Published in final edited form as:

J Am Stat Assoc. 2009 September 1; 104(487): 1251–1260. doi:10.1198/jasa.2009.tm08507.

An Incomplete-Data Quasi-likelihood Approach to Haplotype-Based Genetic Association Studies on Related Individuals

Zuoheng Wang [Ph.D. Candidate] and

Department of Statistics, University of Chicago, Chicago, IL 60637 (E-mail: zwang@galton.uchicago.edu).

Mary Sara McPeck [Professor]

Departments of Statistics and Human Genetics, University of Chicago, Chicago, IL 60637 (E-mail: mcpeek@galton.uchicago.edu).

Abstract

We propose an incomplete-data, quasi-likelihood framework, for estimation and score tests, which accommodates both dependent and partially-observed data. The motivation comes from genetic association studies, where we address the problems of estimating haplotype frequencies and testing association between a disease and haplotypes of multiple tightly-linked genetic markers, using case-control samples containing related individuals. We consider a more general setting in which the complete data are dependent with marginal distributions following a generalized linear model. We form a vector $\mathbf{Z}$ whose elements are conditional expectations of the elements of the complete-data vector, given selected functions of the incomplete data. Assuming that the covariance matrix of $\mathbf{Z}$ is available, we form an optimal linear estimating function based on $\mathbf{Z}$, which we solve by an iterative method. This approach addresses key difficulties in the haplotype frequency estimation and testing problems in related individuals: (1) dependence that is known but can be complicated; (2) data that are incomplete for structural reasons, as well as possibly missing, with different amounts of information for different observations; (3) the need for computational speed in order to analyze large numbers of markers; (4) a well-established null model, but an alternative model that is unknown and is problematic to fully specify in related individuals. For haplotype analysis, we give sufficient conditions for consistency and asymptotic normality of the estimator and asymptotic χ^2 null distribution of the score test. We apply the method to test for association of haplotypes with alcoholism in the GAW 14 COGA data set.

Keywords

Estimating function; Association testing; Dependent data; Missing data; Generalized linear model; Score test

1 INTRODUCTION

Genetic association analysis is an important tool for searching for genetic variants that play a role in human complex disease. Typical approaches compare the frequency distribution of marker alleles, genotypes or haplotypes between cases and controls. A genetic *marker* is a segment of DNA sequence with an identifiable physical location, and it is said to be *polymorphic* if its sequence varies across the population. Each of the variant forms of a polymorphic marker is called an *allele*. For autosomal genetic markers, a human inherits two copies of the marker, one from each parent. An individual's *genotype* at a marker is the observation of the two alleles of the individual at that marker. The *kinship coefficient* between a pair of related individuals is determined by their relationship and is the probability that, at

any given marker, a randomly chosen pair of alleles, one from each individual, is identical by descent, i.e. is an inherited copy of the same founder allele. For example, the kinship coefficient for a parent-offspring or sibling pair is $1/4$, while that for a grandparent-grandchild, avuncular, or half-sibling pair is $1/8$. In genetic association analysis, the most widely-used type of marker is currently the *single nucleotide polymorphism* (SNP), which is a DNA sequence variation that occurs at a single nucleotide and generally has two alleles. When q markers are considered simultaneously, the ordered q -tuple of alleles inherited from the same parent defines a *haplotype*. SNP haplotypes are more informative for disease mapping than are single SNPs. By capturing the dependence structure among multiple markers, haplotypes potentially provide information on untyped variants as well as on interactions among tightly-linked, typed variants. However, standard genotyping methods do not provide the information of the individual's two haplotypes. Only the unphased genotype (i.e. the observation of the two alleles at each marker, but not their parental origin) is generated. As a consequence, we are faced with an incomplete-data problem for haplotype analysis.

There is a vast literature on statistical methods for analysis of haplotype association with genetic traits in samples of unrelated individuals; a relatively recent review paper is Schaid (2004). For related individuals, the approaches can roughly be divided into two types: "case-control," which is the type of approach we pursue, and "family-based." Family-based haplotype association tests, such as FBAT (Horvath et al. 2004), condition on both trait values and parental genotypes. When some or all parental genotype data are missing, the analysis is conditional on sufficient statistics for parental genotypes. Generally, this conditioning approach has the advantage of making the analysis more robust to population structure and the disadvantage of losing power compared to case-control studies (Risch and Teng 1998). Family-based methods also typically require genotype data for family members of an affected individual. Thus they do not allow unrelated individuals in the analysis. In addition to being more powerful, case-control analyses are less restrictive than family-based approaches, because they can allow but do not require genotype data for relatives of affected individuals. For case-control studies that include related individuals, where general relationship types are allowed, methods for single-marker allele frequency estimation (McPeck, Wu, and Ober 2004) and for single-marker association testing (Slager and Schaid 2001; Bourgain et al. 2003; Thornton and McPeck 2007) have been proposed. Haplotype-based association analysis for case-control samples consisting of related individuals has been considered by Browning et al. (2005) and Zhang, Schneider, Ober, and McPeck (2005).

Even for single-marker genetic analysis, sampling of related individuals can introduce complicated dependence structure. In addition, current genetic association analyses typically consider on the order of 10^4 to 10^6 markers throughout the genome, putting a premium on computational speed of inference. Furthermore, although the genetic model of the disease is typically unknown, the first and second moment structure of genotype data, under the null hypothesis of no association and no linkage, is a known function of the pedigree information and the unknown allele frequency. For these reasons, quasi-likelihood estimation and quasi-likelihood score tests have proved useful for single-marker, case-control, genetic association studies with related individuals (Bourgain et al. 2003; McPeck et al. 2004; Thornton and McPeck 2007). Quasi-likelihood inference (reviewed in McCullagh and Nelder 1989, chap. 9; Heyde 1997) was introduced by Wedderburn (1974) and is closely connected to the theory of optimal estimating functions (Godambe 1960). To extend previous quasi-likelihood methods to haplotype-based case-control association testing in related individuals, we need to allow for both incomplete data and dependence of observations. Recent work on estimating equations in the context of incomplete data includes Heyde and Morton (1996) and Elashoff and Ryan (2004).

We propose an incomplete-data quasi-likelihood (IQL) framework to address both dependent and partially-observed data. Under a generalized linear model for the marginal distributions of the complete data, we form an optimal linear estimating function, which we call an incomplete-data quasi-likelihood score (IQLS) function, based on certain conditional expectations of the elements of the complete-data vector. In doing this, we take advantage of the availability of a correctly-specified covariance matrix in the genetics applications we consider. We provide detailed development of the method for the specific contexts of haplotype frequency estimation and testing for association of haplotypes with a genetic disease, and we apply it to test for association of haplotypes with alcoholism in the GAW 14 COGA data set.

2 METHODS

2.1 Incomplete-data Quasi-likelihood Score Function

Our genetics applications involve genotypes and haplotypes, which are modeled using dependent binomial random variables whose parameters depend on covariates. In this section, we develop our IQLS methods for a more general setting of dependent observations, in which the marginal distribution of each observation is a generalized linear model.

Suppose that the complete data consist of pairs $(Y_i, \mathbf{x}_i)$, $i = 1, \dots, n$, where the Y_i 's are scalar responses and the $\mathbf{x}_i$'s are $r \times 1$ covariate vectors that are treated as fixed in the analysis. Suppose the Y_i 's are dependent with each Y_i having marginal log-likelihood given by $l_{Y_i}(y_i | \mathbf{x}_i, \boldsymbol{\beta}) = y_i \theta_i - a(\theta_i) + b(y_i)$, where $\theta_i = u(\eta_i)$, u is a known injective function, $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, and $\boldsymbol{\beta}$ is an $r \times 1$ parameter vector. The marginal score function with respect to the natural parameter θ_i is $\partial l_{Y_i} / \partial \theta_i = y_i - \mu_i$, where $\mu_i = E Y_i = a'(\theta_i)$. Furthermore $-E(\partial^2 l_{Y_i} / \partial \theta_i^2) = \ddot{a}(\theta_i) = \text{Var}(Y_i) = \text{Var}(\partial l_{Y_i} / \partial \theta_i)$. Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$, $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{Y})$, and assume that $\boldsymbol{\Sigma}$ can be specified in terms of $\boldsymbol{\beta}$.

We emphasize the following key features of the inference problems we consider: (1) The effects of interest can be described by a model for $\boldsymbol{\mu}$ as a known function of the unknown parameter $\boldsymbol{\beta}$. (2) The matrix $\boldsymbol{\Sigma}$ is also a known function of $\boldsymbol{\beta}$. This latter point distinguishes our setting from, for example, that assumed in the GEE approach (Liang and Zeger 1986; Zeger and Liang 1986) in which $\boldsymbol{\Sigma}$ is generally unknown. The assumption that $\boldsymbol{\Sigma}$ is a known function of $\boldsymbol{\beta}$ is applicable to practical applications in genetics such as allele and haplotype frequency estimation and association mapping.

