l \neq h^*)$.

Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$, $P_0(G = g)$ the true value of $P(G = g)$, and $\pi_{0j}(g)$ the true values $\pi_j(g)$, $j = 1, 2$. We have the following conclusions: (1) when $\beta_{01} = 0$ and $\boldsymbol{\beta}_{04} = \mathbf{0}$, $m_g(y, \mathbf{x}; \boldsymbol{\theta}) = m_g(y, \mathbf{x}; \boldsymbol{\theta}_0)$ if and only if $\alpha = \alpha_0$, $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $P(G = g) = P_0(G = g)$ for any $g \in \mathcal{G}$; (2) when either β_{01} or $\boldsymbol{\beta}_{04}$ is nonzero, $m_g(y, \mathbf{x}; \boldsymbol{\theta}) = m_g(y, \mathbf{x}; \boldsymbol{\theta}_0)$ if and only if $\alpha = \alpha_0$, $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $P(G = g) = P_0(G = g)$ for $g \in \mathcal{G}_1 \cup \mathcal{G}_2$, and $\pi_j(g) = \pi_{0j}(g)$ for $g \in \mathcal{G}_3$ and $j = 1, 2$. These conclusions hold for any generalized linear model with the linear predictor given in (1).

A.2.2. EM Algorithm

The complete-data likelihood is proportional to $\prod_{i=1}^n \left\{P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, H_i)P_{\boldsymbol{\gamma}}(H_i)\right\}$. The expectation of the logarithm of this function conditional on the observable data $(Y_i, \mathbf{X}_i, G_i)$, $i = 1, \ldots, n$, is

$$\sum_{i=1}^n \sum_{(h_k,h_l)\in S(G_i)} p_{ikl}(\boldsymbol{\theta}) \left\{\log P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, (h_k, h_l)) + \log P_{\boldsymbol{\gamma}}(h_k, h_l)\right\},$$

where

$$p_{ikl}(\boldsymbol{\theta}) = \frac{P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, (h_k, h_l))P_{\boldsymbol{\gamma}}(h_k, h_l)}{\sum_{(h_k,h_l)\in\mathcal{S}(G_i)} P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, (h_k, h_l))P_{\boldsymbol{\gamma}}(h_k, h_l)}.$$

Thus, in the $(m+1)$th iteration of the EM algorithm, we evaluate $p_{ikl}(\boldsymbol{\theta})$ at the current estimate $\widehat{\boldsymbol{\theta}}^{(m)}$, and obtain $\widehat{\boldsymbol{\theta}}^{(m+1)}$ by solving the following equations through the Newton-Raphson algorithm

$$\sum_{i=1}^n \sum_{(h_k,h_l)\in\mathcal{S}(G_i)} p_{ikl}(\widehat{\boldsymbol{\theta}}^{(m)})\nabla_{\alpha,\boldsymbol{\beta},\xi} \log P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, (h_k, h_l)) = \mathbf{0},$$

$$\sum_{i=1}^n \sum_{(h_k,h_l)\in\mathcal{S}(G_i)} p_{ikl}(\widehat{\boldsymbol{\theta}}^{(m)})\nabla_{\boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}}(h_k, h_l) = \mathbf{0}. \qquad (A.1)$$

Under equation (3) with $\rho \geq 0$, the estimate of $\boldsymbol{\gamma} \equiv (\rho, \{\pi_k\})$ can be obtained in a closed form rather than by solving equation (A.1). Let B be a Bernoulli variable with success probability ρ, Q_1 a discrete random variable taking values in H with $P(Q_1 = (h_k, h_l)) = \delta_{kl}\pi_k$, and Q_2 be another discrete random variable taking values in H with $P(Q_2 = (h_k, h_l)) = \pi_k\pi_l$. Then H has the same distribution as $BQ_1 + (1-B)Q_2$. The complete-data likelihood can be represented by

$$\prod_{i=1}^{n}\left\{P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, H_i)\prod_k \pi_k^{I(Q_{1i}=(h_k,h_k))B_i}\prod_{k,l}(\pi_k\pi_l)^{I(Q_{2i}=(h_k,h_l))(1-B_i)}\rho^{B_i}(1-\rho)^{1-B_i}\right\}.$$

The corresponding score equations for $\{\pi_k\}$ and ρ satisfy

$$\pi_k = c^{-1}\left\{\sum_{i=1}^{n} B_i I(Q_{1i} = (h_k, h_k)) + 2\sum_{i=1}^{n}\sum_{l=1}^{K}(1-B_i)I(Q_{2i} = (h_k, h_l))\right\},$$

$$\rho = n^{-1}\sum_{i=1}^{n} B_i,$$

where c is a normalizing constant such that $\sum_k \pi_k = 1$. Define

$$E\{\omega(B_i, Q_{1i}, Q_{2i})|Y_i, \mathbf{X}_i, G_i\} = \frac{\sum_{bq_1+(1-b)q_2\in\mathcal{S}(G_i)}\omega(b,q_1,q_2)P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, bq_1+(1-b)q_2)p(b,q_1,q_2)}{\sum_{bq_1+(1-b)q_2\in\mathcal{S}(G_i)}P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, bq_1+(1-b)q_2)p(b,q_1,q_2)},$$

where $\omega(B, Q_1, Q_2) = BI(Q_1 = (h_k, h_k))$, $(1-B)I(Q_2 = (h_k, h_l))$ or B, and

$$p(b, q_1, q_2) = \prod_k \pi_k^{bI(q_1=(h_k,h_k))}\prod_{k,l}(\pi_k\pi_l)^{(1-b)I(q_2=(h_k,h_l))}\rho^b(1-\rho)^{1-b}.$$

In the $(m+1)$th iteration, the estimates of π_k and ρ are obtained in closed forms

$$\pi_k^{(m+1)} = \frac{1}{c^{(m+1)}}\left[\sum_{i=1}^{n} E^{(m)}\{B_i I(Q_{1i} = (h_k, h_k))\} + 2\sum_{i=1}^{n}\sum_{l=1}^{K} E^{(m)}\{(1-B_i)I(Q_{2i} = (h_k, h_l))\}\right],$$

$$\rho^{(m+1)} = n^{-1}\sum_{i=1}^{n} E^{(m)}(B_i),$$

where $E^{(m)}\{\omega(B_i, Q_{1i}, Q_{2i})\}$ is $E\{\omega(B_i, Q_{1i}, Q_{2i})|Y_i, \mathbf{X}_i, G_i\}$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(m)}$, and $c^{(m+1)}$ is the constant such that $\sum_k \pi_k^{(m+1)} = 1$.

A.3. Case-Control Studies With Known Population Totals

A.3.1. EM Algorithm

This is similar to the EM algorithm for cross-sectional studies, except that, in addition to unknown H on all individuals, $\mathbf{X}$ is missing for the individuals not selected into the case-control sample and that there are nonparametric components $\{F_g(\cdot)\}$. The complete-data likelihood is

$$\prod_{i=1}^{N} P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i, H_i)P_{\boldsymbol{\gamma}}(H_i)\prod_g \{f_g(\mathbf{X}_i)\}^{I(G_i=g)}.$$

