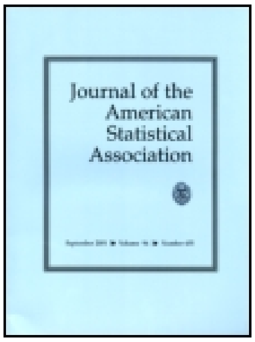

Determining Inheritance Distributions via Stochastic Penetrances

Ola Hössjer

Determining Inheritance Distributions via Stochastic Penetrances

Ola HÖSSJER

The aim of linkage analysis is to map the position of a gene contributing to an inheritable disease. The statistical model contains the disease allele frequency and penetrance parameters. Here I investigate the inheritance distribution, that is, the conditional distribution of the inheritance vector given phenotypes at the disease locus. Based on this, the likelihood and likelihood score function of Whittemore can be defined. As a result, a general semiparametric methodology of choosing score functions in linkage analysis is proposed. The proposed approach is valid for arbitrary pedigrees, and I treat quantitative, dichotomous (binary), and other phenotypes in a unified framework. The resulting score functions can be easily incorporated into existing software for multipoint linkage analysis. I use the fact that the inheritance distribution depend on unknown founder alleles. These are treated as hidden data and give rise to "stochastic penetrance factors." Certain uncorrelated unit variance random variables that are functions of the founder alleles are introduced. I show that the moment-generating function and moments of these play crucial roles in choosing likelihoods and likelihood score functions. Lower-/higher-order moments are more important when the genetic effect is weak/strong, and this corresponds to simultaneous identical by descent (IBD) sharing of few/many individuals. For inbred pedigrees and nonadditive models, the likelihood score function is dominated by individuals homozygous by descent at the disease locus. For outbred pedigrees, the local score function involves pairwise IBD sharing. Relations to existing score functions of nonparametric linkage (S_{pairs}, S_{all}, S_{robdom}) and quantitative trait loci (QTL) are highlighted.

KEY WORDS: Founder alleles; Identical by descent sharing; Inheritance distribution; Linkage analysis; Local score functions; Semiparametric linkage analysis; Stochastic penetrances.

1. INTRODUCTION

The purpose of statistical linkage analysis is to locate the position along the chromosomes of an unknown gene contributing to a certain disease. Disease-related quantities (phenotypes) and parts of DNA (so-called "genetic markers") are collected for a number of families (pedigrees). The phenotype can be binary (affection status) or quantitative (e.g., insulin concentration, body mass index) and involve various covariates (e.g., sex, age). Segregation of DNA is governed by Mendelian laws and the occurence of so-called "crossovers," latter are switching points between grandpaternal and grandmaternal DNA transmission during meiosis, that is, formation of egg or sperm cells.

DNA and phenotype segregation are associated with one another at the disease gene. The stength of this association depends on the strength of the genetic component of the disease. For loci on the same chromosome as the disease gene, the association gradually decreases with the genetic distance from the disease gene because of crossovers. The linkage score function is a stochastic process that at each DNA location measures the degree of association between phenotype segregation and DNA segregation at that position. Regions with high linkage score are declared interesting in further fine mapping of the disease gene.

A genetic model specifies both the frequency of the possible expressions (alleles) of the disease gene and the statistical relationship (penetrance) between disease alleles and phenotype. Given a certain genetic model, one can specify the conditional distribution of inheritance of disease genes given phenotypes. This distribution is important for several reasons. It summarizes the strength of association between phenotypes and disease genes, and thus reflects the strength of the genetic component of the disease. It is needed for performing parametric linkage analysis; it can be used for calculating the power of parametric and nonparametric linkage tests (Feingold, Brown, and Siegmund 1993; Lander and Kruglyak 1995), for determining the asymptotic accuracy of estimators of the disease locus (Liang, Chin, and Beaty 2001; Hössjer 2003a,b); and it generates sampling criteria for pedigrees (Risch and Zhang 1995, 1996; Liang, Huang, and Beaty 2000; Dudoit and Speed 2000). Another advantage of the conditional inheritance distribution is it's independence of the sample mechanism (ascertainment procedure). This holds under the mild assumption that pedigrees are sampled on the basis of disease phenotypes only, regardless of marker data. In contrast, the popular variance components techniques used in linkage analysis (Almasy and Blangero 1998) depend on the sample mechanism.

For each gene, an individual receives one allele from the mother and one from the father. Mendel's law of segregation states that the probability that each of these two alleles are of grandpaternal or grandmaternal origin is .5. For a whole pedigree, allele segregation can be specified at a certain locus through the inhertiance vector $\mathbf{v}$ (see, e.g., Kruglyak, Daly, Reeve-Daly, and Lander 1996), a binary vector whose length equals the number of meioses in the pedigree. I use the fact that the conditional distribution of $\mathbf{v}$ at the disease locus given phenotypes can be written as an expectation, summing over all possible founder alleles, that is, the alleles of individuals with no ancestors in the pedigree. In this way each penetrance factor becomes stochastic, viewed as a function of unknown founder alleles. This function can decomposed into additive and dominance components, analogously to variance components or U statistic techniques. It turns out that the conditional distribution of the inheritance vector can be written as the moment-generating function of a certain array of uncorrelated random variables (depending only on the disease allele frequency) evaluated at a point B, which depends on the occurrence and simultaneous identical by descent (IBD) allele sharing of various founder alleles.

Ola Hössjer is Professor, Department of Mathematics, Stockholm University, S-10691 Stockholm, Sweden (E-mail: *ola@math.su.se*). This research was sponsored by the Swedish Research Council under contract 629-2002-6286. The author thanks Roland Perfekt for suggesting several improvements on an earlier version of this manuscript, particularly the factorization $\xi_{jk} = \xi_j \xi_k$ in (12). An associate editor and two referees further gave several helpful suggestions to improve the manuscript's presentation and readability.

Journal of the American Statistical Association
December 2003, Vol. 98, No. 464, Theory and Methods
DOI 10.1198/016214503000000990

By letting the penetrance factors depend on a parameter reflecting the strength of the genetic component at the disease locus, the local likelihood-based score function of Whittemore (1996), can be evaluated in great generality. This score function depends on the behavior of the inheritance distribution for weak genetic components and involves pairwise IBD sharing of individuals for outbred pedigrees (i.e., pedigrees without loops). For inbred pedigrees (i.e., pedigrees with loops), the score function involves the dominance components of those individuals who are homozygous by descent (HBD) (i.e., both alleles originate from the same founder) at the disease locus. When the genetic component is strong, the inheritance distribution depends to a larger extent on higher-order factors of B, corresponding to simultaneous allele sharing of many individuals.