In our setting, the quasi-likelihood score function is

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{D}^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})). \quad (1)$$

Here $\mathbf{D}$ is the derivative matrix $\partial \boldsymbol{\mu} / \partial \boldsymbol{\beta}$. Even when the marginal log-likelihood l_{Y_i} is not correctly specified, provided that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the correctly-specified moments of $\mathbf{Y}$, the quasi-likelihood score function is optimal in the class of linear unbiased estimating functions $\{\mathbf{H}^T(\boldsymbol{\beta})(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) : \mathbf{H}(\boldsymbol{\beta}) \text{ is an } n \times r \text{ matrix}\}$, in the sense that it has maximum information, where the information of an unbiased estimating function $\mathbf{G}$ is defined as $\Lambda(\mathbf{G}) = E(\partial \mathbf{G} / \partial \boldsymbol{\beta})^T [E(\mathbf{G} \mathbf{G}^T)]^{-1} E(\partial \mathbf{G} / \partial \boldsymbol{\beta})$, and the maximum is taken in the partial order of nonnegative definite matrices. In our setting, $\Lambda(\mathbf{U}) = \mathbf{D}^T \boldsymbol{\Sigma}^{-1} \mathbf{D}$. A quasi-likelihood estimator $\hat{\boldsymbol{\beta}}$ is obtained by solving $\mathbf{U}(\boldsymbol{\beta}) = 0$. Under regularity conditions, $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically efficient.

We are particularly interested in the case when the complete data $\mathbf{Y}$ are partially observed. In that case, let $\mathbf{I}$ be the observed data (incomplete data). In some contexts, it might be natural to express $\mathbf{I}$ as $\mathbf{I} = (I_1, \dots, I_n)$, where I_i is a deterministic function of Y_i , but this is not a requirement.

Indeed, we need not even require that $\mathbf{I}$ be a deterministic function of $\mathbf{Y}$ at all. For instance, we could allow $\mathbf{I}$ to consist of noisy observations on functions of $\mathbf{Y}$.

We use the following approach to form incomplete-data estimating functions. We define $\mathbf{Z} = (Z_1, \dots, Z_n)^T$ by $Z_i = E(Y_i | f_i(\mathbf{I}))$, where $f_i(\mathbf{I})$ is some function of $\mathbf{I}$. As a special case, we could have $f_i(\mathbf{I}) = \mathbf{I}$ for all $1 \leq i \leq n$, leading to $\mathbf{Z} = E(\mathbf{Y} | \mathbf{I})$. In the general case, we note that $E(\mathbf{Z}) = \boldsymbol{\mu}$, and we define $\boldsymbol{\Omega} = \text{Cov}(\mathbf{Z})$ and $\mathbf{F} = -E(\partial(\mathbf{Z} - \boldsymbol{\mu})/\partial\boldsymbol{\beta})$. We then form an IQLS function,

$$\mathbf{U}_I = \mathbf{F}^T \boldsymbol{\Omega}^{-1} (\mathbf{Z} - \boldsymbol{\mu}), \quad (2)$$

where $\boldsymbol{\Omega}^{-1}$ is replaced by the Moore-Penrose generalized inverse if $\boldsymbol{\Omega}$ is singular. An IQL estimator for $\boldsymbol{\beta}$ is a solution of $\mathbf{U}_I(\boldsymbol{\beta}) = 0$. We can typically obtain a solution $\hat{\boldsymbol{\beta}}$ by an iterative algorithm that is analogous to Newton-Raphson with Fisher scoring (Wedderburn 1974). Starting at $\boldsymbol{\beta}^{(0)}$ sufficiently close to $\hat{\boldsymbol{\beta}}$, we iterate according to

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + [(\mathbf{F}^T \boldsymbol{\Omega}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \boldsymbol{\Omega}^{-1} (\mathbf{Z} - \boldsymbol{\mu})]_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}, \quad (3)$$

where $\boldsymbol{\beta}^{(t)}$ is the parameter estimate at the t th iteration.

The IQLS function has information

$$\Lambda(\mathbf{U}_I) = \text{Cov}(\mathbf{U}_I) = -E(\partial \mathbf{U}_I / \partial \boldsymbol{\beta}) = \mathbf{F}^T \boldsymbol{\Omega}^{-1} \mathbf{F} \quad (4)$$

and is the optimal estimating function within the class of linear, unbiased, estimating functions $H_I = \{\mathbf{H}^T(\boldsymbol{\beta})(\mathbf{Z} - \boldsymbol{\mu}) : \mathbf{H} \text{ is an } n \times r \text{ matrix}\}$. The optimality depends on correct specification of the first and second moments of $\mathbf{Z}$ and does not require correct specification of other aspects of the distribution. Yuan and Jennrich (1998) give conditions for existence, consistency, and asymptotic normality of estimators from a wide class of estimating functions that includes Equation (2). The conditions of Xie and Yang (2003) for GEE estimators can also be adapted to IQL estimators. For the haplotype estimation and association testing problems in section 3, we directly verify the conditions of Yuan and Jennrich (1998).

Consider the special case when $\mathbf{I}$ can be expressed as $\mathbf{I} = (I_1, \dots, I_n)$, where $I_i = g_i(Y_i)$, with g_i a deterministic, measurable function not depending on $\boldsymbol{\beta}$. Then the following three properties follow (e.g. from Theorem 1.5.8 of Lehmann and Casella 1998), where l_i is the log-likelihood of I_i : (1) $\partial l_i / \partial \theta_i = E(\partial l_{Y_i} / \partial \theta_i | I_i)$, (2) $E(\partial l_i / \partial \theta_i) = 0$, and (3) $-E(\partial^2 l_i / \partial \theta_i^2) = \text{Var}(\partial l_i / \partial \theta_i)$. Therefore, if we choose $f_i(\mathbf{I}) = I_i$, then $Z_i - \mu_i = \partial l_i / \partial \theta_i$, and $\mathbf{Z} - \boldsymbol{\mu}$ can be viewed as the vector of incomplete-data, marginal score functions. In this special case, $\mathbf{U}_I$ arises naturally as the optimal, linear, un-biased, estimating function based on the vector of incomplete-data, marginal score functions, while the complete-data estimating function $\mathbf{U}$ was the optimal, linear, unbiased, estimating function based on the vector of complete-data, marginal score functions. If, in addition, we restrict Y_i , $1 \leq i \leq n$ to be independent, then $\mathbf{U}_I$ is the incomplete-data, likelihood score function, and the IQL estimators are the maximum likelihood estimators (MLEs).

One would expect that the optimal choice of $f_i(\mathbf{I})$ in the IQLS function would be $f_i(\mathbf{I}) = \mathbf{I}$, $1 \leq i \leq n$. However, in some problems (for example, in many instances of the haplotype inference problems we consider), $E(\mathbf{Y} | \mathbf{I})$ is not computationally practical to obtain, so the IQLS function with $f_i(\mathbf{I}) = \mathbf{I}$ is not practically useful in such problems. In that case, we ideally want to choose

f_i so that (1) $E(Y_i|\mathbf{I})$ is well-approximated by $E(Y_i|f_i(\mathbf{I}))$ and (2) computation of $Z_i = E(Y_i|f_i(\mathbf{I}))$ and the resulting $\mathbf{\Omega}$ and $\mathbf{F}$ are not onerous.

2.2 Comparison with Previous Approaches

Heyde and Morton (1996) consider the problem of quasi-likelihood estimation with incomplete data. Suppose that $\mathbf{Q}$ is the complete-data quasi-score within some convex class H_Y of unbiased, square-integrable, complete-data, estimating equations. Given a convex class H_I of unbiased, square-integrable, incomplete-data, estimating equations, where H_I is a linear subspace of H_Y , Heyde and Morton (1996) show that the quasi-score within H_I is the projection of $\mathbf{Q}$ into H_I . However, dependence across observations would often lead to $H_I \not\subset H_Y$ (e.g. in the haplotype analysis setting we consider), in which case the projection result does not hold.

Elashoff and Ryan (2004) give an approach to constructing an estimating function for incomplete data based on an estimating function for complete data, with a practical computational method for obtaining the estimate. Applying the Elashoff-Ryan approach to our setting would yield estimating function $\mathbf{U}_{ER} = \mathbf{E}(\mathbf{U}|\mathbf{I}) = \mathbf{D}^T \mathbf{\Sigma}^{-1} [\mathbf{E}(\mathbf{Y}|\mathbf{I}) - \boldsymbol{\mu}]$, with information $\mathbf{\Lambda}(\mathbf{U}_{ER}) = \mathbf{F}^T \mathbf{\Sigma}^{-1} \mathbf{D} (\mathbf{D}^T \mathbf{\Sigma}^{-1} \mathbf{\Omega} \mathbf{\Sigma}^{-1} \mathbf{D})^{-1} \mathbf{D}^T \mathbf{\Sigma}^{-1} \mathbf{F}$. One drawback of applying the Elashoff-Ryan approach to the haplotype inference problems we consider is that it is not always practical to obtain $E(\mathbf{Y}|\mathbf{I})$. Another consideration is that $\mathbf{U}_{ER}$ is, in general, sub-optimal in the class of linear unbiased estimating functions $\{\mathbf{H}^T(\boldsymbol{\beta})[E(\mathbf{Y}|\mathbf{I}) - \boldsymbol{\mu}(\boldsymbol{\beta})] : \mathbf{H}(\boldsymbol{\beta}) \text{ is an } n \times r \text{ matrix}\}$. In our setting, when it is practical to obtain $E(\mathbf{Y}|\mathbf{I})$, the IQLS function will be optimal in that class. Thus, for example, if $\hat{\boldsymbol{\beta}}$ is the IQL estimator and $\tilde{\boldsymbol{\beta}}$ solves $\mathbf{U}_{ER} = 0$, then the relative efficiency of $\mathbf{c}^T \hat{\boldsymbol{\beta}}$ compared to the $\mathbf{c}^T \tilde{\boldsymbol{\beta}}$ (or equivalently, the ratio of non-centrality parameters for testing in the direction specified by $\mathbf{c}$), is $[\mathbf{c}^T \mathbf{\Lambda}(\mathbf{U}_I) \mathbf{c}] / [\mathbf{c}^T \mathbf{\Lambda}(\mathbf{U}_{ER}) \mathbf{c}] \geq 1$, for every choice of real vector $\mathbf{c}$ of length r , where $\mathbf{\Lambda}(\mathbf{U}_I)$ is given by Equation (4).