The M-step solves the following equations for $\boldsymbol{\theta}$:

$$\sum_{i=1}^{N} I(R_i=1)E\{\nabla_{\alpha,\boldsymbol{\beta},\xi}\log P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i,H_i)|Y_i,\mathbf{X}_i,G_i\}+\sum_{i=1}^{N} I(R_i=0)E\{\nabla_{\alpha,\boldsymbol{\beta},\xi}\log P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i,H_i)|Y_i\}=\mathbf{0},$$

$$\sum_{i=1}^{N} I(R_i=1)E\{\nabla_{\boldsymbol{\gamma}}\log P_{\boldsymbol{\gamma}}(H_i)|Y_i,\mathbf{X}_i,G_i\}+\sum_{i=1}^{N} I(R_i=0)E\{\nabla_{\boldsymbol{\gamma}}\log P_{\boldsymbol{\gamma}}(H_i)|Y_i\}=\mathbf{0}, \qquad (A.2)$$

and estimates F_g by an empirical function with the following point mass at the $\mathbf{X}_i$ for which $(G_i = g, R_i = 1)$

$$F_g\{\mathbf{X}_i\}=\frac{\sum_{j=1}^{N} I(\mathbf{X}_j=\mathbf{X}_i,G_j=g,R_j=1)+\sum_{j=1}^{N} I(R_j=0)E\{I(\mathbf{X}_j=\mathbf{X}_i,G_j=g)|Y_j\}}{\sum_{j=1}^{N} I(G_j=g,R_j=1)+\sum_{j=1}^{N} I(R_j=0)E\{I(G_j=g)|Y_j\}},$$

where the conditional expectations are evaluated at the current estimates of $\boldsymbol{\theta}$ and $\{F_g\}$ in the E-step. For a random function $\omega(Y_i,\mathbf{X}_i,H_i)$, the conditional expectation takes the form

$$\frac{\sum_{(h_k,h_l)\in\mathcal{S}(G_i)} w(Y_i,\mathbf{X}_i,(h_k,h_l))P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i,(h_k,h_l))P_{\boldsymbol{\gamma}}(h_k,h_l)}{\sum_{(h_k,h_l)\in\mathcal{S}(G_i)} P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{X}_i,(h_k,h_l))P_{\boldsymbol{\gamma}}(h_k,h_l)}$$

for $R_i = 1$, and

$$\frac{\sum_{g\in\mathcal{G}}\sum_{\mathbf{x}\in\{\mathbf{X}_i:G_i=g,R_i=1\}}\sum_{(h_k,h_l)\in\mathcal{S}(g)}\omega(Y_i,\mathbf{x},(h_k,h_l))P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{x},(h_k,h_l))P_{\boldsymbol{\gamma}}(h_k,h_l)F_g\{\mathbf{x}\}}{\sum_{g\in\mathcal{G}}\sum_{\mathbf{x}\in\{\mathbf{X}_i:G_i=g,R_i=1\}}\sum_{(h_k,h_l)\in\mathcal{S}(g)} P_{\alpha,\boldsymbol{\beta},\xi}(Y_i|\mathbf{x},(h_k,h_l))P_{\boldsymbol{\gamma}}(h_k,h_l)F_g\{\mathbf{x}\}}$$

for $R_i = 0$. Under equation (3) with $\rho \geq 0$, the idea described in A.2.2 can be applied to (A.2) to obtain a closed-form estimate of $\boldsymbol{\gamma}$.

A.3.2. Proof of Theorem 1

The case-control design with known population totals is a special case of the two-phase designs studied by Breslow et al. (2003). The likelihood given in (6) resembles (2.3) of Breslow et al. The key difference is that the former involves several nonparametric components $\{F_g(\cdot)\}$ whereas the latter involves only a single nonparametric function. Despite this difference, the arguments of Breslow et al. can be used to prove Theorem 1 with little modifications. Specifically, the regularity conditions of Breslow et al. hold under our Conditions 1–4. Thus, the consistency of $(\widehat{\boldsymbol{\theta}},\{\widehat{F}_g(\cdot)\})$ follows from the results of van der Vaart and Wellner (2001), while the weak convergence and asymptotic efficiency can be established by applying the results of Murphy and van der Vaart (2000) via a least favorable submodel, which can be constructed along the lines of Breslow et al. (2003, §3).

A.4. Case-Control Studies With Unknown Population Totals

A.4.1. Equivalence Class

Suppose that two sets of parameters $(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\})$ and $(\widetilde{\boldsymbol{\theta}}, \{\widetilde{F}_g^\dagger\}, \{\widetilde{q}_g\})$ yield the same likelihood, i.e.,

$$RL(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\}) = RL(\widetilde{\boldsymbol{\theta}}, \{\widetilde{F}_g^\dagger\}, \{\widetilde{q}_g\}). \tag{A.3}$$

Since $\eta(0, \mathbf{x}, g; \boldsymbol{\theta}) = 1$, (A.3) with $y = 0$ implies that $f_g^\dagger(\mathbf{x}) q_g / \sum_{\widetilde{g} \in \mathcal{G}} q_{\widetilde{g}} = \widetilde{f}_g^\dagger(\mathbf{x}) \widetilde{q}_g / \sum_{\widetilde{g} \in \mathcal{G}} \widetilde{q}_{\widetilde{g}}$. Thus, $f_g^\dagger(\mathbf{x}) = \widetilde{f}_g^\dagger(\mathbf{x})$ and $q_g = \widetilde{q}_g$. It then follows from (A.3) that

$$\eta(y, \mathbf{x}, g; \boldsymbol{\theta}) = C(y) \eta(y, \mathbf{x}, g; \widetilde{\boldsymbol{\theta}}), \tag{A.4}$$

where $C(y)$ depends only on y. By setting $\mathbf{x} = \mathbf{x}_0$ and $g = g_0$ in (A.4) and noting that $\eta(y, \mathbf{x}_0, g_0; \boldsymbol{\theta}) = 1$, we conclude that $C(y) = 1$. Hence, the equivalence class for $(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\})$ is $\big\{(\widetilde{\boldsymbol{\theta}}, \{F_g^\dagger\}, \{q_g\}) : \eta(y, \mathbf{x}, g; \widetilde{\boldsymbol{\theta}}) = \eta(y, \mathbf{x}, g; \boldsymbol{\theta})\big\}$.

A.4.2. Identifiability for Logistic Link Function

Suppose that

$$\eta(y, \mathbf{x}, g; \widetilde{\boldsymbol{\theta}}) = \eta(y, \mathbf{x}, g; \boldsymbol{\theta}) \tag{A.5}$$

for two set of parameters $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. Let $g_0 = 0$. As in Appendix A.2.1, we partition $\mathcal{G}$ into $(\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3)$. For $g \in \mathcal{G}_1$, $\mathcal{S}(g)$ is a singleton, so that the generalized odds ratio reduces to the ordinary odds ratio of Y given $\mathbf{X}$ and H. Thus, equation (A.5) is equivalent to $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$ under Condition 8. For $g \in \mathcal{G}_2$, $P(Y = 0 | \mathbf{X} = \mathbf{x}, H = (h_k, h_l)) = \{1 + \exp(\alpha + \boldsymbol{\beta}_3^T \mathbf{x})\}^{-1}$. Thus, (A.5) holds if and only if $\widetilde{\boldsymbol{\beta}}_3 = \boldsymbol{\beta}_3$. For $g \in \mathcal{G}_3$, both $\pi_1(g)$ and $\pi_2(g)$ are non-zero. Then equation (A.5) becomes

$$\begin{aligned} &\frac{\widetilde{\pi}_1(g)(1 + e^{\widetilde{\alpha} + \psi_2(\mathbf{x})}) / \widetilde{\pi}_2(g)(1 + e^{\widetilde{\alpha} + \psi_1(\mathbf{x})}) + e^{\psi_2(\mathbf{x}) - \psi_1(\mathbf{x})}}{\widetilde{\pi}_1(g)(1 + e^{\widetilde{\alpha} + \psi_2(\mathbf{x})}) / \widetilde{\pi}_2(g)(1 + e^{\widetilde{\alpha} + \psi_1(\mathbf{x})}) + 1} \\ &= \frac{\pi_1(g)(1 + e^{\alpha + \psi_2(\mathbf{x})}) / \pi_2(g)(1 + e^{\alpha + \psi_1(\mathbf{x})}) + e^{\psi_2(\mathbf{x}) - \psi_1(\mathbf{x})}}{\pi_1(g)(1 + e^{\alpha + \psi_2(\mathbf{x})}) / \pi_2(g)(1 + e^{\alpha + \psi_1(\mathbf{x})}) + 1}, \end{aligned} \tag{A.6}$$

where $\psi_1(\mathbf{x}) = \beta_1 + \boldsymbol{\beta}_3^T \mathbf{x} + \boldsymbol{\beta}_4^T \mathbf{x}$ and $\psi_2(\mathbf{x}) = \boldsymbol{\beta}_3^T \mathbf{x}$.