I also derive general expressions for the inheritance distribution when the disease allele frequency tends to 0, meaning that at most one disease allele is present in the pedigree.

My approach is valid for any kind of penetrance factors, particularly binary and quantitative phenotypes, with or without covariates. In the context of dichotomous phenotypes and affected pedigree members, McPeek (1999) has derived general expressions for the inheritance distribution and in particular studied which score functions are optimal for various genetic models. Several of her results can be derived as special cases in my setting with arbitrary phenotypes.

The practical implications of my approach is a general semiparametric way of choosing score functions and weighting of pedigrees, valid for arbitrary kinds of phenotypes. These can be easily incorporated into existing nonparametric linkage software, such as the Genehunter program (Kruglyak et al. 1996).

In Section 2 I define the genetic model, and in Section 3 I introduce stochastic penetrances and bivariate functions of founder alleles. I use these in Sections 4 and 5 to expand the logarithm of the inheritance distribution or the inheritance distribution itself. In the latter case, I apply the results to binary phenotypes and nonparametric linkage. I consider local genetic models and apply them to Gaussian phenotypes in Section 6. In Section 7, I define the limiting score function when the disease allele frequency tends to 0. In Section 8 I discuss how to put parametric and nonparametric linkage analysis into a general framework and how to choose score functions of practical use in linkage analysis. This is accompanied by a simulation study in Section 9 and a discussion of possible extensions in Section 10. The more technical parts of the article are collected in a series of appendixes.

2. LINKAGE SCORES AND THE GENETIC MODEL

Consider a chromosome with genetic map length of l Morgans. This means that on average, l crossovers occur during each meiosis. Further, the position of each locus $t \in [0, l]$ on the chromosome refers to genetic map distance, meaning that on average, t and $l - t$ crossovers occur to the left and right of that locus. Let τ be the unknown disease locus. Then, for each $t \in [0, l]$ we with to test $H_0 : \tau = \infty$ against $H_1(t) : \tau = t$, where $\tau = \infty$ means that the disease gene is located on another chromosome. In the former case, t is unlinked to τ, and in the latter case the two loci are perfectly linked. Given a number of families (pedigrees) $\mathcal{P}_1, \dots, \mathcal{P}_N$ with unusually large occurrence of the disease, linkage analysis proceeds by comparing the inheritance pattern of DNA at t (deduced from marker alleles) with inheritance of phenotypes. Following Kruglyak et al. (1996), a possible test statistic for choosing between H_0 and $H_1(t)$ is

$$Z(t) = \sum_{i=1}^{N} \gamma_i Z_i(t) \Bigg/ \sqrt{\sum_{i=1}^{N} \gamma_i^2}, \tag{1}$$

where $\{Z_i(t)\}_{i=1}^N$ are the *family scores* at locus t and $\{\gamma_i\}_{i=1}^N$ are weights assigned to the various pedigrees. The marker information at t is perfect if the inheritance pattern can be unambigously determined from the marker genes surrounding t. If the family scores are normalized to have 0 expectation and unit variance under H_0 under perfect marker information, it follows that $Z(t)$ also has 0 mean and unit variance under H_0. The whole process $\{Z(t); 0 \le t \le l\}$ can be used for testing H_0 against the composite alternative hypothesis $H_1 : \tau \in [0, l]$, by comparing the test statistic

$$Z_{\max} = \sup_{0 \le t \le l} Z(t) \tag{2}$$

with a given threshold. H_0 is rejected when $Z_{\max}$ exceeds the threshold. (See, e.g., Feingold et al. 1993; Lander and Kruglyak 1995; and Ängquist and Hössjer 2003 for guidance in choosing thresholds to control the significance level.) If H_0 is rejected, then a confidence region for τ can be computed. This region consists of those loci for which $Z_{\max} - Z(t)$ is smaller than a given constant, which is chosen to control the coverage probability (see Hössjer 2003a, b for details). Unless otherwise stated, from here on I consider a fixed pedigree $\mathcal{P}$ with n individuals. Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ir})$ and $\mathbf{x}_i = (x_{i1}, \dots, x_{is})$ be the set of phenotypes and covariates of the ith individual, $i = 1, \dots, n$. If f is the number of founders and each nonfounder has both of its parents in the pedigree, then the alleles at a certain locus for all pedigree members originate from the $2f$ founder alleles. For each nonfounder, there are two meioses that determine which parental alleles are transmitted. Hence the total number of meioses needed to describe transmission of founder alleles is $m = 2(n - f)$. The inheritance vector $\mathbf{v}(t) = (v_1(t), \dots, v_m(t))$ at locus t is is a binary vector of length m, where $v_k(t)$ is zero or one depending on whether a grandpaternal or grandmaternal allele is transmitted at locus t during the kth meiosis.

Notice that $\mathbf{v}(t)$ is an element of $\mathbb{Z}_2^m$, the additive vector space over the field of two elements. Let $S : \mathbb{Z}_2^m \to \mathbb{R}$ be a score function, where large values of $S(\mathbf{v}(t)) = S(\mathbf{v}(t); \mathcal{P}, \mathbf{Y})$ indicate a high degree of similarity between the inheritance vector $\mathbf{v}(t)$ and the observed phenotypes. The family score of $\mathcal{P}$ at locus t was defined by Kruglyak et al. (1996) as

$$Z(t) = \sum_{\mathbf{w} \in \mathfrak{Z}_2^m} S(\mathbf{w}) P\big(\mathbf{v}(t) = \mathbf{w} | \text{marker data for } \mathcal{P}\big), \tag{3}$$

if S has been standardized to have 0 mean and unit variance under H_0. The probability distribution of the inheritance vector given the marker data can be calculated by means of the forward–backward algorithm for hidden Markov models (cf. Lander and Green 1987). Under perfect marker information, (3) reduces to $Z(t) = S(\mathbf{v}(t))$, the score function evaluated at the inheritance vector at locus t. Even under perfect marker information, $\mathbf{v}(t)$ is not completely known, due to uncertainty of founder phases. But because all score functions of practical interest are invariant with repect to this uncertainty, $S(\mathbf{v}(t))$ is known even though $\mathbf{v}(t)$ is not (Kruglyak et al. 1996, app. B). An alternative method of calculating $Z(t)$ was proposed by Kong and Cox (1997), using pointwise likelihood ratio tests for an empirical model defined by S and the marker data.