2.3 Hypothesis Testing

We also consider the problem of hypothesis testing in the case when the unknown parameters $\boldsymbol{\beta}$ can be partitioned as $\boldsymbol{\beta} = (\boldsymbol{\gamma}^T, \boldsymbol{\alpha}^T)^T$, where $\boldsymbol{\gamma}$ are the parameters of interest and $\boldsymbol{\alpha}$ are nuisance parameters. Suppose we want to test the null hypothesis that $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$. We partition the vector $\mathbf{U}_I$ into $\mathbf{U}_{I\gamma} = \mathbf{F}_\gamma^T \mathbf{\Omega}^{-1} (\mathbf{Z} - \boldsymbol{\mu})$ and $\mathbf{U}_{I\alpha} = \mathbf{F}_\alpha^T \mathbf{\Omega}^{-1} (\mathbf{Z} - \boldsymbol{\mu})$, where $\mathbf{F}_\gamma = -E(\partial(\mathbf{Z} - \boldsymbol{\mu})/\partial\boldsymbol{\gamma})$ and $\mathbf{F}_\alpha = -E(\partial(\mathbf{Z} - \boldsymbol{\mu})/\partial\boldsymbol{\alpha})$. Under the null hypothesis, the nuisance parameter $\boldsymbol{\alpha}$ is estimated by $\hat{\boldsymbol{\alpha}}_0$ which solves $\mathbf{U}_{I\alpha} = \mathbf{0}$ when $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$. Consider the information matrix $\mathbf{\Lambda}(\mathbf{U}_I)$ of Equation (4), which we assume is positive definite. We can think of $\mathbf{\Lambda}$ as being composed of four blocks: $\mathbf{\Lambda}_{\gamma\gamma}$ which is the $(\boldsymbol{\gamma}, \boldsymbol{\gamma})$ th diagonal block of $\mathbf{\Lambda}$, $\mathbf{\Lambda}_{\alpha\alpha}$ which is the $(\boldsymbol{\alpha}, \boldsymbol{\alpha})$ th diagonal block of $\mathbf{\Lambda}$, and two off-diagonal blocks: $\mathbf{\Lambda}_{\gamma\alpha}$ and $\mathbf{\Lambda}_{\alpha\gamma}$, which are transposes of each other. Similarly, we partition $\mathbf{\Lambda}^{-1}$ into $\mathbf{\Lambda}^{\gamma\gamma}$, $\mathbf{\Lambda}^{\alpha\alpha}$, and $\mathbf{\Lambda}^{\gamma\alpha}$, $\mathbf{\Lambda}^{\alpha\gamma}$. Then, in analogue to the classical likelihood-based score statistic in the presence of nuisance parameters, the IQLS statistic has the following form

$$W = \mathbf{U}_{I\gamma}^T(\boldsymbol{\gamma}_0, \hat{\boldsymbol{\alpha}}_0) \mathbf{\Lambda}^{\gamma\gamma}(\boldsymbol{\gamma}_0, \hat{\boldsymbol{\alpha}}_0) \mathbf{U}_{I\gamma}(\boldsymbol{\gamma}_0, \hat{\boldsymbol{\alpha}}_0), \quad (5)$$

where $\mathbf{\Lambda}^{\gamma\gamma}$ can be expressed as $\mathbf{\Lambda}^{\gamma\gamma} = (\mathbf{\Lambda}_{\gamma\gamma} - \mathbf{\Lambda}_{\gamma\alpha} \mathbf{\Lambda}_{\alpha\alpha}^{-1} \mathbf{\Lambda}_{\alpha\gamma})^{-1}$. More explicitly, we obtain the following equivalent form of the IQLS test statistic

$$W = \{(\mathbf{Z} - \boldsymbol{\mu})^T \mathbf{\Omega}^{-1} \mathbf{F}_\gamma [\mathbf{F}_\gamma^T \mathbf{\Omega}^{-1} \mathbf{F}_\gamma - \mathbf{F}_\gamma^T \mathbf{\Omega}^{-1} \mathbf{F}_\alpha (\mathbf{F}_\alpha^T \mathbf{\Omega}^{-1} \mathbf{F}_\alpha)^{-1} \mathbf{F}_\alpha^T \mathbf{\Omega}^{-1} \mathbf{F}_\gamma]^{-1} \mathbf{F}_\gamma^T \mathbf{\Omega}^{-1} (\mathbf{Z} - \boldsymbol{\mu})\}_{(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = (\boldsymbol{\gamma}_0, \hat{\boldsymbol{\alpha}}_0)},$$

where the entire right-hand side is evaluated at $(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = (\boldsymbol{\gamma}_0, \hat{\boldsymbol{\alpha}}_0)$. Under regularity conditions, the asymptotic null distribution of W is χ^2 with degrees of freedom equal to the dimension of $\boldsymbol{\gamma}$. (For the haplotype association testing problem, sufficient conditions are given in the

Appendix.) Note that in order to obtain W , it is not necessary to have a fully specified joint alternative distribution. It is sufficient to provide μ under the alternative and $(\partial \mathbf{Z} / \partial \beta)_{(\gamma, \alpha) = (\gamma_0, \hat{\alpha}_0)}$, which, in the haplotype testing application, can be derived under relatively weak assumptions on the alternative model.

3 APPLICATION TO HAPLOTYPE ANALYSIS

We apply the IQLS method to the problems of haplotype frequency estimation (sub-section 3.1) and testing for association between haplotypes and a binary genetic trait (subsection 3.2) in samples that contain related individuals. The following general set-up applies to both problems.

Suppose we have a sample of m outbred individuals, some of whom may be related with relationships specified by known pedigrees. (We assume a random sampling scheme under which the event that any particular set of individuals in the population is sampled is conditionally independent of their haplotypes, given the pedigree information and covariate values $\mathbf{x}$.) Consider a set of q ordered, tightly-linked SNPs, where each SNP is a binary marker with alleles labeled 0 and 1. (Here, *tightly-linked* means that recombination among the SNPs is observed only over large time scales of many generations and is not observed within families.) Let $\mathcal{H} = \{\mathbf{h}_1, \dots, \mathbf{h}_{a+1}\} \subset \{0, 1\}^q$ be the set of $a + 1$ distinct possible observed haplotypes for the q SNPs. In practice, we can have $a + 1 < 2^q$ when certain haplotypes do not occur in nature. For simplicity, we assume that $\mathcal{H}$ is known. The complete-data vector of length $n = ma$ is taken to be $\mathbf{Y} = (Y_{11}, \dots, Y_{1a}, \dots, Y_{m1}, \dots, Y_{ma})^T$, where $Y_{ij} = .5 \times$ (the number of copies of haplotype $\mathbf{h}_j$ held by individual i), for $1 \leq i \leq m$, $1 \leq j \leq a + 1$. Under the assumption of Hardy-Weinberg equilibrium, the marginal model for Y_{ij} is $.5 \times \text{Bin}(2, p_{ij})$, where

$p_{ij} > 0$, $\sum_{j=1}^{a+1} p_{ij} = 1$. The resulting marginal log-likelihood in terms of the natural parameter, $\theta_{ij} = 2 \logit(p_{ij})$, is $l_{Y_{ij}}(y) = y\theta_{ij} - 2 \log(1 + \exp(\theta_{ij}/2))$ plus a term not depending on θ_{ij} , and we use the canonical link function, so that $\theta_{ij} = 2\mathbf{x}_{ij}^T \beta$.

For individual i , $1 \leq i \leq m$, let $\mathbf{H}_i = (\mathbf{H}_{i1}, \mathbf{H}_{i2})$ denote i 's ordered pair of haplotypes. Instead of observing $\mathbf{H}_i$ for each i , we observe only the q -vector of SNP genotypes $\mathbf{G}_i = \mathbf{H}_{i1} + \mathbf{H}_{i2}$, provided i 's genotype is not missing at any marker. If i 's genotype is missing at one or more markers, then the corresponding entries of the vector $\mathbf{G}_i$ are not directly observed. Let $\mathbf{M}$ be the mq -vector $\mathbf{M} = (M_{11}, \dots, M_{1q}, \dots, M_{m1}, \dots, M_{mq})$, where M_{ij} is the indicator of missingness of i 's genotype at the j th SNP. We assume that genotype data are missing at random, i.e. conditional on the pedigree information and on $\mathbf{x}$, the vector $\mathbf{M}$ is independent of $\mathbf{Y}$, and we condition the analyses on $\mathbf{M}$, $\mathbf{x}$, and the pedigree information. We denote the observed data by $\mathbf{I} = (\mathbf{I}_1, \dots, \mathbf{I}_m)$, where $\mathbf{I}_i$ is the information of the values of the nonmissing entries of $\mathbf{G}_i$.