Without loss of generality, assume that 0 is in the support of $\mathbf{X}$. We have the following results.

1. $\beta_1 = 0$ and $\boldsymbol{\beta}_4 = \mathbf{0}$. Then (A.6) holds naturally.

2. $\beta_1 \neq 0, \boldsymbol{\beta}_4 = \mathbf{0}$ and $\boldsymbol{\beta}_3 = \mathbf{0}$. Then since the function $(\lambda + c)/(\lambda + 1)$ is strictly monotone in λ for $c \neq 1$, (A.6) yields

$$\frac{\widetilde{\pi}_1(g)}{\widetilde{\pi}_2(g)} \frac{1 + e^{\widetilde{\alpha}}}{1 + e^{\widetilde{\alpha} + \beta_1}} = \frac{\pi_1(g)}{\pi_2(g)} \frac{1 + e^{\alpha}}{1 + e^{\alpha + \beta_1}}.$$

Thus, (A.6) is equivalent to

$$\frac{\widetilde{\pi}_1(g) / \widetilde{\pi}_2(g)}{\widetilde{\pi}_1(\widetilde{g}) / \widetilde{\pi}_2(\widetilde{g})} = \frac{\pi_1(g) / \pi_2(g)}{\pi_1(\widetilde{g}) / \pi_2(\widetilde{g})}, \quad \text{for all } g, \widetilde{g} \in \mathcal{G}_3.$$

3. $\beta_1 \neq 0, \boldsymbol{\beta}_4 = \mathbf{0}$ and $\beta_{3,z} \neq 0$, where $\beta_{3,z}$ is the component of $\boldsymbol{\beta}_3$ associated with a continuous covariate Z. For $\mathbf{x}$ such that $\beta_{3,z} z \neq 0$, (A.6) yields

$$\frac{\widetilde{\pi}_1(g)}{\widetilde{\pi}_2(g)} \frac{1 + e^{\widetilde{\alpha} + \beta_{3,z} z}}{1 + e^{\widetilde{\alpha} + \beta_1 + \beta_{3,z} z}} = \frac{\pi_1(g)}{\pi_2(g)} \frac{1 + e^{\alpha + \beta_{3,z} z}}{1 + e^{\alpha + \beta_1 + \beta_{3,z} z}}.$$

The above equation holds for any $z \in (-\infty, \infty)$ since the functions on the two sides are analytic in z and z is continuous. Without loss of generality, assume that $\beta_{3,z} > 0$. By letting $z = -\infty$, we have $\widetilde{\pi}_1(g)/\widetilde{\pi}_2(g) = \pi_1(g)/\pi_2(g)$. Then by letting $z = 0$, we have $\widetilde{\alpha} = \alpha$. Thus, (A.6) is equivalent to $\{\widetilde{\alpha} = \alpha,\ \widetilde{\pi}_1(g)/\widetilde{\pi}_2(g) = \pi_1(g)/\pi_2(g)\}$.

4. $\beta_{4,z} \neq 0$, where $\beta_{4,z}$ is the component of $\boldsymbol{\beta}_4$ pertaining to z. Then (A.6) is equivalent to

$$\frac{\widetilde{\pi}_1(g)}{\widetilde{\pi}_2(g)} \frac{1 + e^{\widetilde{\alpha} + \psi_2(\mathbf{x})}}{1 + e^{\widetilde{\alpha} + \psi_1(\mathbf{x})}} = \frac{\pi_1(g)}{\pi_2(g)} \frac{1 + e^{\alpha + \psi_2(\mathbf{x})}}{1 + e^{\alpha + \psi_1(\mathbf{x})}} \tag{A.7}$$

for any $\mathbf{x}$ such that $\beta_1 + \boldsymbol{\beta}_4^T \mathbf{x} \neq 0$. We set $\mathbf{x}$ except the component z to 0. By letting $z \to -\beta_1/\beta_{4,z}$, we have $\widetilde{\pi}_1(g)/\widetilde{\pi}_2(g) = \pi_1(g)/\pi_2(g)$. Then by differentiating both sides of (A.7) with respect to z and letting $z \to -\beta_1/\beta_{4,z}$, we obtain $\alpha = \widetilde{\alpha}$. Thus, (A.6) is equivalent to $\{\widetilde{\alpha} = \alpha,\ \widetilde{\pi}_1(g)/\widetilde{\pi}_2(g) = \pi_1(g)/\pi_2(g)\}$.

A.4.3. Identifiability for Probit and Complementary Log-Log Link Functions

Assume that $|\beta_1| + |\boldsymbol{\beta}_4| \neq 0$. Also, there exists a continuous covariate in $\mathbf{X}$, denoted by Z, such as that the corresponding regression parameter β_z is non-zero. Let $\mathbf{x}_0 = \mathbf{0}$ and $g_0 = 0$. We claim that under the probit and complementary log-log regression models, $\eta(1, \mathbf{x}, g; \boldsymbol{\theta}) = \eta(1, \mathbf{x}, g; \widetilde{\boldsymbol{\theta}})$ for two sets of parameters $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$ if and only if $\alpha = \widetilde{\alpha}$, $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$ and $\pi_1(g)/\pi_2(g) = \widetilde{\pi}_1(g)/\widetilde{\pi}_2(g)$ for $g \in \mathcal{G}_3$.

We first prove the above claim for the probit model. Suppose that $\eta(1, \mathbf{x}, g; \boldsymbol{\theta}) = \eta(1, \mathbf{x}, g; \widetilde{\boldsymbol{\theta}})$. Without loss of generality, assume that h^* is a non-zero vector. Let $g = 2h^*, h^* + \widetilde{h}^*$ and 0 in turn. Since $\mathcal{S}(g)$ has a single element for such g, we obtain

$$\begin{aligned} &\frac{\Phi(\alpha)}{1 - \Phi(\alpha)} \left\{1/\Phi(\alpha + 2\beta_1 + \beta_2 + \boldsymbol{\beta}_3^T \mathbf{x} + 2\boldsymbol{\beta}_4^T \mathbf{x} + \boldsymbol{\beta}_5^T \mathbf{x}) - 1\right\} \\ &\quad = \frac{\Phi(\widetilde{\alpha})}{1 - \Phi(\widetilde{\alpha})} \left\{1/\Phi(\widetilde{\alpha} + 2\widetilde{\beta}_1 + \widetilde{\beta}_2 + \widetilde{\boldsymbol{\beta}}_3^T \mathbf{x} + 2\widetilde{\boldsymbol{\beta}}_4^T \mathbf{x} + \widetilde{\boldsymbol{\beta}}_5^T \mathbf{x}) - 1\right\}, \end{aligned} \tag{A.8}$$

$$\frac{\Phi(\alpha)}{1 - \Phi(\alpha)} \left\{1/\Phi(\alpha + \beta_1 + \boldsymbol{\beta}_3^T \mathbf{x} + \boldsymbol{\beta}_4^T \mathbf{x}) - 1\right\} = \frac{\Phi(\widetilde{\alpha})}{1 - \Phi(\widetilde{\alpha})} \left\{1/\Phi(\widetilde{\alpha} + \widetilde{\beta}_1 + \widetilde{\boldsymbol{\beta}}_3^T \mathbf{x} + \widetilde{\boldsymbol{\beta}}_4^T \mathbf{x}) - 1\right\}, \tag{A.9}$$

$$\frac{\Phi(\alpha)}{1 - \Phi(\alpha)} \left\{1/\Phi(\alpha + \boldsymbol{\beta}_3^T \mathbf{x}) - 1\right\} = \frac{\Phi(\widetilde{\alpha})}{1 - \Phi(\widetilde{\alpha})} \left\{1/\Phi(\widetilde{\alpha} + \widetilde{\boldsymbol{\beta}}_3^T \mathbf{x}) - 1\right\}, \tag{A.10}$$

where Φ is the normal distribution function. In (A.10), we let $\mathbf{x}$ except the component z be 0. Then

$$\frac{\Phi(\alpha)}{1 - \Phi(\alpha)} \left\{1/\Phi(\alpha + \beta_z z) - 1\right\} = \frac{\Phi(\widetilde{\alpha})}{1 - \Phi(\widetilde{\alpha})} \left\{1/\Phi(\widetilde{\alpha} + \widetilde{\beta}_z z) - 1\right\}.$$