Example 1 (Two score functions). With binary and scalar ($r = 1$) phenotypes, the two possible values $Y_i = 1$ and $Y_i = 0$ refer to "affected" and "unaffected." This is the setup of nonparametric linkage (NPL) analysis. If the disease is rare, including only the affected pedigree members generally involves little loss of information. Then the score function quantifies the degree of IBD allele sharing among the affected individuals, $\mathcal{P}_1 \subset \mathcal{P}$, with known disease status. Let $\text{IBD}_{i_1 i_2} = \text{IBD}_{i_1 i_2}(\mathbf{v})$ be the number of founder alleles shared IBD by i_1 and i_2, which is either 0, 1, or 2. Whittemore and Halpern (1994) introduced a score function that is a constant multiple of

$$S_{\text{pairs}}(\mathbf{v}) = \sum_{i_1 < i_2 \in \mathcal{P}_1} \text{IBD}_{i_1 i_2}, \tag{4}$$

and focuses on pairwise allele sharing of affected pedigree members. Whittemore and Halpern (1994) also defined another score function S_{all}, by putting higher weight than S_{pairs} on inheritance vectors with simultaneous IBD sharing of many individuals. Given $\mathbf{v}$ and $i \in \mathcal{P}$, we can trace backward in the pedigree, the two (possibly identical if there are loops) founder alleles that were transmitted from the father and mother to i. In this way, there is a founder allele associated with each one of the $2|\mathcal{P}_1|$ alleles in $\mathcal{P}_1$. Each binary vector $\mathbf{u} = (u_1, \dots, u_{|\mathcal{P}_1|}) \in \{0, 1\}^{\mathcal{P}_1}$ picks out $|\mathcal{P}_1|$ of these alleles by taking the paternal and maternal allele of i when $u_i = 0$ and 1. Let $b_{jj}(\mathbf{u})$ be the number of times that the jth founder allele is picked out. Then, for a randomly chosen $\mathbf{u}$,

$$S_{\text{all}}(\mathbf{v}) = 2^{-|\mathcal{P}_1|} \sum_{\mathbf{u} \in \{0,1\}^{\mathcal{P}_1}} \prod_{j=1}^{2f} b_{jj}(\mathbf{u})! \tag{5}$$

is the average number of permutations of $\{1, \dots, |\mathcal{P}_1|\}$ that leave the terms of $|\mathcal{P}_1| = \sum_{j=1}^{2f} b_{jj}(\mathbf{u})$ intact.

A priori, without making use of information from the phenotypes, one has

$$P\big(\mathbf{v}(t) = \mathbf{w}\big) = 2^{-m} \qquad \forall\, \mathbf{w} \in \mathbb{Z}_2^m, \tag{6}$$

whether or not t is linked to the disease locus. This reflects the Mendelian mode of inheritance: for each parent, both grandparental alleles are equally likely to be transmitted to an offspring, and the outcomes of different meioses are independent. For loci unlinked to the disease, the conditional distribution of $\mathbf{v}(t)$ given phenotypes coincides with (6), because $\mathbf{v}(t)$ is then independent of the phenotypes.

Let $\mathbf{v} = \mathbf{v}(\tau)$ be the inheritance vector of $\mathcal{P}$ at the disease locus. When conditioning on the phenotype vector $\mathbf{Y} = (Y_1, \dots, Y_n)$, I regard all covariates as fixed (nonrandom) and use (6) and Bayes's rule to get

$$\begin{aligned} P(\mathbf{v} = \mathbf{w} | \mathbf{Y}) &= \frac{P(\mathbf{Y} | \mathbf{v} = \mathbf{w}) 2^{-m}}{\sum_{\mathbf{w}' \in \mathfrak{Z}_2^m} P(\mathbf{Y} | \mathbf{v} = \mathbf{w}') 2^{-m}} \\ &\propto P(\mathbf{Y} | \mathbf{v} = \mathbf{w}), \end{aligned} \tag{7}$$

for the a posteriori distribution of $\mathbf{v}$. The conditional distribution in (3) describes knowledge of the inheritance vector for a fixed pedigree, as given by the marker genes. In contrast, the conditional distribution in (7) concerns the statistical variation of $\mathbf{v}$ over a (thought to be) large population of pedigrees of the same structure as $\mathcal{P}$ and with the same phenotype vector $\mathbf{Y}$.

The classical lod score in parametric linkage analysis uses the base-10 logarithm of the likelihood ratio $P(\mathbf{Y} | \mathbf{v}(t), \tau = t) / P(\mathbf{Y})$ at locus t for perfect marker data. It follows from Bayes's rule (7) that the latter is proportional to $P(\mathbf{v}(t) | \mathbf{Y}, \tau = t)$. Hence the lod score is a function of the conditional inheritance distribution. The same is true when maximizing lod scores over genetic model parameters, the so-called "mod scores" (see Risch 1984). On the other hand, variance components techniques maximize the numerator and denominator of an approximate likelihood ratio separately (see Almasy and Blangero 1998). The result is a score function that is not a function of the conditional inheritance distribution. Recently, several alternative linkage methods have been proposed for quantitiative traits that, in contrast to variance components techniques, are functions of the conditional inheritance distributions (cf. Dudoit and Speed 1999; Sham, Zhao, Cerney, and Hewitt 2000; Goldstein, Dudoit, and Speed 2001; Sham, Purcell, Cherny, and Abecasis 2002).