3.1 Haplotype Frequency Estimation

We first consider the problem of estimating the haplotype frequency distribution $\mathbf{p} = (p_1, \dots, p_a)^T$ for the population from which the m sampled individuals are drawn, where we assume P

$(\mathbf{H}_{ik} = \mathbf{h}_j) = p_j$, for $i = 1, \dots, m$, $k = 1, 2$, $j = 1, \dots, a + 1$ and $p_{a+1} = 1 - \sum_{i=1}^a p_i$. We take $\mathbf{x}_{ij} = \mathbf{e}_j$, the vector of length a having j th entry equal to 1 and all other entries 0, for $1 \leq i \leq m$, $1 \leq j \leq a$. Then $\theta_{ij}/2 = \logit(p_j) = \beta_j$, so β_j is a one-to-one function of p_j for $1 \leq j \leq a$. We have $\mu = \mathbf{1}_m \otimes \mathbf{p}$, where $\mathbf{1}_m$ is an m -vector of 1's, and $\otimes$ is Kronecker product. Mendelian inheritance in the pedigree implies that $\Sigma = \mathbf{K} \otimes \mathbf{B}$, where $\mathbf{K}_{m \times m}$, which is the correlation matrix of the m -vector $(Y_{1j}, \dots, Y_{mj})^T$ for each $1 \leq j \leq a$, is given by $K_{ij} = 2\phi_{ij}$, where ϕ_{ij} is the kinship coefficient between individuals i and j , with $2\phi_{ii} = 1$. (We will also refer to $\mathbf{K}$ as the *kinship matrix*.) Here $\mathbf{B}_{a \times a}$, which is the covariance matrix of the a -vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ia})^T$ for each $1 \leq i \leq m$, is given by $B_{ij} = p_i(1 - p_i)/2$ if $i = j$ and $-p_i p_j/2$ if $i \neq j$. We have $\mathbf{D} = \mathbf{1}_m \otimes (\partial \mathbf{p} / \partial \beta)$,

where $\partial \mathbf{p} / \partial \boldsymbol{\beta}$ is the $a \times a$ diagonal matrix with j th diagonal element $p_j(1 - p_j)$. The complete-data quasi-likelihood score function is $\mathbf{U}(\boldsymbol{\beta}) = (\partial \mathbf{p} / \partial \boldsymbol{\beta}) \mathbf{U}(\mathbf{p})$, where $\mathbf{U}(\mathbf{p})$ is the quasi-likelihood score function in terms of $\mathbf{p}$ (used by McPeck et al. 2004) which is given by

$\mathbf{U}(\mathbf{p}) = \sum_{i=1}^m w_i \mathbf{B}^{-1}(\mathbf{Y}_i - \mathbf{p})$, where $w_i = (\mathbf{1}_m^T \mathbf{K}^{-1})_i$. $\mathbf{U}(\mathbf{p})$ and $\mathbf{U}(\boldsymbol{\beta})$ are equivalent in that they produce the same estimators of $\mathbf{p}$. Here $\mathbf{K}$ is invertible provided the sample does not contain MZ twins (Thornton and McPeck 2007), and the Moore-Penrose generalized inverse is used when the sample does contain MZ twins. For the case when complete haplotype data are available, the quasi-likelihood estimator, $\hat{\mathbf{p}}$ which solves $\mathbf{U}(\mathbf{p}) = 0$, is given explicitly by McPeck et al. (2004).

With incomplete data, we form $\mathbf{Z} = (Z_{11}, \dots, Z_{1a}, \dots, Z_{m1}, \dots, Z_{ma})^T$, where $Z_{ij} = E[Y_{ij} | f_{ij}(\mathbf{I})]$, for some functions f_{ij} , $1 \leq i \leq m$, $1 \leq j \leq a$. In the special case of unrelated individuals, provided $f_{ij}(\mathbf{I})$ contains at least the information of $\mathbf{I}_i$, the IQL haplotype frequency estimator is the same as the MLE, and, when the iterative algorithm of Equation (3) is applied to solve $\mathbf{U}_f(\mathbf{p}) = 0$, it is the same as the EM algorithm. For large complex pedigrees, the optimal choice, $f_{ij}(\mathbf{I}) = \mathbf{I}_i$, is not practical from a computational point of view. We consider two other choices: (a) $f_{ij}(\mathbf{I}) = \mathbf{I}_i$, the non-missing SNP genotypes of individual i , and (b) $f_{ij}(\mathbf{I}) = (\mathbf{I}_i, \mathbf{I}_{m(i)}, \mathbf{I}_{f(i)})$, the non-missing SNP genotypes of individuals $i, m(i)$, and $f(i)$, where $m(i)$ and $f(i)$ are the mother and father, respectively, of i . We expect choice (b) to improve efficiency of estimation over choice (a), and this is explored in the simulations of section 4. In the Appendix, we give sufficient conditions for existence, consistency, and asymptotic normality of the IQL haplotype frequency estimator.

3.2 Case-Control Association Testing: M_{IQLS}

We consider the problem of testing for association between a binary trait, which we call the *phenotype* (typically a disease with some genetic component), and haplotypes for a set of tightly-linked SNPs. We use the IQLS approach to extend the single-marker, complete-data M_{QLS} test of Thornton and McPeck (2007) to the M_{IQLS} test based on haplotypes with incomplete data. We allow for missing phenotypes and missing genotypes. Phenotype is treated as a covariate, and the analysis is conditioned on it. Let $\mathbf{A}$ be the m -vector of phenotype information, with $A_i = 1 - k$ if i is affected, $-k$ if i is unaffected, and 0 if i 's phenotype is missing, where k is an estimate of prevalence of the disease in the relevant reference population. $\mathbf{A}$ can be thought of as the mean-corrected phenotype indicator with adjustment for retrospective sampling. The resulting hypothesis test will be valid for any fixed $0 < k < 1$.

For association studies of complex traits in samples of related individuals, there is an "enrichment effect," namely, that affected individuals with affected relatives have a higher expected frequency of the alleles that increase susceptibility for a complex disease than do unaffected individuals who do not have affected relatives. We allow for this enrichment effect by including $\tilde{\mathbf{A}} = \mathbf{K}\mathbf{A}$ instead of $\mathbf{A}$ as a covariate in our generalized linear model, where $\mathbf{K}$ is the kinship matrix. We take $Y_{ij} \sim .5 \times \text{Bin}(2, p_{ij})$ as before, where we now let

$$\text{logit}(p_{ij}) = \alpha_j + \tilde{A}_i \gamma_j = \alpha_j + \gamma_j \sum_{k=1}^m 2\phi_{ik} A_k. \quad (6)$$

The $2a \times 1$ parameter vector is $\boldsymbol{\beta} = (\boldsymbol{\gamma}^T, \boldsymbol{\alpha}^T)^T$ where $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_a)^T$ is the parameter of interest, which measures trait-haplotype association, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_a)^T$ is a nuisance parameter whose entries are the logits of the haplotype frequencies unconditional on phenotype. The form of Equation (6) is justified by an asymptotic argument applied to a single-gene model with effect size tending to zero (see Thornton and McPeck 2007). Note that the definition of $\tilde{\mathbf{A}}$ can be easily extended to allow for contributions from relatives who are phenotyped but not

genotyped, by extending the summation over k in Equation (6) to include these additional relatives. The null hypothesis of no trait-haplotype association is $H_0: \gamma = \mathbf{0}$, while the alternative is $\gamma \neq \mathbf{0}$. Under H_0 , assuming no linkage, the first and second moment structure for $\mathbf{Y}$ is the same as in the previous subsection. Applying Equation (5) to the special case of complete data ($\mathbf{Z} = \mathbf{Y}$), we obtain a quasi-likelihood score statistic that is equal to the M_{QLS} test statistic of Thornton and McPeck (2007).

To form the M_{QLS} test statistic with incomplete data, we consider the same two choices of f_{ij} as in the previous subsection. If we view $Z_{ij}(\gamma, \alpha)$ as a function of (γ, α) , then the calculation of M_{QLS} requires $Z_{ij}(\mathbf{0}, \hat{\alpha}_0)$, which is the same as $Z_{ij}(\hat{\mathbf{p}})$ in the previous subsection. We require $\text{Var}(\mathbf{Z})$ only under the null hypothesis, which equals the Ω that would be obtained for the estimation problem of the previous subsection. We keep μ the same as in the complete data case (specified by inverting the logit of Equation (6)). To obtain $\mathbf{F}(\mathbf{0}, \hat{\alpha}_0) = -E(\partial(\mathbf{Z} - \mu) / \partial \beta)_{(\mathbf{0}, \hat{\alpha}_0)}$, we also need an expression for $E(\partial \mathbf{Z} / \partial \beta)_{(\mathbf{0}, \hat{\alpha}_0)}$. For any given alternative model, consider a power series expansion for $Z_{ij}(\gamma, \alpha) = E(Y_{ij} | f_{ij}(\mathbf{I}))$ around $\gamma = \mathbf{0}$. In fact, we only need the first-order term of the series to obtain the derivative at $\gamma = \mathbf{0}$. It turns out that this first-order term, which depends on $\tilde{\mathbf{A}}$, will be the same for a single-gene multiplicative model and a single-gene additive model, and that is what we use to obtain $E(\partial \mathbf{Z} / \partial \beta)_{(\mathbf{0}, \hat{\alpha}_0)}$. The M_{QLS} test is the resulting score test for the null hypothesis $\gamma = \mathbf{0}$, obtained from Equation (5). Note that to obtain an asymptotic χ_a^2 null distribution for M_{QLS} , it is sufficient to have correct specification of μ , Ω , and $\mathbf{F}$ under the null hypothesis (under conditions given in the Appendix).