By letting $z \to \infty$ or $-\infty$, we conclude that β_z and $\widetilde{\beta}_z$ must have the same sign. Without loss of generality, assume that $\beta_z > \widetilde{\beta}_z > 0$. Then the left-hand side divided by the right-hand side goes to 0 as $z \to \infty$. This is a contradiction. Therefore, $\beta_z = \widetilde{\beta}_z$. We differentiate both sides to obtain

$$\frac{\Phi(\alpha)}{1-\Phi(\alpha)}\frac{\phi(\alpha+\beta_z z)}{\Phi(\alpha+\beta_z z)^2} = \frac{\Phi(\widetilde{\alpha})}{1-\Phi(\widetilde{\alpha})}\frac{\phi(\widetilde{\alpha}+\beta_z z)}{\Phi(\widetilde{\alpha}+\beta_z z)^2}.$$

By taking the ratio of the two sides and letting $z \to \mathrm{sgn}(\beta_z)\infty$, we immediately conclude that $\alpha = \widetilde{\alpha}$. Applying this result to (A.8)–(A.10), we obtain $2\beta_1 + \beta_2 + \boldsymbol{\beta}_3^T\mathbf{x} + 2\boldsymbol{\beta}_4^T\mathbf{x} + \boldsymbol{\beta}_5^T\mathbf{x} = 2\widetilde{\beta}_1 + \widetilde{\beta}_2 + \widetilde{\boldsymbol{\beta}}_3^T\mathbf{x} + 2\widetilde{\boldsymbol{\beta}}_4^T\mathbf{x} + \widetilde{\boldsymbol{\beta}}_5^T\mathbf{x}$, $\beta_1 + \boldsymbol{\beta}_3^T\mathbf{x} + \boldsymbol{\beta}_4^T\mathbf{x} = \widetilde{\beta}_1 + \widetilde{\boldsymbol{\beta}}_3^T\mathbf{x} + \widetilde{\boldsymbol{\beta}}_4^T\mathbf{x}$, and $\boldsymbol{\beta}_3^T\mathbf{x} = \widetilde{\boldsymbol{\beta}}_3^T\mathbf{x}$. Therefore, $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$. For $g \in \mathcal{G}_3$,

$$\eta(1,\mathbf{x},g;\boldsymbol{\theta}) = \frac{1-\Phi(\alpha)}{\Phi(\alpha)}\frac{\Phi(\alpha+\beta_1+\boldsymbol{\beta}_3^T\mathbf{x}+\boldsymbol{\beta}_4^T\mathbf{x})\pi_1(g)/\pi_2(g) + \Phi(\alpha+\boldsymbol{\beta}_3^T\mathbf{x})}{\{1-\Phi(\alpha+\beta_1+\boldsymbol{\beta}_3^T\mathbf{x}+\boldsymbol{\beta}_4^T\mathbf{x})\}\pi_1(g)/\pi_2(g) + 1 - \Phi(\alpha+\boldsymbol{\beta}_3^T\mathbf{x})}. \qquad (A.11)$$

It follows that $\pi_1(g)/\pi_2(g) = \widetilde{\pi}_1(g)/\widetilde{\pi}_2(g)$. The other direction of the claim is obvious in view of (A.11) and the expressions of $\eta(1,\mathbf{x},g)$ for $g \in \mathcal{G}_1$ and $g \in \mathcal{G}_2$.

For the complementary log-log model, we obtain the same equations as (A.8)–(A.11) with $\Phi(x)$ replaced by $1-\exp(-e^x)$. In particular, $e^{-e^\alpha}(e^{e^{\alpha+\beta_z z}}-1)/(1-e^{-e^\alpha}) = e^{-e^{\widetilde{\alpha}}}(e^{e^{\widetilde{\alpha}+\widetilde{\beta}_z z}}-1)/(1-e^{-e^{\widetilde{\alpha}}})$. Taking the first and second derivatives of the two sides with respect to z and forming the ratio of them, we obtain $\beta_z(e^{\alpha+\beta_z z}+1) = \widetilde{\beta}_z(e^{\widetilde{\alpha}+\widetilde{\beta}_z z}+1)$. Thus, $\alpha = \widetilde{\alpha}$ and $\beta_z = \widetilde{\beta}_z$. The rest of the proof is the same as that of the probit model.

A.4.4. Profile Likelihood of $\boldsymbol{\theta}$ Based on (8)

Suppose that there are J distinct observed values of $(\mathbf{X}, G)$, denoted by $(\mathbf{x}_1, g_1), \ldots, (\mathbf{x}_J, g_J)$. Let n_{1j} and n_{0j} be the numbers of times $(\mathbf{x}_j, g_j)$ is observed in the cases and controls, respectively, and let δ_j be the jump size of the estimated distribution of $(\mathbf{X}, G)$ at $(\mathbf{x}_j, g_j)$. Then the log-likelihood based on (8) can be written as

$$l_n(\boldsymbol{\theta}, \{\delta_j\}) = \sum_{j=1}^{J} n_{1j}\log\eta(1,\mathbf{x}_j,g_j;\boldsymbol{\theta}) - n_1\log\Big\{\sum_{j=1}^{J}\eta(1,\mathbf{x}_j,g_j;\boldsymbol{\theta})\delta_j\Big\} + \sum_{j=1}^{J} n_{+j}\log\delta_j,$$

where $n_{+j} = n_{0j} + n_{1j}$. Following Scott and Wild (1997), we introduce a Lagrange multiplier λ for the constraint $\sum_j \delta_j = 1$ and set the derivative with respect to δ_j to 0. Then we obtain

$$\frac{n_{+j}}{\delta_j} - \frac{n_1\eta(1,\mathbf{x}_j,g_j;\boldsymbol{\theta})}{\sum_{j=1}^{J}\eta(1,\mathbf{x}_j,g_j;\boldsymbol{\theta})\delta_j} + \lambda = 0.$$

Multiplying both sides by δ_j and summing over j entails $\lambda = n_1 - n$. Thus,

$$\delta_j = \frac{n_{+j}}{n - n_1 + n_1\eta(1,\mathbf{x}_j,g_j;\boldsymbol{\theta})/\mu}, \qquad (A.12)$$

where $\mu = \sum_{j=1}^{J} \eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta})\delta_j$. Plugging (A.12) into $l_n(\boldsymbol{\theta}, \{\delta_j\})$, we see that the objective function to be maximized is, up to a constant C_n, equal to

$$l_n^*(\boldsymbol{\theta}, \mu) = \sum_{j=1}^{J} n_{1j} \log \eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta}) - \sum_{j=1}^{J} n_{+j} \log \{(n_1/n)\eta(1, \mathbf{x}_j, g_j; \boldsymbol{\theta}) + (1 - n_1/n)\mu\} + (n - n_1) \log \mu.$$