The factor $P(\mathbf{Y} | \mathbf{v})$ in (7) can be expanded further by conditioning on the vector of genotypes $\mathbf{G} = (\mathbf{G}_1, \dots, \mathbf{G}_n)$. Here $\mathbf{G}_i = (a_{2i-1}, a_{2i})$ consists of the two alleles of i transmitted from the father and the mother. I label the founders $1, \dots, 2f$. It is clear that $\mathbf{G}$ is a function of the founder alleles $\mathbf{a} = (a_1, \dots, a_{2f})$ and $\mathbf{v}$. Assuming no segregation distortion (i.e., $\mathbf{a}$ and $\mathbf{v}$ are independent), independence of founder alleles (random mating), and conditional independence of phenotypes given genotypes, I get

$$\begin{aligned} P(\mathbf{Y} | \mathbf{v}) &= \sum_{\mathbf{G}} P(\mathbf{Y} | \mathbf{G}) P(\mathbf{G} | \mathbf{v}) = \sum_{\mathbf{a}} P(\mathbf{Y} | \mathbf{a}, \mathbf{v}) P(\mathbf{a}) \\ &= \sum_{a_1, \dots, a_{2f}} \prod_{i=1}^{n} P(Y_i | \mathbf{a}, \mathbf{v}) \prod_{j=1}^{2f} P(a_j) \\ &= E\left(\prod_{i=1}^{n} P(Y_i | \mathbf{a}, \mathbf{v}) \right). \end{aligned} \tag{8}$$

For simplicity, I assume that each $a_j \in \{0, 1\}$, where 0 is the normal allele and 1 is the disease allele, with $p = P(1)$ and $q = 1 - p = P(0)$. Depending on the application, each *penetrance factor*, $P(Y_i | \mathbf{G}_i) = P(Y_i | \mathbf{a}, \mathbf{v})$, can be either a probability or a density, and I put $P(Y_i | \mathbf{G}_i) = 1$ for individuals with unobserved phenotypes. If the covariates are random, then I condition on them in (7) and (8) as well. The representation (8) is crucial for the rest of the article. Notice that expectation is taken with respect to unknown founder alleles, which are treated as hidden data. Thus the penetrance factors $P(Y_i | \mathbf{a}, \mathbf{v})$ are stochastic, because they depend on $\mathbf{a}$.

Example 2 (Binary phenotypes). For binary phenotypes, as in Example 1, a logistic regression model incorporating covariates can be formulated according to

$$P(Y_i = 1|\mathbf{G}_i) = 1\Bigg/\left(1 + \exp\left(-\boldsymbol{\alpha}_{|\mathbf{G}_i|} - \sum_{l=1}^{s} \beta_l x_{il}\right)\right),$$

where $|\mathbf{G}_i| = a_{2i-1} + a_{2i}$ is the number of disease alleles of $\mathbf{G}_i$. Thus the triplet $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$ determines the genetic contribution, whereas the numbers $\{\beta_l\}$ are regression coefficients determining the influence of the corresponding covariates. The penetrance factor becomes

$$P(Y_i|\mathbf{G}_i) = \left(1 + \exp\left(-\boldsymbol{\alpha}_{|\mathbf{G}_i|} - \sum_{l=1}^{s} \beta_l x_{il}\right)\right)^{-Y_i} \times \left(1 + \exp\left(\boldsymbol{\alpha}_{|\mathbf{G}_i|} + \sum_{l=1}^{s} \beta_l x_{il}\right)\right)^{-(1-Y_i)}.$$

The dimension of $\boldsymbol{\alpha}$ might be reduced. For a dominant model, put $\alpha_1 = \alpha_2$ and for a recessive model, put $\alpha_0 = \alpha_1$.

Example 3 (Gaussian phenotypes). When the phenotypes are quantitative and scalar ($r = 1$), it is common to assume that the conditional distribution of phenotypes given genotypes (and covariates) is Gaussian,

$$Y_i|\mathbf{G}_i \in N\left(m_{|\mathbf{G}_i|} + \sum_{l=1}^{s} \beta_l x_{il}, \sigma^2\right), \tag{9}$$

where $\{\beta_l\}_{l=1}^{q}$ are regression coefficients and σ^2 is the residual (environmentally caused) variance. In this case, the penetrance factor is a density

$$P(Y_i|\mathbf{G}_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(Y_i - m_{|\mathbf{G}_i|} - \sum_{l=1}^{s} \beta_l x_{il}\right)^2\right).$$

Example 4 (Survival analysis). Let $r = 1$ and let Y_i be a dichotomous disease status indicator as in Example 2. The first covariate, $x_{i1} = t_i$, is the age of disease onset if $Y_i = 1$ and the last known unaffected age if $Y_i = 0$. The remaining covariates, $x_{i2}, \dots, x_{ip}$, are individual characteristics for pedigree member i. A proportional hazards model for the penetrance function is

$$P(Y_i|\mathbf{G}_i) = \begin{cases} \exp\bigl(-\Lambda(t_i)\bigr), & Y_i = 0 \\ \lambda(t_i)\exp\bigl(-\Lambda(t_i)\bigr), & Y_i = 1, \end{cases}$$

where $\Lambda(t) = \int_0^t \lambda(u)\,du$ is the cumulative hazard function,

$$\lambda(t) = \lambda_0(t) \exp\left(\boldsymbol{\alpha}_{|\mathbf{G}_i|} + \sum_{l=2}^{p} \beta_l x_{il}\right),$$

is the hazard function, $\lambda_0(\cdot)$ is the baseline hazard and $\{\beta_l\}$ is the regression coefficients. As in Example 2, the triplet $(\alpha_0, \alpha_1, \alpha_2)$ determines the genetic contribution to the disease. Note that the penetrance factor is a probability for unaffected individuals and a density for affected ones. More details on and extensions of this model have been given by Thomas and Gauderman (1996).