When expected counts of some haplotypes are small, one can usually increase power by pooling haplotypes to reduce the degrees of freedom, a . In the analysis of 5-SNP haplotypes in section 5, we combine haplotypes by a variation on the “clustering” method of Tzeng, Devlin, Wasserman, and Roeder (2003). This involves grouping rare haplotypes with more common ones that differ at a single site until every category has estimated frequency $\geq .05$, while rare haplotypes that differ at two or more sites from every other sampled haplotype are pooled to form a new category.

3.3 Connection between M_{QLS} and Retrospective Likelihood

To clarify the relationship of IQL to likelihood-based haplotype methods (e.g. Lin and Zeng (2006), developed for unrelated individuals), we describe a close connection between M_{QLS} and the score test based on a retrospective likelihood. Assume without loss of generality that the first $\rho \leq m$ of the m ordered individuals have known phenotype, and for these individuals, let $X_i = 1$ if i is affected, and $X_i = 0$ if i is unaffected. Here we treat $\mathbf{Y}$ and $\mathbf{A}$ as m -vectors, as before, while $\mathbf{X}$ is a ρ -vector which is a function of $\mathbf{A}$. We condition the analysis on the pattern of missing phenotype information among the sampled individuals. Consider the following prospective logistic regression model for phenotype given complete haplotype data:

$$P(\mathbf{A} | \mathbf{Y}) = P(\mathbf{X} | \mathbf{Y}) = \prod_{i=1}^{\rho} \frac{e^{(\beta_0 + \pi_i) X_i}}{1 + e^{(\beta_0 + \pi_i)}}, \quad (7)$$

where $\pi_i = \sum_{j=1}^a r_j Y_{ij}$, with the parameter $\mathbf{r} = (r_1, \dots, r_a)^T$ representing association of the haplotypes with the phenotype, and with $\beta_0 = \text{logit}(k)$ representing logit of disease prevalence under the null hypothesis. Model (7) is not appropriate for complex traits in related individuals, because under the null hypothesis of no association, $H_0: \mathbf{r} = \mathbf{0}$, the phenotypes of related individuals are modeled as independent, which is unrealistic. Therefore, we generalize Equation (7) to the following larger class of models, which allows dependence of the phenotypes under the null hypothesis:

$$P(\mathbf{A}|\mathbf{Y}) = P(\mathbf{X}|\mathbf{Y}) = P_0(\mathbf{X}) c(\mathbf{Y}, \mathbf{r}) \prod_{i=1}^{\rho} e^{\pi_i X_i}, \quad (8)$$

where $c(\mathbf{Y}, \mathbf{r}) = [\sum_{\mathbf{x} \in \{0,1\}^{\rho}} P_0(\mathbf{x}) \prod_{i=1}^{\rho} \exp(\pi_i x_i)]^{-1}$ is the normalizing constant and $P_0(\mathbf{X})$ is the joint distribution of $\mathbf{X}$ when $\mathbf{r} = \mathbf{0}$. Here $P_0(\mathbf{X})$ allows for dependence among the X_i 's and need not be specified. We assume that $P_0(\mathbf{X})$ has the property that all the marginals are the same, i.e. that $P_0(X_i = 1) = \sum_{\mathbf{x} \in \{0,1\}^m} P_0(\mathbf{x}) x_i = k$ for all $1 \leq i \leq m$, where this is the same prevalence, k , that we use in the definition of the phenotype vector $\mathbf{A}$. Equation (8) represents exponential tilting of $P_0(\mathbf{X})$. Equation (7) is obtained as a special case of Equation (8) by choosing

$P_0(\mathbf{X}) = \prod_{i=1}^{\rho} \exp(\beta_0 X_i) / [1 + \exp(\beta_0)]$, $P(\mathbf{Y})$ is determined by Mendelian inheritance in pedigrees and the assumption of Hardy-Weinberg equilibrium in the general population. Combining $P(\mathbf{Y})$ with Equation (8), we can obtain the retrospective likelihood for complete haplotype data, $L_Y = P(\mathbf{Y}|\mathbf{A})$. Under assumptions on missing genotype data stated in the beginning of section 3, the retrospective likelihood for incomplete haplotype data is modeled by

$$L_I = P(\mathbf{I}|\mathbf{A}) = \sum_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{A}) P(\mathbf{I}|\mathbf{Y}) = \sum_{\mathbf{Y} \subset \mathbf{I}} P(\mathbf{Y}|\mathbf{A}) = \sum_{\mathbf{Y} \subset \mathbf{I}} L_Y, \quad (9)$$

where the second and third summations are over all complete-data vectors $\mathbf{Y}$ that are compatible with the observed data $\mathbf{I}$. As a consequence, the incomplete haplotype data retrospective likelihood score function, l_I satisfies

$$\dot{l}_I = E(\dot{l}_Y | \mathbf{I}), \quad (10)$$

where l_Y is the complete haplotype data retrospective likelihood score function. Using the above observations, it is easy to verify that when the nuisance parameter, $\mathbf{p}$, is known and with k chosen as above, the retrospective likelihood score test for every model in the class given by Equation (8), with related individuals and complete or incomplete data, is the same as the M_{IQLS} with choice $f_{ij}(\mathbf{I}) = \mathbf{I}$. However, the nuisance parameter is generally not known, and in the likelihood score test, $\mathbf{p}$ is estimated by the MLE under the null hypothesis, while in the M_{IQLS} , it is estimated by the computationally-simpler IQL estimator under the null hypothesis. The different estimators of the nuisance parameter, $\mathbf{p}$, cause the variance portions of the two test statistics to differ as well because they account for estimation of $\mathbf{p}$. For unrelated individuals, the MLE and IQL estimators of $\mathbf{p}$ are the same, and the retrospective likelihood score test is identical to the M_{IQLS} , under the weak assumption that $f_{ij}(\mathbf{I})$ contains at least the information of $\mathbf{I}_i$ for each $1 \leq i \leq m$.

3.4 Estimators and Tests to be Compared by Simulation

In section 4, we perform simulation studies to assess our IQL haplotype frequency estimators and M_{IQLS} disease-haplotype association tests, and we compare them to a few other tests and estimators, which are described in this subsection.

In our IQLS haplotype tests and estimators, calculation of the covariance matrix $\mathbf{\Omega}$ is one of the more challenging aspects. Particularly in the estimation context, it is natural to wonder how much is gained by accounting for relatedness of individuals in the calculation of $\mathbf{\Omega}$. Therefore, we also consider what we call the *equal-weight* (EW) estimating function, $\mathbf{U}_{EW}$. Let $\mathbf{Z}_i =$

$(Z_{i1}, \dots, Z_{ia})^T$ be the portion of $\mathbf{Z}$ corresponding to individual i , $\boldsymbol{\mu}_i = E(\mathbf{Z}_i)$, $\boldsymbol{\Omega}_i = \text{Cov}(\mathbf{Z}_i)$, and $\mathbf{F}_i = -E(\partial(\mathbf{Z}_i - \boldsymbol{\mu}_i)/\partial\boldsymbol{\beta})$ for $1 \leq i \leq m$. Then $\mathbf{U}_{EW} = \sum_{i=1}^m \mathbf{F}_i^T \boldsymbol{\Omega}_i^{-1} (\mathbf{Z}_i - \boldsymbol{\mu}_i)$, which can be viewed as Equation (2) with $\boldsymbol{\Omega}$ replaced by the block-diagonal matrix $\boldsymbol{\Omega}_{diag} = \text{diag}(\boldsymbol{\Omega}_1, \dots, \boldsymbol{\Omega}_m)$. In the special case of unrelated individuals, $\mathbf{U}_{EW} = \mathbf{U}_J$, and the EW estimator coincides with the IQL estimator. When relatives are sampled, if we choose $f_{ij}(\mathbf{I}) = \mathbf{I}_i$ (choice (a) of subsection 3.1), the EW estimator is the estimator obtained by applying the MLE for unrelated individuals to a sample of related individuals. We also use $\mathbf{U}_{EW}$ to obtain a score test for $\boldsymbol{\gamma} = \mathbf{0}$ by Equation (5). (Note that $\boldsymbol{\Lambda}(\mathbf{U}_{EW})$, which is needed in Equation (5), does involve the off-diagonal blocks of $\boldsymbol{\Omega}$.)