Thus, $\max_{\{\delta_j\}} l_n(\boldsymbol{\theta}, \{\delta_j\}) \leq \max_{\mu} l_n^*(\boldsymbol{\theta}, \mu) + C_n$. If μ maximizes $l_n^*(\boldsymbol{\theta}, \mu)$, then $\partial l_n^*(\boldsymbol{\theta}, \mu)/\partial \mu = 0$ and the δ_j given in (A.12) satisfy $\sum_{j=1}^{J} \delta_j = 1$. Thus, $\max_{\mu} l_n^*(\boldsymbol{\theta}, \mu) + C_n \leq \max_{\{\delta_j\}} l_n(\boldsymbol{\theta}, \{\delta_j\})$. Therefore, the profile log-likelihood function for $\boldsymbol{\theta}$ based on $l_n(\boldsymbol{\theta}, \{\delta_j\})$ equals the profile function based on $l_n^*(\boldsymbol{\theta}, \mu)$, up to a constant C_n. We maximize $l_n^*(\boldsymbol{\theta}, \mu)$ via the Newton-Raphson algorithm to yield $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mu}$, where $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. It can be shown that, up to a constant, $l_n^*(\boldsymbol{\theta}, \mu)$ is the log-likelihood based on a random sample of size n from a conditional distribution of Y given $\mathbf{X}$ and G. Hence, the covariance matrix of $(\widehat{\boldsymbol{\theta}}, \widehat{\mu})$ can be estimated by the inversed information matrix of $l_n^*(\boldsymbol{\theta}, \mu)$.

A.4.5. Profile Likelihood of $\boldsymbol{\theta}$ Based on (9)

Suppose that equation (3) holds. Write $\boldsymbol{\theta} = (\boldsymbol{\beta}, \{\pi_k\}, \rho)$. Also, define

$$\zeta_1(\mathbf{x}, g; \boldsymbol{\theta}) = \sum_{(h_k, h_l) \in \mathcal{S}(g)} e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{x}, h_k, h_l)} \{\rho \pi_k \delta_{kl} + (1 - \rho)\pi_k \pi_l\}, \; \zeta_0(g; \boldsymbol{\theta}) = \sum_{(h_k, h_l) \in \mathcal{S}(g)} \{\rho \pi_k \delta_{kl} + (1 - \rho)\pi_k \pi_l\}.$$

By a derivation similar to that of Appendix A.4.4, profiling (9) over $\{F_g(\cdot)\}$ is equivalent to profiling the following function over $\{\mu_g\}$

$$\widetilde{l}_n^*(\boldsymbol{\theta}, \{\mu_g\}) = \sum_{i=1}^{n} \big\{ y_i \log \zeta_1(\mathbf{X}_i, G_i; \boldsymbol{\theta}) + (1 - y_i) \log \zeta_0(G_i; \boldsymbol{\theta}) \big\}$$

$$- \sum_{i=1}^{n} \sum_{g} I(G_i = g) \log \Big\{ \zeta_1(\mathbf{X}_i, G_i; \boldsymbol{\theta}) + n_1^{-1} \widetilde{n}_g \sum_{\widetilde{g}} \mu_{\widetilde{g}} - \mu_g \Big\} + \sum_{i=1}^{n} (1 - y_i) \log \Big\{ \sum_{g} \mu_g \Big\},$$

where $\widetilde{n}_g$ is the number of times $G = g$ in the sample. The covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be estimated by the sandwich estimator or the profile likelihood method.

If $\mathbf{X}$ is independent of G, then we obtain the MLE $\widehat{\boldsymbol{\theta}}$ by maximizing the following function

$$\widetilde{l}_n^*(\boldsymbol{\theta}, \mu) = \sum_{i=1}^{n} y_i \log \zeta_1(\mathbf{X}_i, G_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} (1 - y_i) \log \zeta_0(G_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} (1 - y_i) \log \mu$$

$$- \sum_{i=1}^{n} \log \Big\{ (1 - r)\mu + r \sum_{g} \zeta_1(\mathbf{X}_i, g; \boldsymbol{\theta}) \Big\},$$

where $r = n_1/n$. Let $H = BQ_1 + (1-B)Q_2$, where B is a Bernoulli variable, Q_1 takes values in $\{(h_k, h_k); k = 1, \ldots, K\}$ and Q_2 takes values in $\{(h_k, h_l); k, l = 1, \ldots, K\}$. Suppose that Y is a binary variable and that the conditional distribution of (B, Q_1, Q_2, Y) given $\mathbf{X}$ is characterized by

$$P(B, Q_1, Q_2, Y|\mathbf{X}) = \frac{\exp\{\boldsymbol{\vartheta}^T \mathcal{W}(B, Q_1, Q_2, Y, \mathbf{X})\}}{\sum_{B,Q_1,Q_2,Y} \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(B, Q_1, Q_2, Y, \mathbf{X})\}},$$

where $\boldsymbol{\vartheta} = (-\log\mu + \log r/(1-r), \boldsymbol{\beta}, \log\pi_1 - \log\rho/(1-\rho), \ldots, \log\pi_K - \log\rho/(1-\rho))^T$ and $\mathcal{W}(B, Q_1, Q_2, Y, \mathbf{X}) = (Y, Y\mathcal{Z}(\mathbf{X}, H), BI(Q_1 = (h_1, h_1)) + 2(1-B)\sum_l I(Q_2 = (h_1, h_l)), \ldots, BI(Q_1 = (h_K, h_K)) + 2(1-B)\sum_l I(Q_2 = (h_K, h_l)))^T$. It can be verified that $\widetilde{l}_n^*(\boldsymbol{\theta}, \mu)$ is equivalent to the log-likelihood

$$\widetilde{l}_n^*(\boldsymbol{\vartheta}) = \sum_{i=1}^n \log \left[\sum_{BQ_1+(1-B)Q_2 \in \mathcal{S}(G_i)} \frac{\exp\{\boldsymbol{\vartheta}^T \mathcal{W}(B, Q_1, Q_2, Y_i, \mathbf{X}_i)\}}{\sum_{b,q_1,q_2,y} \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}} \right].$$

We maximize $\widetilde{l}_n^*(\boldsymbol{\vartheta})$ through the EM-algorithm, in which (B, Q_1, Q_2) is treated as missing. The estimation of the covariance matrix of $\widehat{\boldsymbol{\theta}}$ is based on the information matrix of $\widetilde{l}_n^*(\boldsymbol{\vartheta})$.

The complete-data score function is

$$\sum_{i=1}^n \left[\mathcal{W}(B_i, Q_{1i}, Q_{2i}, Y_i, \mathbf{X}_i) - \sum_{i=1}^n \frac{\sum_{b,q_1,q_2,y} \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i) \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}}{\sum_{b,q_1,q_2,y} \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}} \right].$$

Thus, in E-step, we calculate the conditional expectation of $\mathcal{W}(B_i, Q_{1i}, Q_{2i}, Y_i, \mathbf{X}_i)$ given $(Y_i, \mathbf{X}_i, G_i)$ and the current parameter estimates:

$$E[\mathcal{W}(B_i, Q_{1i}, Q_{2i}, Y_i, \mathbf{X}_i)|Y_i, \mathbf{X}_i, G_i]$$

$$= \frac{\sum_{b,q_1,q_2} I(bq_1 + (1-b)q_2 \in \mathcal{S}(G_i)) \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, Y_i, \mathbf{X}_i)\} \mathcal{W}(b, q_1, q_2, Y_i, \mathbf{X}_i)}{\sum_{b,q_1,q_2} I(bq_1 + (1-b)q_2 \in \mathcal{S}(G_i)) \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, Y_i, \mathbf{X}_i)\}}.$$

In the M-step, we use the one-step Newton-Raphson iteration to update the parameter estimates:

$$\boldsymbol{\vartheta}^{(k+1)} = \boldsymbol{\vartheta}^{(k)} - \boldsymbol{\Sigma}^{-1} \times \sum_{i=1}^n \Big[E[\mathcal{W}(B, Q_1, Q_2, Y_i, \mathbf{X}_i)|Y_i, \mathbf{X}_i, G_i]$$

$$- \sum_{i=1}^n \frac{\sum_{b,q_1,q_2,y} \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i) \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}}{\sum_{b,q_1,q_2,y} \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}} \Big],$$

where

$$\boldsymbol{\Sigma} = -\left[\sum_{i=1}^n \frac{\sum_{b,q_1,q_2,y} \mathcal{W}^{\otimes 2}(b, q_1, q_2, y, \mathbf{X}_i) \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}}{\sum_{b,q_1,q_2,y} \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\}} \right]$$

$$+ \sum_{i=1}^n \left[\frac{\left\{ \sum_{b,q_1,q_2,y} \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i) \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\} \right\}^{\otimes 2}}{\left\{ \sum_{b,q_1,q_2,y} \exp\{\boldsymbol{\vartheta}^T \mathcal{W}(b, q_1, q_2, y, \mathbf{X}_i)\} \right\}^2} \right],$$

and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$.