3. DEFINING STOCHASTIC PENETRANCES

For each pedigree member i, I introduce $j_i = j_i(\mathbf{v})$ and $k_i = k_i(\mathbf{v})$ as the founder allele indeces of the two alleles that i receives from the father and the mother. The vector $\{(j_i, k_i)\}_{i=1}^{n}$ of length $2n$ is the *gene identity state* of $\mathcal{P}$ at the disease locus (cf. Thompson 1974). Because

$$P(\mathbf{Y}_i|\mathbf{G}_i) = \begin{cases} P\bigl(\mathbf{Y}_i|(00)\bigr) \\ \qquad \text{if } a_{j_i} = a_{k_i} = 0 \\ P\bigl(\mathbf{Y}_i|(10)\bigr) \\ \qquad \text{if } a_{j_i} = 0, a_{k_i} = 1, \text{ or } a_{j_i} = 1, a_{k_i} = 0 \\ P\bigl(\mathbf{Y}_i|(11)\bigr) \\ \qquad \text{if } a_{j_i} = a_{k_i} = 1, \end{cases} \tag{10}$$

$P(\mathbf{Y}_i|\mathbf{G}_i)$ is a symmetric bivariate function of a_{j_i} and a_{k_i}. Introduce

$$\xi_j = (a_j - p)/\sqrt{pq} \tag{11}$$

and, if $j \neq k$,

$$\xi_{jk} = \xi_j \xi_k. \tag{12}$$

Under random mating, note that $\{a_j\}_{j=1}^{2f}$ are iid with a binomial $(1, p)$ distribution. Hence ξ_j is the standardized version of a_j, and

$$\begin{aligned} E(\xi_j) &= E(\xi_{jk}) = 0, \\ E(\xi_j \xi_k) &= \mathcal{N}_{\{j=k\}}, \\ E(\xi_j \xi_{kl}) &= 0, \\ E(\xi_{jk} \xi_{lm}) &= \mathcal{N}_{\{(j,k)=(l,m)\}} + \mathcal{N}_{\{(j,k)=(m,l)\}}. \end{aligned} \tag{13}$$

Appendix A, describes how to expand bivariate functions of a_j and a_k as linear combinations of the random variables in (11) and (12). This expansion is similar to the variance components technique of Kempthorne (1957), which has been used by a number of authors to derive IBD probabilities for dichotomous phenotypes. Suarez, Rice, and Reich (1978), and Risch (1990b) consideed affected sib pairs, and Feingold et al. (1993), Feingold and Siegmund (1997), and Teng and Siegmund (1997) treated other kinds of pedigrees with up to four affected members. In the following two sections I generalize these results by deriving explicit expressions for the inheritance distribution $P(\mathbf{v}|\mathbf{Y})$ for arbitrary pedigrees and phenotypes.

4. EXPONENTIAL EXPANSION OF PENETRANCES

In this section I expand the logarithm of $P(\mathbf{Y}_i|\mathbf{G}_i)$, which depends on the unknown and stochastic $\mathbf{G}_i$. By symmetry, $\mathbf{G}_i$ can attain one of three values, (00), (10), or (11). I introduce

$$\begin{aligned} \kappa_{ai} &= \sqrt{pq}\Bigl(p\bigl(\log P\bigl(\mathbf{Y}_i|(11)\bigr) - \log P\bigl(\mathbf{Y}_i|(10)\bigr)\bigr) \\ &\qquad + q\bigl(\log P\bigl(\mathbf{Y}_i|(10)\bigr) - \log P\bigl(\mathbf{Y}_i|(00)\bigr)\bigr)\Bigr), \\ \kappa_{di} &= pq\bigl(\log P\bigl(\mathbf{Y}_i|(11)\bigr) \\ &\qquad - 2\log P\bigl(\mathbf{Y}_i|(10)\bigr) + \log P\bigl(\mathbf{Y}_i|(00)\bigr)\bigr), \\ \kappa_{li} &= \sqrt{pq}\bigl(\log P\bigl(\mathbf{Y}_i|(11)\bigr) - \log P\bigl(\mathbf{Y}_i|(00)\bigr)\bigr), \end{aligned} \tag{14}$$

as the additive, dominant, and loop random fluctuations of $\log P(\mathbf{Y}_i|\mathbf{G}_i)$.

Given any fixed (nonrandom) array $\mathbf{B} = (B_{jk})_{1\le j\le k\le 2f}$, I introduce

$$M_\xi(\mathbf{B}) = E\left(\exp\left(\sum_{j=1}^{2f} B_{jj}\xi_j + \sum_{1\le j<k\le 2f} B_{jk}\xi_{jk}\right)\right),$$

as the moment-generating function of

$$\xi = \{\xi_j\}_{j=1}^{2f} \cup \{\xi_{jk}\}_{1\le j<k\le 2f}$$

defined in (11) and (12). It turns out that the conditional inheritance distribution involves $M_\xi(\mathbf{B})$, with coefficients $\mathbf{B}$ depending on the quantities κ_{ai}, κ_{di}, and κ_{li}.

Proposition 1. The conditional distribution of the inheritance vector is given by

$$P(\mathbf{v}|\mathbf{Y}) \propto \exp\left(\sum_{i\in\bar{\mathcal{R}}} \kappa_{di}\right) M_\xi(\mathbf{B}), \tag{15}$$

where the components of $\mathbf{B} = (B_{jk})_{1\le j\le k\le 2f}$ are defined by

$$B_{jk} = B_{jk}(\mathbf{v}) = \begin{cases} \sum_{i\in\mathcal{R}_{jk}} \kappa_{di}, & j<k \\ \sum_{i\in\mathcal{R}_j} \kappa_{ai} + \sum_{i\in\bar{\mathcal{R}}_j} \kappa_{li}, & j=k, \end{cases} \tag{16}$$

$$\begin{aligned} \mathcal{R} &= \mathcal{R}(\mathbf{v}) = \{i \in \mathcal{P}_{\text{known}}, j_i \neq k_i\}, \\ \bar{\mathcal{R}} &= \bar{\mathcal{R}}(\mathbf{v}) = \{i \in \mathcal{P}_{\text{known}}, j_i = k_i\}, \\ \mathcal{R}_j &= \mathcal{R}_j(\mathbf{v}) = \{i \in \mathcal{R};\ j_i = j \text{ or } k_i = j\}, \\ \mathcal{R}_{jk} &= \mathcal{R}_{jk}(\mathbf{v}) = \{i \in \mathcal{R};\ (j_i,k_i) = (j,k) \text{ or } (k,j)\}, \\ \bar{\mathcal{R}}_j &= \bar{\mathcal{R}}_j(\mathbf{v}) = \{i \in \bar{\mathcal{R}};\ j_i = k_i = j\}. \end{aligned} \tag{17}$$

and $\mathcal{P}_{\text{known}}$ is the collection of all pedigree members with known disease phenotypes.