Because of its close connection to prior work (Elashoff and Ryan 2004; Browning et al. 2005; Zhang et al. 2005), the estimating function, $\mathbf{U}_{CDW}$, which we call the *complete-data-weight* (CDW) estimating function, is also considered. We define

$\mathbf{U}_{CDW} = \sum_{i=1}^m w_i \mathbf{F}_i^T \boldsymbol{\Omega}_i^{-1} (\mathbf{Z}_i - \boldsymbol{\mu}_i)$, where $w_i = (\mathbf{1}_m^T \mathbf{K}^{-1})_i$. In the special case of complete data, we have $\mathbf{U}_{CDW} = \mathbf{U}$, where $\mathbf{U}$ is the complete-data, quasi-likelihood score function given in subsection 3.1. The CDW estimator is closely related to the estimator of Zhang et al. (2005), who applied essentially the same estimating function to a somewhat different context of control haplotype-frequency estimation based on the non-transmitted haplotypes of parent/affected-child trios, under the further constraint that the haplotype frequencies follow a Markov model. Furthermore, $\mathbf{U}_{CDW}$ is equivalent to the estimating function, $\mathbf{U}_{ER}$, of Elashoff and Ryan (2004) in the special case when both of the following hold: (1) $f_{ij}(\mathbf{I}) = \mathbf{I}$ for $1 \leq i \leq m$, $1 \leq j \leq a$ and (2) $\mathbf{F}_i^T \boldsymbol{\Omega}_i^{-1}$ is constant in $1 \leq i \leq m$ and of full rank.

For complete-data, case-control association testing with related individuals, Bourgain et al. (2003) formed a quasi-likelihood score test statistic (W_{QLS}) using Equation (5), but instead of modeling the mean by Equation (6), they used the simpler model

$$\text{logit}(p_{ij}) = \alpha_j + X_i \gamma_j, \quad (11)$$

where $X_i = 1$ if i is affected and 0 if i is unaffected, and individuals with missing phenotype are dropped from the analysis. The model of Equation (11) ignores the enrichment effect described in subsection 3.2. For any of the estimating functions, $\mathbf{U}_J$, $\mathbf{U}_{EW}$, or $\mathbf{U}_{CDW}$, when we form a score test for $\boldsymbol{\gamma} = \mathbf{0}$ by Equation (5), we can use either Equation (6) or Equation (11) for the mean model. The score test constructed from $\mathbf{U}_{CDW}$ with the mean model of Equation (11) is the score-test version of the composite-likelihood ratio test of Browning et al. (2005) for case-control haplotype association in related individuals.

4 SIMULATION STUDIES

To assess MSE of the IQL haplotype frequency estimators and type I error and power of the IQLS haplotype association tests, we consider three study designs, which are similar to those of Thornton and McPeck (2007). In the first, 60 outbred, three-generation, 16-person pedigrees are sampled, where 20 of the pedigrees have 4 affected individuals, 20 have 5, and 20 have 6. The pattern of affected and unaffected individuals varies randomly according to the trait models described in the next paragraph. In each sampled pedigree, phenotypes for all 16 individuals are observed. An individual's genotypes are observed if and only if at least 30% of the individual's siblings, parents, and offspring are affected. The second study design is similar to the first with two differences: (1) an additional 200 unrelated, unaffected controls are included in the study; and (2) for each individual in a sampled pedigree, the individual's genotypes are observed if and only if at least half of the individual's siblings, parents, and offspring are affected. The third study design consists of 120 pedigrees of the above structure, where exactly

40 have 1 affected individual, 40 have 2, and 40 have 3. Phenotypes are observed for all individuals, and genotypes are observed for all affected individuals plus two unaffected individuals randomly chosen in each pedigree. To assess the impact of parental genotype information, we further simulate parental genotypes for all genotyped individuals and compare the results with and without this additional parental genotype information.

We simulate under multilocus trait models I and III of Thornton and McPeck (2007), where we replace their SNP 2 with a trio, T_2 , of tightly-linked SNPs for which there are 6 possible haplotypes. We replace alleles 1 and 0 of their SNP 2 by the sets of haplotypes {101, 110} and {000, 001, 010, 100}, respectively, for T_2 . Table 1 gives haplotype frequencies for T_2 with resulting population prevalence, k , and sibling risk ratio, $\lambda_s = k_s/k$, where k_s is the prevalence conditioned on having an affected sibling.

4.1 MSE Comparison for Haplotype Frequency Estimation

In Table 2, we compare 6 incomplete-data, haplotype-frequency estimators from section 3, where the subscript a or b denotes the choice for $f_{ij}(\mathbf{I})$. We generate 5,000 data sets under design 1 with model Ia. Haplotype frequencies are estimated at a trio of tightly-linked SNPs that are unlinked and unassociated with any causal variants. For each setting, we use the method of Nicolae (2006) to quantify the extent to which genotype data are informative about the given haplotype. Specifically, we measure the proportion of an individual's haplotype- j information that is provided by his genotype as $m_{\mathbf{G}} = \text{Var}(Z_{ij})/[p_j(1 - p_j)/2]$, where Z_{ij} is calculated with $f_{ij}(\mathbf{I}) = \mathbf{G}_i$. The proportion of haplotype- j information on an individual obtained when the genotype data of both the individual and his parents are available is $m_{\mathbf{G}_{trio}}$, which equals the above ratio calculated with $f_{ij}(\mathbf{I}) = \mathbf{G}_{trio} = (\mathbf{G}_i, \mathbf{G}_{m(i)}, \mathbf{G}_{f(i)})$.

IQL_b , which incorporates parental genotypes and uses optimal weights, has the smallest MSE in all cases. CDW_b performs almost as well, because $\mathbf{G}_{trio}$ provides close to complete haplotype information ($m_{\mathbf{G}_{trio}}$ close to 1). The use of optimal weights provides a more marked improvement over the use of complete-data weights when the amount of information is lower (IQL_a vs. CDW_a for which the information is $m_{\mathbf{G}}$). The EW estimators perform worse than the CDW and IQL estimators under both choices (a) and (b).

4.2 Power and Type I error of Haplotype Association Tests

We compare score tests based on $\mathbf{U}_I$, $\mathbf{U}_{EW}$, and $\mathbf{U}_{CDW}$, where we denote the resulting tests by $IQLS$, EWS , and $CDWS$, respectively. We consider each of these with options (a) and (b) for f_{ij} , where these are denoted by subscripts a and b , respectively, and we consider two alternative mean models, given by Equation (6) and Equation (11), which we denote by M and W , respectively. Thus, e.g., W_{IQLS_b} denotes the score test based on $\mathbf{U}_I$ with choice (b) for f_{ij} and the mean model of Equation (11).

To further investigate the effect of the choice of f_{ij} on the power of the $IQLS$ tests, in addition to choices (a) and (b) described above, we would ideally like to compare the choice $f_{ij}(\mathbf{I}) = \mathbf{I}$. Because this choice presents computational difficulties, we instead consider choice (c), which gives an upper bound on the power when $f_{ij}(\mathbf{I}) = \mathbf{I}$. For choice (c), if individual i 's typed relatives are a subset of $\{i, f(i), m(i)\}$, then we take $f_{ij}(\mathbf{I}) = \mathbf{I}$, and if individual i has typed relatives other than parents, we use complete haplotype data for individual i .

To assess type I error, association is tested with haplotypes at a trio of tightly-linked SNPs that are unlinked and unassociated with any causal variant, where individuals are sampled according to design 1 with model Ia. Based on 5,000 replicates, we find that for all 14 tests, type I error is not significantly different from the nominal when $\alpha = .05$ or $.01$ (results not shown).

In Table 3, we compare power to detect trait-haplotype association at the trio T_2 of tightly-linked SNPs that are associated with the trait, under designs 2 and 3. In every setting, the M_{IQLS} tests perform best, with M_{IQLS_c} and M_{IQLS_b} having, respectively, the highest and second-highest power among all tests for all simulation settings, and with M_{IQLS_a} having highest power among the tests in which only the individual's genotype data are conditioned on (choice (a)). This suggests that with related individuals, all three of the following are important for power: optimal weighting of individuals, accounting for the enrichment effect, and incorporating parental genotype information. The comparison of choices (a), (b), and (c) for f_{ij} in the IQLS tests indicates that little power is lost when $f_{ij}(\mathbf{I}) = (\mathbf{I}_i, \mathbf{I}_{m(i)}, \mathbf{I}_{f(i)})$ is used compared to the more computationally intensive or infeasible approach of using $f_{ij}(\mathbf{I}) = \mathbf{I}$. The power difference between these is no more than .05 across all simulation settings. This is not surprising, because adding parental genotype information allows one to capture most of the haplotype information for an individual, as reflected, for example, in the high values of the measure of information, $m_{\mathbf{G}_{\text{trio}}}$ in Table 2.

Interestingly, although the *CDW* approach performs substantially better than the *EW* approach for estimation, it performs noticeably worse for association testing, probably because the *EW* approach tends to have a built-in enrichment effect in these study designs, i.e. it effectively gives extra weight to affected individuals with affected relatives in the sample. For the *IQLS* and *CDWS* tests, use of the mean model of Equation (6) always provided at least as much power, and usually more, than did use of the mean model of Equation (11), illustrating the enrichment effect. For the *EWS* test, because there is already a built-in enrichment effect, one can often have worse power with Equation (6) than with Equation (11).