A.4.6. Proof of Theorem 2

Write $F_{\mathbf{x},g}(\mathbf{x},g) = F_g^\dagger(\mathbf{x})q_g$ and $\widehat{F}_{\mathbf{x},g}(\mathbf{x},g) = \widehat{F}_g^\dagger(\mathbf{x})\widehat{q}_g$. Since $\widehat{\boldsymbol{\theta}}$ is bounded and $\widehat{F}_{\mathbf{x},g}$ is a probability distribution, we can choose a subsequence such that $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ and $\widehat{F}_{\mathbf{x},g}(\mathbf{x},g) \to F^*_{\mathbf{x},g}(\mathbf{x},g) \equiv F_g^*(\mathbf{x})q_g^*$, where $q_g^* > 0$ for any g.

Since $\widehat{F}_g^\dagger$ maximizes the likelihood, there exists some Lagrange multiplier $\widehat{\lambda}_g$ such that

$$\frac{I(G_i = g)}{\widehat{F}_g^\dagger\{\mathbf{X}_i\}} - \frac{n_1\eta(1,\mathbf{X}_i,g;\widehat{\boldsymbol{\theta}})\widehat{q}_g}{\int_{\mathbf{x},\widetilde{g}}\eta(1,\mathbf{x},\widetilde{g};\widehat{\boldsymbol{\theta}})d\widehat{F}_{\mathbf{x},g}(\mathbf{x},\widetilde{g})} - n\widehat{\lambda}_g = 0,$$

where $\widehat{F}_g^\dagger\{\mathbf{X}_i\}$ denotes the point mass of $\widehat{F}_g^\dagger$ at $\mathbf{X}_i$, and the integral is interpreted as integration over $\mathbf{x}$ and summation over g. Since $\sum_{i=1}^n \widehat{F}_g^\dagger\{\mathbf{X}_i\} = 1$, $\widehat{\lambda}_g$ satisfies the equation

$$n^{-1}\sum_{i=1}^n \frac{I(G_i = g)}{\widehat{\lambda}_g + n_1\eta(1,\mathbf{X}_i,g;\widehat{\boldsymbol{\theta}})\widehat{q}_g\{n\int_{\mathbf{x},\widetilde{g}}\eta(1,\mathbf{x},\widetilde{g};\widehat{\boldsymbol{\theta}})d\widehat{F}_{\mathbf{x},g}(\mathbf{x},\widetilde{g})\}^{-1}} = 1, \tag{A.13}$$

and

$$\min_{1\le i\le n}\left\{\widehat{\lambda}_g + \frac{n_1\eta(1,\mathbf{X}_i,g;\widehat{\boldsymbol{\theta}})\widehat{q}_g}{n\int_{\mathbf{x},\widetilde{g}}\eta(1,\mathbf{x},\widetilde{g};\widehat{\boldsymbol{\theta}})d\widehat{F}_{\mathbf{x},g}(\mathbf{x},\widetilde{g})}\right\} > 0.$$

Clearly, $\widehat{\lambda}_g$ must be bounded asymptotically. Thus, by choosing a subsequence, we assume that $\widehat{\lambda}_g \to \lambda_g^*$.

By (A.13) and the Lipschitz continuity of $\eta(1,\mathbf{x},g;\boldsymbol{\theta}^*)$ in the continuous components of $\mathbf{x}$, we can show that there exists a positive constant δ such that

$$\min_{g,\mathbf{x}}\left\{\left|\lambda_g^* + \frac{\varrho\eta(1,\mathbf{x},g;\boldsymbol{\theta}^*)q_g^*}{\int_{\mathbf{x},\widetilde{g}}\eta(1,\mathbf{x},\widetilde{g};\boldsymbol{\theta}^*)dF^*_{\mathbf{x},g}(\mathbf{x},\widetilde{g})}\right|\right\} > \delta.$$

Consequently, when n is sufficiently large,

$$\widehat{F}_g^\dagger(\mathbf{x}) = n^{-1}\sum_{i=1}^n \frac{I(G_i = g, \mathbf{X}_i \le \mathbf{x})}{\max\left[|\widehat{\lambda}_g + \eta(1,\mathbf{X}_i,g;\widehat{\boldsymbol{\theta}})\widehat{q}_g n_1\{n\int_{\mathbf{x},\widetilde{g}}\eta(1,\mathbf{x},\widetilde{g};\widehat{\boldsymbol{\theta}})d\widehat{F}_{\mathbf{x},g}(\mathbf{x},\widetilde{g})\}^{-1}|, \delta\right]}.$$

We define an empirical function $\widetilde{F}_g^\dagger$ whose jump size at $\mathbf{X}_i$ is proportional to

$$\frac{n^{-1}I(G_i = g)}{P(G = g, Y = 0) + \eta(1,\mathbf{X}_i,g;\boldsymbol{\theta}_0)q_g\varrho\{\int_{\mathbf{x},\widetilde{g}}\eta(1,\mathbf{x},\widetilde{g};\boldsymbol{\theta}_0)dF_{\mathbf{x},g}(\mathbf{x},\widetilde{g})\}^{-1}}.$$

Then it can be verified that $\widetilde{F}_g^\dagger$ converges uniformly to $F_g^\dagger$. In addition, $\widehat{F}_g^\dagger$ is absolutely continuous with respect to $\widetilde{F}_g^\dagger$, and the Radon-Nikodym derivative $d\widehat{F}_g^\dagger(\mathbf{x})/d\widetilde{F}_g^\dagger(\mathbf{x})$ is bounded and converges uniformly to $dF_g^*(\mathbf{x})/dF_g^\dagger(\mathbf{x})$. Let $\widetilde{F}_{\mathbf{x},g}(\mathbf{x},g) = \widetilde{F}_g^\dagger(\mathbf{x})q_g$, and let $l_n(\boldsymbol{\theta},\{F_g^\dagger\},\{q_g\})$ be the log-likelihood based on (8). By the definition of the MLE, $n^{-1}l_n(\widehat{\boldsymbol{\theta}},\{\widehat{F}_g^\dagger\},\{\widehat{q}_g\}) - n^{-1}l_n(\boldsymbol{\theta}_0,\{\widetilde{F}_g^\dagger\},\{q_g\}) \ge 0$. The limit of this difference is the negative Kullback-Leibler information of the distribution for

$(\boldsymbol{\theta}^*, \{F_g^*\}, \{q_g^*\})$ with respect to $(\boldsymbol{\theta}_0, \{F_g^\dagger\}, \{q_g\})$ under $P(Y=1)=\varrho$. The identifiability conditions then yield $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0, F_g^* = F_g^\dagger$ and $q_g^* = q_g$. Thus, the consistency of $\widehat{\boldsymbol{\theta}}$ is established. Since $F_{\mathbf{x},g}$ is continuous, $\sup_{\mathbf{x},g} |\widehat{F}_{\mathbf{x},g}(\mathbf{x},g) - F_{\mathbf{x},g}(\mathbf{x},g)| \to 0$ almost surely.