Note that $\bar{\mathcal{R}}(\mathbf{v}) = \varnothing$ for all v if the pedigree contains no loops. Thus, the first factor on the right side of (15) vanishes for outbred pedigrees, and then $P(\mathbf{v}|\mathbf{Y})$ can be expressed solely in terms of the moment-generating function of the array ξ.

Example 5 (Survival analysis, continued). Consider the survival analysis model of Example 4. Then

$$\log P(Y_i|\mathbf{G}_i) = \begin{cases} -e^{\boldsymbol{\alpha}_{|\mathbf{G}_i|}}\Lambda_i, & Y_i = 0 \\ \boldsymbol{\alpha}_{|\mathbf{G}_i|} + \log\lambda_i(t_i) - e^{\boldsymbol{\alpha}_{|\mathbf{G}_i|}}\Lambda_i, & Y_i = 1, \end{cases}$$

where $\lambda_i(t) = \lambda_0(t)\exp(\sum_{l=2}^p \beta_l x_{il})$ and $\Lambda_i = \int_0^{t_i}\lambda_i(u)\,du$. Inserting this into (14) yields

$$\kappa_{ai} = -\Lambda_i\kappa_a^I + \mathcal{N}_{\{Y_i=1\}}\kappa_a^{II}, \tag{18}$$

with $\kappa_a^I = \sqrt{pq}(p(e^{\alpha_2} - e^{\alpha_1}) + q(e^{\alpha_1} - e^{\alpha_0}))$ and $\kappa_a^{II} = \sqrt{pq}(p(\alpha_2 - \alpha_1) + q(\alpha_1 - \alpha_0))$. The dominance and loop components κ_{di} and κ_{li} are defined analogously in terms of κ_d^I, κ_d^{II}, κ_l^I, and κ_l^{II}. For an outbred pedigree, $P(\mathbf{v}|\mathbf{Y}) \propto M_\xi(\mathbf{B})$, with

$$B_{jk} = \begin{cases} -\kappa_a^I \sum_{i\in\mathcal{R}_j} \Lambda_i + \kappa_a^{II}|\mathcal{R}_j^1|, & j=k \\ -\kappa_d^I \sum_{i\in\mathcal{R}_{jk}} \Lambda_i + \kappa_d^{II}|\mathcal{R}_{jk}|, & j<k, \end{cases}$$

with $\mathcal{R}_j^1 = \{i \in \mathcal{R}_j;\ Y_i = 1\}$ and $\mathcal{R}_{jk}^1 = \{i \in \mathcal{R}_{jk};\ Y_i = 1\}$.

We know that $\{\xi_j\}_{j=1}^{2f}$ are iid and, from (13), all components of ξ are uncorrelated. Thus it is tempting to treat ξ as if it had independent components and try the approximation

$$P(\mathbf{v}|\mathbf{Y}) \overset{\approx}{\propto} \exp\left(\sum_{i\in\bar{\mathcal{R}}} \kappa_{di}\right) \prod_{j=1}^{2f} M_{\xi_j}(B_{jj}) \cdot \prod_{1\le j<k\le 2f} M_{\xi_{jk}}(B_{jk}), \tag{19}$$

where $M_{\xi_j}(t) = E(\exp(t\xi_j))$ and $M_{\xi_{jk}}(t) = E(\exp(t\xi_{jk}))$ are the moment generating functions of ξ_j and ξ_{jk}. The approximation in (18) is better with smaller dominance components κ_{di}, because then all B_{jk}, $j<k$, are small as well. When $\kappa_{di} \equiv 0$, the easily-computed formula

$$P(\mathbf{v}|\mathbf{Y}) \propto \prod_{j=1}^{2f} M_{\xi_j}(B_{jj}) \tag{20}$$

is exact. This is the case for binary phenotypes of Example 6 when $\mathcal{P}_{\text{known}}$ consists only of affecteds and $f_1 = \sqrt{f_0 f_2}$. Further, $\kappa_{di} \approx 0$ in Example 3, when $m_1 = (m_0 + m_2)/2$ and $|m_2 - m_0| \ll \sigma$, that is, an additive model with a small genetic component. Formula (20) is also of interest for rare diseases $p \to 0$, as is explored further in Section 7.

5. LINEAR EXPANSION OF PENETRANCES

For some genetic models, it is more convenient to expand the penetrance factor $P(\mathbf{Y}_i|\mathbf{G}_i)$, rather than its logarithm, in terms of $\{\xi_j\}$. This is true for binary phenothypes when the penetrance probabilities are parameterized by affection probabilities given genotypes. More importantly, the linear expansion shows more clearly how the conditional inheritance distribution $P(\mathbf{v}|\mathbf{Y})$ is split into different terms involving simultaneous allele sharing of a varying number of individuals.

I start by introducing

$$\tilde{\mu}_i = q^2 P\big(\mathbf{Y}_i|(00)\big) + 2pqP\big(\mathbf{Y}_i|(10)\big) + p^2 P\big(\mathbf{Y}_i|(11)\big) \tag{21}$$

as the average penetrance of individual i and

$$\begin{aligned} \tilde{\kappa}_{ai} &= \sqrt{pq}\big(p\big(P(\mathbf{Y}_i|(11)) - P(\mathbf{Y}_i|(10))\big) \\ &\qquad + q\big(P(\mathbf{Y}_i|(10)) - P(\mathbf{Y}_i|(00))\big)\big)/\tilde{\mu}_i, \\ \tilde{\kappa}_{di} &= pq\big(P(\mathbf{Y}_i|(11)) \\ &\qquad - 2P(\mathbf{Y}_i|(10)) + P(\mathbf{Y}_i|(00))\big)/\tilde{\mu}_i, \\ \tilde{\kappa}_{li} &= \sqrt{pq}\big(P(\mathbf{Y}_i|(11)) - P(\mathbf{Y}_i|(00))\big)/\tilde{\mu}_i, \end{aligned} \tag{22}$$

as the amount of additive, dominant, and loop random fluctuation of the ith penetrance factor.