5 DATA ANALYSIS

We apply our IQLS score tests to study association of haplotypes with alcoholism in the GAW 14 COGA data set described in detail in Edenberg et al. (2005). Of the 1,614 individuals from 143 families, we include only those coded as “white, non-Hispanic.” After exclusion of two individuals with apparent relationship errors and five who appear to have mixed or misclassified ethnicity, 1,067 individuals remain. We classify as affected those who are “affected with ALDX1” or who have “symptoms of ALDX1.” We classify as unaffected those coded as “pure unaffected,” and we classify as missing phenotype those coded as “never drank alcohol.” By these criteria, there are 855 affected, 199 unaffected and 13 missing phenotype. Among these 1,067 individuals, there are 1,023 individuals with available SNP data. We analyze the 10,407 autosomal SNPs in the data set having minor allele frequency $> .01$.

We test for association with haplotypes based every adjacent pair of SNPs, using the M_{IQLS} with $k = .05$. Table 4 lists three pairs of SNPs for which the M_{IQLS} test is significant at the 5% level after Bonferroni correction. In each of these examples, the detected association is driven by a rare two-SNP haplotype that occurs only in individuals with large values of $\tilde{A}$. Association is not detected for any of the SNPs marginally, because the individual SNPs do not have rare alleles, and their alleles are not strongly associated with $\tilde{A}$. This could occur, for example, if a predisposing allele at an untyped variant were in linkage disequilibrium with the haplotype.

We follow up the 2-SNP haplotype analysis with a 5-SNP haplotype analysis, based on a sliding window, across all SNPs on each of the three chromosomes on which we detected association. For each 5-SNP window, we reduce the degrees of freedom of the test by clustering haplotypes as described in subsection 3.2. In addition, we compute p -values using a wide range of values for the estimated population prevalence, k . Table 5 gives the three windows for which the resulting M_{IQLS} association test has p -value $< 10^{-6}$. The effect of k on the p -value is slight, suggesting that the results are quite robust to choice of k . If we instead use a p -value threshold of 10^{-5} , we gain two additional significant windows, one that overlaps by 4 SNPs with the

third window in Table 5 and one that overlaps by 4 SNPs with the second window in Table 5, so that the same 3 regions are still identified. Two of the regions identified by 2-SNP haplotype analysis in Table 4 are identical to regions identified by 5-SNP haplotype analysis in Table 5, with comparable p -values between the two analyses. A region on chromosome 10 was detected by 2-SNP but not 5-SNP haplotype analysis. This is a result of low linkage disequilibrium among typed SNPs in the region: r^2 between SNPs tsc0041916 and tsc0048856 is .98, while r^2 of either of these two with other typed SNPs is less than .16. Thus, the 3 additional SNPs, in a 5-SNP haplotype containing tsc0041916 and tsc0048856, mainly introduce noise to the analysis.

The first set of five SNPs in Table 5 is in the intron of a candidate tumor suppressor gene that encodes protein tyrosine phosphatase receptor type G (PTPRG [Affymetrix; Entrez Gene]), located at 3p21-p14. The second set of five SNPs in Table 5 is 251 kb from the gene encoding vestigial like 3 (VGLL3 [Affymetrix; Entrez Gene]), located at 3p12.1. The set of five SNPs on chromosome 2 is in the intron of the gene encoding solute carrier family 8 member 1 (SLC8A1 [Affymetrix; Entrez Gene]), located at 2p23-p22. The two SNPs on chromosome 10 are 178 kb from the gene encoding sortilin-related VPS10 domain containing receptor 1 (SORCS1 [Affymetrix; Entrez Gene]), located at 10q23-q25. To our knowledge, none of these genes are obvious candidates for alcoholism.

6 DISCUSSION

Genetic association analysis is widely used to identify genetic variants that underlie human complex disease. Haplotypes are more informative for disease mapping than are single markers, because haplotypes can provide additional information on untyped variants as well as on interactions among tightly-linked, typed variants. However, we generally have only partial haplotype information on each individual. Dependence among observations in an association study, due to relatedness of sampled individuals, commonly arises when families sampled for a linkage study are included in an association study. We introduce the IQLS method for assessment of disease association with haplotypes in case-control studies in which some individuals may be related. The method is applicable to completely general samples of related and unrelated individuals, including large complex pedigrees as well as samples that combine small families with unrelated individuals. The IQLS method effectively addresses the problems of dependent and partially informative observations in a way that is computationally practical and that does not require specification of a full joint distribution.

In simulation studies, the M_{IQLS} tests perform well for a range of complex disease models. We consider the impact of each of the following features of our M_{IQLS} haplo-type association test: (1) incorporation of the “enrichment effect” that is characteristic of complex genetic models; (2) use of optimal weights (as opposed to complete-data or equal weights); and (3) use of parental genotype data to obtain more information about an individual’s haplotype. All three of these features seem to have a noticeable impact on power. In an analysis of the GAW 14 COGA data set, we detect significant association of haplotypes with alcoholism in a case-control sample consisting of families, where these associations are not apparent from single-SNP analysis.

The IQLS method for parameter estimation and hypothesis testing is presented in a somewhat more general setting in which the complete data are dependent with marginal distributions following a generalized linear model. When the first and second moment structure can be correctly specified, the quasi-likelihood score method provides a way to draw inference without specifying the full joint distribution. This is advantageous when the MLE is very difficult to calculate or when there is insufficient information to construct a likelihood function. With incomplete data, we form an IQLS function, which is a quasi-likelihood score function based

on a vector $\mathbf{Z}$ whose elements are conditional expectations of the elements of the complete-data vector, given selected functions of the incomplete data. We use an iterative algorithm to obtain the IQL parameter estimates. The computational complexity of the method will depend on the choice of conditioning events. There is wide latitude in this choice, provided that the covariance matrix and expected derivative of the conditional expectations are correctly specified. (For the testing problem, these need only be evaluated under the null hypothesis.) An estimated variance-covariance matrix for the IQL estimator, appropriately adjusted for incomplete data, is obtained as part of the iterative algorithm. Under regularity conditions, the IQL estimator is consistent and asymptotically normal with minimum asymptotic variance among estimators that are solutions to linear unbiased estimating equations based on $\mathbf{Z}$.

Acknowledgments

This study was supported by National Institutes of Health grants R01HG001645 and U01HL084715. Data were provided by the Collaborative Study on the Genetics of Alcoholism (U10AA008401) through the Genetics Analysis Workshop (R01GM031575). The authors thank the editor, an associate editor and two anonymous referees for helpful comments.

APPENDIX

For haplotype analysis, we give sufficient conditions for consistency and asymptotic normality of the IQL estimators and asymptotic χ^2 null distribution of the IQLS score tests. Suppose that the sampled individuals form b independent families, where we let $b \rightarrow \infty$. (Note that a single individual unrelated to anyone else in the sample is considered a special case of a family.) Define a family configuration to be a particular constellation of typed relatives with a particular pattern of missing genotype information at the given tightly-linked SNPs. Assume that there are only a finite number of possible configurations allowed, indexed by $c = 1, \dots, C$. Each configuration has an associated information Λ_c for a family having that configuration. Assume that there exist $p_c \geq 0$, for $1 \leq c \leq C$ with $\sum_c p_c = 1$ such that

$P\left\{b^{-1} \sum_{i=1}^b 1\{\text{family } i \text{ of type } c\} \rightarrow p_c\right\} = 1$. We require the existence of at least one c such that $p_c > 0$ and Λ_c is positive definite. Under these assumptions, it is straightforward to verify conditions 4 and 5 of Yuan and Jennrich (1998) and the Theorem of Yuan (1997) which implies condition 2 of Yuan and Jennrich (1998).

REFERENCES

- Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeck MS. Novel Case-Control Test in a Founder Population Identifies P-Selectin as an Atopy-Susceptibility Locus. *American Journal of Human Genetics* 2003;73:612–626. [PubMed: 12929084]
- Browning SR, Briley JD, Briley LP, Chandra G, Charnecki JH, Ehm MG, Johansson KA, Jones BJ, Karter AJ, Yarnall DP, Wagner MJ. Case-Control Single-Marker and Haplotypic Association Analysis of Pedigree Data. *Genetic Epidemiology* 2005;28:110–122. [PubMed: 15578751]
- Edenberg HJ, Bierut LJ, Boyce P, Cao M, Cawley S, Chiles R, Doheny KF, Hansen M, Hinrichs T, Jones K, Kelleher M, Kennedy GC, Liu G, Marcus G, McBride C, et al. Description of the Data from the Collaborative Study on the Genetics of Alcoholism (COGA) and Single-Nucleotide Polymorphism Genotyping for Genetic Analysis Workshop 14. *BMC Genetics* 2005;6:S2. [PubMed: 16451628]
- Elashoff M, Ryan L. An EM Algorithm for Estimating Equations. *Journal of Computational and Graphical Statistics* 2004;13:48–65.
- Godambe VP. An Optimum Property of Regular Maximum Likelihood Estimation. *The Annals of Mathematical Statistics* 1960;31:1208–1211.
- Heyde, CC. Quasi-likelihood and its Application: a General Approach to Optimal parameter Estimation. New York: Springer; 1997.