The derivation of the asymptotic distribution is similar to the proof of Theorem 1.2 in Murphy and van der Vaart (2001). We first obtain a score function by differentiating $l_n(\boldsymbol{\theta}, \{F_g^\dagger\}, \{q_g\})$ with respect to $\widehat{\boldsymbol{\theta}}$ along the direction v and with respect to $\widehat{F}_{\mathbf{x},g}$ along the path $\widehat{F}_\epsilon = \widehat{F}_{\mathbf{x},g} + \epsilon \int \psi(\mathbf{x},g) d\widehat{F}_{\mathbf{x},g}$, where v has a unit norm and $\psi(\cdot, g)$ is any function whose total variation is bounded by 1. The linearization of the score function around the true parameter value yields

$$
\begin{aligned}
& n^{1/2} \left\{ (\mathbf{v}^T \boldsymbol{\Omega}_{11} + \boldsymbol{\Omega}_{21}[\psi]^T)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \int (\mathbf{v}^T \boldsymbol{\Omega}_{12} + \Omega_{22}[\psi]) d(\widehat{F}_{\mathbf{x},g} - F_{\mathbf{x},g}) \right\} \\
&= n^{-1/2} \sum_{i=1}^n y_i \left\{ \mathbf{v}^T l_{\boldsymbol{\theta}}(1, \mathbf{X}_i, G_i; \boldsymbol{\theta}_0, F_{\mathbf{x},g}) + l_F(1, \mathbf{X}_i, G_i; \boldsymbol{\theta}_0, F_{\mathbf{x},g}) \Big[\int \psi dF_{\mathbf{x},g} \Big] \right\} \\
&+ n^{-1/2} \sum_{i=1}^n (1-y_i) \left\{ \mathbf{v}^T l_{\boldsymbol{\theta}}(0, \mathbf{X}_i, G_i; \boldsymbol{\theta}_0, F_{\mathbf{x},g}) + l_F(0, \mathbf{X}_i, G_i; \boldsymbol{\theta}_0, F_{\mathbf{x},g}) \Big[\int \psi dF_{\mathbf{x},g} \Big] \right\} + o_p(1),
\end{aligned}
$$

where $\boldsymbol{\Omega}_{11}$ is a constant matrix, $\boldsymbol{\Omega}_{12}$ is a vector function of $\mathbf{x}$, $\boldsymbol{\Omega}_{21}[\psi]$ and $\Omega_{22}[\psi]$ are linear operators of ψ, and $l_{\boldsymbol{\theta}}$ and l_F are the scores with respect to $\boldsymbol{\theta}$ and $F_{\mathbf{x},g}$. The right-hand side of the above equation converges weakly to a Gaussian process, which depends on $(y_1, y_2, \ldots)$ only through ϱ. We can show that the operator $B[v,\psi] \equiv \left\{ \mathbf{v}^T \boldsymbol{\Omega}_{11} + \boldsymbol{\Omega}_{21}[\psi]^T, \mathbf{v}^T \boldsymbol{\Omega}_{12} + \Omega_{22}[\psi] \right\}^T$ is invertible along the lines of Murphy and van der Vaart (2001). It then follows from Theorem 3.3.1 of van der Vaart and Wellner (1996) that $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{F}_{\mathbf{x},g} - F_{\mathbf{x},g})$ converges weakly to a Gaussian process.

Since the asymptotic distribution depends on $(y_1, y_2, \ldots)$ only via ϱ, we assume that $(y_1, y_2, \ldots)$ are independent realizations from a Bernoulli distribution with mean ϱ. By choosing some ψ such that $B[\mathbf{v}, \psi] = (\mathbf{v}^T, 0)^T$ for all v, we see that $\widehat{\boldsymbol{\theta}}$ is an asymptotically linear estimator for $\boldsymbol{\theta}_0$ with the influence function in the score space. 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.

A.4.7. Proof of Theorem 3

We call the probability distribution induced by (9) the pseudo-probability law, denoted by $\widetilde{P}_n$. Let $f(y, \mathbf{x}, g; \boldsymbol{\theta}, \{F_g\}, a_n)$ be the density function under the true probability law P_n. Since $a_n = o(n^{-1/2})$,

$$
\frac{dP_n}{d\widetilde{P}_n} = \exp \left\{ a_n \sum_{i=1}^n \left. \frac{\partial \log f(y_i, \mathbf{X}_i, G_i; \boldsymbol{\theta}, \{F_g\}, a)}{\partial a} \right|_{a=0} + o(1) \right\} \to^{\widetilde{P}_n} 1.
$$

Thus, any weak convergence under $\widetilde{P}_n$ also holds for P_n. On the other hand, by the arguments in the proof of Theorem 2, we can easily verify the results of Theorem 3 when the data is generated from $\widetilde{P}_n$. Thus, Theorem 3 holds when the data is generated from P_n.

A.5. Cohort Studies

A.5.1. Identifiability

We shall show that if two sets of parameters $(\boldsymbol{\theta}, \Lambda)$ and $(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda})$ yield the same joint distribution, then $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$ and $\Lambda = \widetilde{\Lambda}$. First, it follows from Lemma 1 that $\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}}$. Suppose that

$$\sum_{H \in \mathcal{S}(G)} \left\{ \widetilde{\dot{\Lambda}}(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)} \dot{Q}(\widetilde{\Lambda}(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)}) \right\}^{\Delta} \left\{ 1 - Q(\widetilde{\Lambda}(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)}) \right\}^{1-\Delta} P_{\boldsymbol{\gamma}}(H)$$

$$= \sum_{H \in \mathcal{S}(G)} \left\{ \dot{\Lambda}(\widetilde{Y}) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X},H)} \dot{Q}(\Lambda(\widetilde{Y}) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X},H)}) \right\}^{\Delta} \left\{ 1 - Q(\Lambda(\widetilde{Y}) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X},H)}) \right\}^{1-\Delta} P_{\boldsymbol{\gamma}}(H).$$

By choosing $\Delta = 1$ and integrating Y from 0 to τ on both sides, we obtain

$$\sum_{H \in \mathcal{S}(G)} Q(\widetilde{\Lambda}(\tau) e^{\widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)}) P_{\boldsymbol{\gamma}}(H) = \sum_{H \in \mathcal{S}(G)} Q(\Lambda(\tau) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X},H)}) P_{\boldsymbol{\gamma}}(H).$$

Since $Q(\cdot)$ is strictly increasing, the above equation implies that $\widetilde{\Lambda}(\widetilde{Y}) e^{\widetilde{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)} = \Lambda(\widetilde{Y}) e^{\boldsymbol{\beta}^T \mathcal{Z}(\mathbf{X},H)}$ for $H = (h, h)$ and $H = (h, \widetilde{h})$. It then follows from Condition 8 that $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ and $\widetilde{\Lambda} = \Lambda$.

A.5.2. Proof of Theorem 4

Our problem is the same as that of Zeng et al. (2004) except that the integration over random effects in that paper is replaced by the sum over $H \in \mathcal{S}(G)$. The asymptotic properties stated in the theorem will follow from the identifiability shown in Appendix A.5.1 and the proofs of Zeng et al. (2004) provided that we can verify the following result: if there exist a vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\mu}_{\boldsymbol{\gamma}})$ and a function $\psi(t)$ such that

$$\boldsymbol{\mu}^T l_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) + l_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0) \Big[\int \psi d\Lambda_0 \Big] = 0, \tag{A.14}$$

where $l_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$ and $l_{\Lambda}[\int \psi d\Lambda_0]$ is the score function for Λ along the submodel $\Lambda_0 + \epsilon \int \psi d\Lambda_0$, then $\boldsymbol{\mu} = \mathbf{0}$ and $\psi = 0$.