The conditional inheritance distribution in this section is built on statistics $\tilde{T}_{\mathcal{Q}} = \tilde{T}_{\mathcal{Q}}(\mathbf{v})$, which involve simultaneous allele sharing of the pedigree members in $\mathcal{Q} \subset \mathcal{P}_{\text{known}}$. When $\mathcal{Q} = \{i\}$ consists of just one individual,

$$\tilde{T}_i = \tilde{\kappa}_{di}\mathrm{HBD}_i, \tag{23}$$

where $\mathrm{HBD}_i = \mathrm{HBD}_i(\mathbf{v})$ is 1 if i has both its alleles homozygous by descent and 0 otherwise. For pairs $\mathcal{Q} = \{i_1, i_2\}$, the statistic is

$$\tilde{T}_{i_1 i_2} = \begin{cases} \tilde{\kappa}_{ai_1}\tilde{\kappa}_{ai_2}\mathrm{IBD}_{i_1 i_2} \\ \quad + \tilde{\kappa}_{di_1}\tilde{\kappa}_{di_2}\mathcal{N}_{\{\mathrm{IBD}_{i_1 i_2}=2\}}, & i_1, i_2 \in \mathcal{R} \\ \tilde{\kappa}_{ai_1}\tilde{\kappa}_{li_2}\mathrm{IBD}_{i_1 i_2}, & i_1 \in \mathcal{R}, i_2 \in \bar{\mathcal{R}} \\ \tilde{\kappa}_{di_1}\tilde{\kappa}_{di_2} + \tilde{\kappa}_{li_1}\tilde{\kappa}_{li_2}\mathrm{IBD}_{i_1 i_2}/2, & i_1, i_2 \in \bar{\mathcal{R}}. \end{cases} \tag{24}$$

When $|\mathcal{Q}| \geq 3$, $\tilde{T}_{\mathcal{Q}}$ has quite a complicated expression when $\mathcal{Q} \cap \bar{\mathcal{R}} \neq \varnothing$. I thus restrict the investigation to outbred pedigrees, $\bar{\mathcal{R}} = \varnothing$: Given any array $b = (b_{jk}; 1 \leq j \leq k \leq 2f)$ of natural numbers, I define the bth central moment of ξ,

$$\begin{aligned}\nu_b &= E(\xi^b)\\ &= E\left(\prod_{j=1}^{2f} \xi_j^{b_{jj}} \cdot \prod_{1\leq j<k\leq 2f} \xi_{jk}^{b_{jk}}\right) = \prod_{j=1}^{2f} \nu_{b_j},\end{aligned} \tag{25}$$

where $\nu_k = E(\xi_j^k) = (pq)^{-(k-2)/2}((-1)^k p^{k-1} + q^{k-1})$ for natural numbers k and $b_j = b_{jj} + \sum_{k<j} b_{kj} + \sum_{k>j} b_{jk}$. In the last step of (25), I used (12) and the independence of $\{\xi_j\}_{j=1}^{2f}$.

For any $\mathcal{Q} \subset \mathcal{P}_{\text{known}}$ with $|\mathcal{Q}| \geq 2$, I define the vector $\mathbf{u} = (u_i, i \in \mathcal{Q})$. Here u_i decides which alleles of i to pick the grandpaternal allele ($u_i = 0$), the grandmaternal allele ($u_i = 1$), or both alleles ($u_i = 2$). Then

$$\tilde{T}_{\mathcal{Q}} = \tilde{T}_{\mathcal{Q}}(\mathbf{v}) = \sum_{\mathbf{u}\in\{0,1,2\}^{\mathcal{Q}}} \prod_{i\in\mathcal{Q}_{01}(\mathbf{u})} \tilde{\kappa}_{ai} \prod_{i\in\mathcal{Q}_2(\mathbf{u})} \tilde{\kappa}_{di} \cdot \nu_{b(\mathbf{u})}, \tag{26}$$

where $\mathcal{Q}_{01}(\mathbf{u}) \cup \mathcal{Q}_2(\mathbf{u})$ is a disjoint decomposition of $\mathcal{Q}$ into subsets with $u_i \in \{0,1\}$ and $u_i \in \{2\}$, and $b(\mathbf{u}) = (b_{jk}(\mathbf{u}), 1 \leq j \leq k \leq 2f)$ is defined by

$$b_{jk}(\mathbf{u}) = \begin{cases} \big|\{i \in \mathcal{Q};\ u_i = 0,\ j_i = j \text{ or } u_i = 1, k_i = j\}\big|, \\ \qquad j = k \\ \big|\{i \in \mathcal{Q};\ u_i = 2 \text{ and } (j_i, k_i) = (j,k) \text{ or } (k,j)\}\big|, \\ \qquad j < k. \end{cases}$$

For additive models ($\tilde{\kappa}_{di} \equiv 0$), all terms with $\mathcal{Q}_2(\mathbf{u}) \neq \varnothing$ will vanish in (26). Hence, only one allele is picked from each individual. Notice further that (26) also holds for $|\mathcal{Q}| = 1, 2$ when $\bar{\mathcal{R}} = \varnothing$, yielding $T_i = 0$ and $T_{i_1 i_2}$ equal to the first line of (24). I gave an alternative definition of $T_{\mathcal{Q}}$ in terms of IBD-sharing quantities in earlier work (Hössjer 2001).

I can now state the following result.

Proposition 2. The conditional distribution of the inheritance vector can be expanded as

$$P(\mathbf{v}|\mathbf{Y}) \propto 1 + \sum_{l=1}^{|\mathcal{P}_{\text{known}}|} \tilde{S}_l(\mathbf{v}), \tag{27}$$

where

$$\tilde{S}_l(\mathbf{v}) = \sum_{\substack{\mathcal{Q}\subset\mathcal{P}_{\text{known}}\\ |\mathcal{Q}|=l}} \tilde{T}_{\mathcal{Q}}, \tag{28}$$

involves the simultaneous allele sharing of l individuals, $l = 1, \ldots, |\mathcal{P}_{\text{known}}|$.