- Heyde CC, Morton R. Quasi-Likelihood and Generalizing the EM Algorithm. *Journal of the Royal Statistical Society, Ser. B* 1996;58:317–327.
- Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM. Family-Based Tests for Associating Haplotypes with General Phenotype Data: Application to Asthma Genetics. *Genetic Epidemiology* 2004;26:61–69. [PubMed: 14691957]
- Lehmann, EL.; Casella, G. *Theory of Point Estimation*. 2nd edition. New York: Springer; 1998.
- Liang KY, Zeger SL. *Longitudinal Data Analysis Using Generalized Linear Models*. *Biometrika* 1986;73:13–22.
- Lin DY, Zeng D. Likelihood-Based Inference on Haplotype Effects in Genetic Association Studies (with discussion). *Journal of the American Statistical Association* 2006;101:89–118.
- McCullagh, P.; Nelder, JA. *Generalized Linear Models*. 2nd edition. New York: Chapman and Hall; 1989.
- McPeck MS, Wu X, Ober C. Best Linear Unbiased Allele-Frequency Estimation in Complex Pedigrees. *Biometrics* 2004;60:359–367. [PubMed: 15180661]
- Nicolae DL. Quantifying the Amount of Missing Information in Genetic Association Studies. *Genetic Epidemiology* 2006;30:703–717. [PubMed: 16986163]
- Risch N, Teng J. The Relative Power of Family-Based and Case-Control Designs for Linkage Disequilibrium Studies of Complex Human Diseases I: DNA Pooling. *Genome Research* 1998;8:1273–1288. [PubMed: 9872982]
- Schaid DJ. Evaluating Associations of Haplotypes with Traits. *Genetic Epidemiology* 2004;27:348–364. [PubMed: 15543638]
- Slager SL, Schaid DJ. Evaluation of Candidate Genes in Case-Control Studies: a Statistical Method to Account for Related Subjects. *American Journal of Human Genetics* 2001;68:1457–1462. [PubMed: 11353403]
- Thornton T, McPeck MS. Case-Control Association Testing with Related Individuals: A More Powerful Quasi-Likelihood Score Test. *American Journal of Human Genetics* 2007;81:321–337. [PubMed: 17668381]
- Tzeng J-Y, Devlin B, Wasserman L, Roeder K. On the Identification of Disease Mutations by the Analysis of Haplotype Similarity and Goodness of Fit. *American Journal of Human Genetics* 2003;72:891–902. [PubMed: 12610778]
- Wedderburn RWM. Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika* 1974;61:439–447.
- Xie M, Yang Y. Asymptotics for Generalized Estimating Equations with Large Cluster Sizes. *The Annals of Statistics* 2003;31:310–347.
- Yuan K-H. A Theorem of Uniform Convergence of Stochastic Functions with Applications. *Journal of Multivariate Analysis* 1997;62:100–109.
- Yuan K-H, Jennrich RI. Asymptotics of Estimating Equations under Natural Conditions. *Journal of Multivariate Analysis* 1998;65:245–260.
- Zeger SL, Liang KY. Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics* 1986;42:121–130. [PubMed: 3719049]
- Zhang J, Schneider D, Ober C, McPeck MS. Multilocus linkage disequilibrium mapping by the decay of haplotype sharing with samples of related individuals. *Genetic Epidemiology* 2005;29:128–140. [PubMed: 16049966]

Table 1

Parameter settings for simulation models

Model	Haplotype frequencies for T_2										k	λ_{ns}
	p_{000}	p_{001}	p_{010}	p_{100}	p_{101}	p_{110}	p_{111}	p_{100}	p_{101}	p_{110}		
Ia	.15	.2	.05	.1	.3	.2	.071	.1	.3	.2	.071	1.224
Ib	.07	.13	.17	.23	.15	.25	.123	.15	.23	.25	.123	1.077
Ic	.15	.3	.1	.2	.15	.1	.256	.15	.1	.1	.078	2.001
IIIa	.1	.2	.35	.15	.1	.1	.047	.15	.1	.1	.047	4.209
IIIb	.07	.13	.17	.23	.15	.25	.123	.15	.23	.25	.123	1.077

Table 2

Ratios of MSE for estimation of haplotype frequency p , based on 5,000 simulated replicates, where, for each estimator, the ratio of its MSE to the MSE of IQ_{L_b} is reported

p	Information		MSE Ratio				
	m_G	$m_{G_{\text{info}}}$	EW_a	EW_b	CDW_a	CDW_b	IQ_{L_a}
0.40	0.90	0.96	1.27	1.19	1.14	1.006	1.09
0.30	0.88	0.96	1.30	1.20	1.19	1.004	1.11
0.20	0.85	0.95	1.34	1.21	1.23	1.007	1.14
0.15	0.86	0.96	1.32	1.22	1.19	1.003	1.12
0.10	0.80	0.94	1.37	1.21	1.28	1.001	1.19

Note: The full haplotype frequency distribution is (0.4, 0.3, 0.2, 0.1) for row 1 and (0.3, 0.2, 0.2, 0.15, 0.1, 0.05) for the remaining rows. The average of the s.e.'s for the MSE ratios are .018, .015, .014, .002 and .010 for, respectively, EW_a , EW_b , CDW_a , CDW_b and IQ_{L_a} .

Table 3

Empirical power results, based on 5,000 simulated replicates, $\alpha = .05$

Statistic	Estimated Power							
	Design 2 Models				Design 3 Models			
	Ia	Ib	Ic	IIIa	IIIb	Ib	Ic	IIIb
W_{CDWS_a}	0.24	0.52	0.38	0.32	0.40	0.55	0.40	0.41
W_{CDWS_b}	0.27	0.59	0.45	0.36	0.46	0.60	0.47	0.45
M_{CDWS_a}	0.42	0.72	0.50	0.40	0.55	0.57	0.50	0.41
M_{CDWS_b}	0.46	0.77	0.57	0.44	0.60	0.62	0.57	0.46
W_{EWS_a}	0.40	0.69	0.51	0.43	0.68	0.64	0.46	0.55
W_{EWS_b}	0.46	0.74	0.59	0.48	0.74	0.68	0.54	0.58
M_{EWS_a}	0.46	0.72	0.48	0.39	0.61	0.57	0.49	0.45
M_{EWS_b}	0.51	0.77	0.56	0.44	0.67	0.62	0.56	0.50
W_{QLS_a}	0.21	0.40	0.38	0.32	0.46	0.52	0.36	0.46
W_{QLS_b}	0.23	0.46	0.44	0.35	0.53	0.55	0.43	0.49
W_{QLS_c}	0.25	0.49	0.47	0.37	0.56	0.60	0.46	0.54
M_{QLS_a}	0.52	0.81	0.61	0.52	0.73	0.68	0.57	0.55
M_{QLS_b}	0.57	0.86	0.68	0.57	0.78	0.73	0.65	0.60
M_{QLS_c}	0.58	0.89	0.71	0.59	0.81	0.76	0.69	0.65

Note: The standard error is .007 for power in the range (0.31, 0.69), .006 for (0.19, 0.30) or (0.70, 0.81), .005 for (0.82, 0.88), and .004 for (0.89, 0.93).

Table 4

Results for the COGA data based on 2-SNP haplotypes

Chr	Marker 1	Marker 2	Position (cM)	Haplotype frequencies	Marginal <i>p</i> -values	<i>M_{QLS}</i> <i>p</i> -value
10	tsc0041916	tsc0048856	125.3	(.56, .43, .003, .001)	9.8e-1	6.2e-1
3	tsc0521126	tsc0521130	109.15	(.65, .33, .018, .003)	3.8e-1	8.0e-1
2	tsc0049214	tsc0049215	64.885	(.46, .43, .106, .002)	6.2e-1	2.1e-2
						4.6e-7

Table 5

Results for the COGA data based on 5-SNP haplotypes

Chr	Marker	Position (cM)	Allele frequency	Marginal <i>p</i> -value	<i>MQTL5 p</i> -value for values of <i>k</i>			
					0.01	0.05	0.1	0.2
3	tsc0417991	84.0252	.59	5.6e-1	5.2e-8	5.7e-8	6.4e-8	9.6e-8
	tsc0274859	84.0648	.86	5.2e-1				
	tsc0883753	84.688	.73	5.9e-2				
	tsc0060693	84.6886	.63	3.5e-1				
	tsc0293065	85.3441	.81	3.1e-1				
3	tsc0942736	108.684	.62	2.2e-1	5.9e-7	6.2e-7	6.7e-7	9.0e-7
	tsc1391538	108.708	.64	8.8e-1				
	tsc0048444	109.15	.70	9.0e-1				
	tsc0149354	109.15	.61	3.3e-1				
	tsc0521126	109.15	.66	3.8e-1				
2	tsc0049213	64.8849	.90	8.1e-1	1.6e-7	2.1e-7	2.9e-7	6.8e-7
	tsc0049214	64.8849	.89	6.2e-1				
	tsc0049215	64.885	.56	2.1e-2				
	tsc0301404	64.8882	.90	6.8e-1				
	tsc0301406	64.8883	.81	1.6e-1				