To prove the desired result, we write out equation (A.14). We then let $\Delta = 1$ and integrate Y from 0 to τ to obtain

$$\sum_{H \in \mathcal{S}(G)} \left\{ Q(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}) \right\} P_{\boldsymbol{\gamma}}(H) \left\{ \frac{\dot{Q}(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}) \Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)} \boldsymbol{\mu}_{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X}, H)}{Q(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})} \right.$$

$$\left. + \frac{\dot{Q}(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}) \int_0^{\tau} \psi(t) d\Lambda_0(t) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}}{Q(\Lambda_0(\tau) e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})} + \boldsymbol{\mu}_{\boldsymbol{\gamma}}^T \nabla_{\boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}}(H) \right\} = 0. \tag{A.15}$$

On the other hand, by letting $\Delta = 0$ and $Y = \tau$ in (A.14), we have

$$\sum_{H\in\mathcal{S}(G)} \left\{1 - Q(\Lambda_0(\tau)e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})\right\} P_{\boldsymbol{\gamma}}(H) \left\{ -\frac{\dot{Q}(\Lambda_0(\tau)e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})\Lambda_0(\tau)e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}\boldsymbol{\mu}_{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)}{1 - Q(\Lambda_0(\tau)e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})} \right.$$

$$\left. -\frac{\dot{Q}(\Lambda_0(\tau)e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}) \int_0^\tau \psi(t)d\Lambda_0(t)e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}}{1 - Q(\Lambda_0(\tau)e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})} + \boldsymbol{\mu}_{\boldsymbol{\gamma}}^T \nabla_{\boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}}(H) \right\} = 0. \qquad (A.16)$$

The summation of (A.15) and (A.16) entails $\boldsymbol{\mu}_{\boldsymbol{\gamma}}^T \nabla_{\boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}}(H) = 0$. From the proof of Lemma 1, $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \mathbf{0}$. We choose $G = 2h$ or $h+\widetilde{h}$ and let $\Delta = 1$ and $Y = 0$ in (A.14) to obtain $\boldsymbol{\mu}_{\boldsymbol{\beta}}^T \mathcal{Z}(\mathbf{X},H)+\psi(0) = 0$ for $H = (h,h)$ and $(h,\widetilde{h})$. Thus, $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$ and $\psi(0) = 0$ under Condition 8. Finally, equation (A.14) with $\Delta = 1$ implies that

$$\psi(\widetilde{Y}) + \frac{\ddot{Q}(\Lambda_0(\widetilde{Y})e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}) \int_0^{\widetilde{Y}} \psi(t)d\Lambda_0(t)e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)}}{\dot{Q}(\Lambda_0(\widetilde{Y})e^{\boldsymbol{\beta}_0^T \mathcal{Z}(\mathbf{X},H)})} = 0$$

for $H = (h,h)$. Therefore, $\psi = 0$.

REFERENCES

Akaike, H. (1985) Prediction and entropy. In *A Celebration of Statistics* (eds A. C. Atkinson and S. E. Fienberg), pp. 1-24. New York, Springer.

Akey, J., Jin, L. and Xiong, M. (2001) Haplotypes vs. single marker linkage disequilibrium tests: what do we gain? *Eur. J. Hum. Genet.*, **9**, 291-300.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore, Johns Hopkins University Press.

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

Breslow, N., McNeney, B., and Wellner, J. A. (2003) Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Ann. Stat.*, **31**, 1110-1139.

Clark, A. G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.*, **7**, 111-122.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.,* **20**, 37-46.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Stat. Soc. B.*, **34**, 187-220.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B.*, **39**, 1-38.

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data*, 2nd Ed. Oxford University Press.

Epstein, M. P. and Satten, G. A. (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *Am. J. Hum. Genet.*, **73**, 1316-1329.

Excoffier, L. and Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **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 Res.*, **11**, 143-151.

Fisher, R. A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh,* **52**, 399-433.

Hallman, D. M., Groenemeijer, B. E., Jukema, J. W. and Boerwinkle, E. (1999) Analysis of lipoprotein lipase haplotypes reveals associations not apparent from analysis of the constitute loci. *Ann. Hum. Genet.*, **63**, 499-510.

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

International SNP Map Working Group. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928-933.

Lake, S. L., Lyon, H., Tantisira, K., Silverman, E. K., Weiss, S. T., Laird, N. M., and Schaid, D. J. (2003) Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum. Hered.*, **55**, 56-65.

Li, H. (2001) A permutation procedure for the haplotype method for identification of disease-predisposing variants. *Ann. Hum. Genet.*, **65**, 180-196.

Liang, K.-Y. and Qin, J. (2000) Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *J. R. Statist. Soc. B*, **62**, 773-786.

Lin, D. Y. (2004) Haplotype-based association analysis in cohort studies of unrelated individuals. *Genet. Epidemiol.*, **26**, 255-264.

Lin, D. Y. (2004) An efficient Monte Carlo approach to assessing statistical significance in genomic

studies. *Bioinformatics*, in press.

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

Morris, R. W. and Kaplan, N. L. (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet. Epidemiol.*, **23**, 221-233.

Murphy, S. A. and van der Vaart, A. W. (2000) On profile likelihood. *J. Am. Statist. Ass.*, **95**, 449-465.

Murphy, S. A. and van der Vaart, A. W. (2001) Semiparametric mixtures in case-control studies. *J. Mult. Var. Anal.*, **79**, 1-32.

Niu, T., Qin, Z. S., Xu, X. and Liu, J. S. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **70**, 157-169.

Patil, N., Berno, A. J., Hinds, D. A., et al. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719-1723.

Pettitt, A. N. (1984) Proportional odds models for survival data and estimates using ranks. *Applied Statist.*, **26**, 183-214.

Prentice, R. L. and Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika,* **66**, 403-411.

Qin, Z. S., Niu, T. and Liu, J. S. (2002) Partition-ligation-expectation maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **71**, 1242-1247.

Risch, N. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847-856.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996) A semiparametric mixture approach to case-control studies with errors in covariables. *J. Am. Statist. Assoc.*, **91**, 722-732.

Satten, G. A. and Epstein, M. P. (2004) Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet. Epidemiol.*, **27**, 192-201.

Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. *Genet. Epidemiol.*, **27**, 348-364.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. and Poland G. A. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, **70**, 425-434.

Scott, A. J. and Wild, C. J. (1997) Fitting regression models to case-control data by maximum likelihood. *Biometrika*, **84**, 57-71.

Seltman, H., Roeder, K. and Devlin, B. (2003) Evolutionary-based association analysis using haplotype data. *Genet. Epidemiol.*, **25**, 48-58.

Stephens, M., Smith, N. J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.,* **68**, 978-989.

Stram, D. O., Pearce, C. L., Bretsky, P., et al. (2003) Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum. Hered.*, **55**, 179-190.

Valle, T., Tuomilehto, J., Bergman, R., et al. (1998) Mapping genes for NIDDM. *Diabetes Care,* **21**, 949-958.

van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes.* New York, Springer.

van der Vaart, A. W. and Wellner, J. A. (2001) Consistency of semiparametric maximum likelihood estimators for two-phase sampling. *Canad. J. Statist.* **29**, 269-288.

Venter, J., Adams, M., Myers, E., Li, P., Mural, R., Sutton, G., Smith, H., et al. (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.

Wang, S., Kidd, K. and Zhao, H. (2003) On the use of DNA pooling to estimate haplotype frequencies. Genet Epidemiol, **24**, 74-82.

Weir, B. S. (1996) *Genetic Data Analysis II.* Sunderland, Sinauer Associates, Inc. Publishers.

Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J. and Ehm, M. G. (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.*, **53**, 79-91.

Zeng, D. Lin, D. Y. and Lin, X. (2004) Semiparametric transformation models with random effects for clustered failure time data. Unpublished technical report.

Zhang, S., Pakstis, A. J., Kidd, K. K. and Zhao, H. (2001) Comparisons of two methods for haplotype recontruction and haplotype frequency estimation from population data. *Am. J. Hum. Genet.*, **69**, 906-912.

Zhao, L. P., Li, S. 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.

Am. J. Hum. Genet., **72**, 1231-1250.