Example 6 (Binary phenotypes, continued). As a particular case of Example 2 without covariates, I consider binary phenotypes

$$P(Y_i = 1|\mathbf{G}_i) = f_{|\mathbf{G}_i|}, \tag{29}$$

where f_0, f_1, and f_2 are the penetrance probabilites of the disease [i.e., $f_i = 1/(1 + \exp(-\alpha_i))$ in Example 2]. Based on this binary model, Feingold et al. (1993), and Teng and Siegmund (1997) have established expansions of the kind given in (27) for additive models and outbred pedigrees with up to four affected members. The prevalence and the additive and dominant genetic variance components are

$$\begin{aligned} K_p &= P(Y_i = 1) = q^2 f_0 + 2pqf_1 + p^2 f_2,\\ \tilde{\sigma}_a^2 &= 2pq\big(p(f_2 - f_1) + q(f_1 - f_0)\big)^2,\\ \tilde{\sigma}_d^2 &= p^2 q^2 (f_2 - 2f_1 + f_0)^2. \end{aligned} \tag{30}$$

Then, with $R = K_p/(1 - K_p)$ the prevalence odds ratio, as shown in Appendix D,

$$\tilde{S}_1(\mathbf{v}) \propto \sum_{i\in\mathcal{P}_1} \mathrm{HBD}_i - R \sum_{i\in\mathcal{P}_0} \mathrm{HBD}_i, \tag{31}$$

where $\mathcal{P}_0$ and $\mathcal{P}_1$ are the subsets of individuals in $\mathcal{P}_{\text{known}}$ with $Y_i = 0$ and $Y_i = 1$. For small prevalences, $\tilde{S}_1$ approximately equals $S_{\sharp\text{aff HBD}} = \sum_{i\in\mathcal{P}_1} \mathrm{HBD}_i$. This is the number of affected individuals homozygous by descent, a score function introduced by McPeek (1999). For outbred pedigrees,

$$\begin{aligned} \tilde{S}_2(\mathbf{v}) \quad &\propto \quad .5(1-c) \cdot \Bigg(\sum_{i_1<i_2\in\mathcal{P}_1} \mathrm{IBD}_{i_1 i_2} - R \sum_{\substack{i_1\in\mathcal{P}_1\\ i_2\in\mathcal{P}_0}} \mathrm{IBD}_{i_1 i_2} \\ &\qquad + R^2 \sum_{i_1<i_2\in\mathcal{P}_0} \mathrm{IBD}_{i_1 i_2} \Bigg) \\ &\qquad + c \cdot \Bigg(\sum_{i_1<i_2\in\mathcal{P}_1} \mathcal{N}_{\{\mathrm{IBD}_{i_1 i_2}=2\}} - R \sum_{\substack{i_1\in\mathcal{P}_1\\ i_2\in\mathcal{P}_0}} \mathcal{N}_{\{\mathrm{IBD}_{i_1 i_2}=2\}} \\ &\qquad + R^2 \sum_{i_1<i_2\in\mathcal{P}_0} \mathcal{N}_{\{\mathrm{IBD}_{i_1 i_2}=2\}} \Bigg) \\ &\overset{K_p\ll 1}{\approx} \quad .5(1-c) S_{\text{pairs}}(\mathbf{v}) + c S_{\text{g-prs}}(\mathbf{v}), \end{aligned} \tag{32}$$

where $c = \tilde{\sigma}_d^2/(\tilde{\sigma}_a^2 + \tilde{\sigma}_d^2)$ is the proportion of total genetic variance due to dominance effects, S_{pairs} is defined in (4) and $S_{\text{g-prs}} = \sum_{i_1<i_2\in\mathcal{P}_1} \mathcal{N}_{\{\mathrm{IBD}_{i_1 i_2}=2\}}$ counts the number of pairs of affected pedigree members that have the same genotype IBD (see McPeek 1999). In other words, $\tilde{S}_2$ is essentially a weighted average of $.5S_{\text{pairs}}$ and $S_{\text{g-prs}}$ when the prevalence is small. The weight for $.5S_{\text{pairs}}$ is 1 in the additive case. I gave results for higher-order score functions earlier (Hössjer 2001). In particular, I showed that S_{all} in (5) is somewhat related to $T_{\mathcal{P}_1} = \tilde{S}_{|\mathcal{P}_1|}$ in the additive case.

6. LOCAL PENETRANCE MODELS

As illustrated in the preceding examples, the penetrance function commonly includes a set of parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)$, which are either known (i.e., estimated before the linkage analysis) or unknown and then estimated simultaneously with the linkage analysis. Thus, applying penetrance factors $P_{\boldsymbol{\psi}}(\mathbf{Y}_i|\mathbf{G}_i)$ in (8), using (6) and Bayes's rule (7), yields an inheritance distribution $P_{\boldsymbol{\psi}}(\mathbf{v}|\mathbf{Y})$ depending on $\boldsymbol{\psi}$.

Assume a one-dimensional submodel, that is, a family of penetrance parameters

$$\{\boldsymbol{\psi}_\varepsilon; \varepsilon \geq 0\}, \tag{33}$$

where $\boldsymbol{\psi}_0$ corresponds to no genetic contribution. This means that $P_{\boldsymbol{\psi}_0}(\mathbf{Y}_i|\mathbf{G}_i)$ is independent of $\mathbf{G}_i$, and hence (6) and (7) imply that $P_{\boldsymbol{\psi}_0}(\mathbf{v}=\mathbf{w}|\mathbf{Y}) \equiv 2^{-m}$. The larger the ε, the stronger the genetic effect and the more nonuniform the $P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y})$ distribution. The degree of association between the inheritance vector and the disease phenotypes, that is, the degree of "nonuniformity" of $P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y})$, was termed "specificity" by Thompson (1997). Thus ε may be considered a specificity parameter of the genetic model.

6.1 Exponential Expansions

Suppose that there exists a positive integer ρ such that

$$P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y}) = 2^{-m}\exp\bigl(\varepsilon^\rho S(\mathbf{v})/\rho! + o(\varepsilon^\rho)\bigr) \tag{34}$$

as $\varepsilon \to 0$. Following Whittemore (1996) and Commenges (1994), this can be used to define a likelihood score function. In fact, suppose that $P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y}) = L(\varepsilon;\mathbf{v})$ is interpreted as a likelihood function for ε. The likelihood score function at $\varepsilon = 0$ is $d\log^\rho L(\varepsilon;\mathbf{v})/d\varepsilon^\rho|_{\varepsilon=0} = S(\mathbf{v})$, provided the reminder term in (34) is sufficiently smooth. Under the idealized assumption that the disease locus is known, $S(\cdot)$ yields a *locally most powerful* test for testing $\varepsilon = 0$ against local alternatives $\varepsilon > 0$ (see, e.g., Cox and Hinkley 1974). Expansions of the kind shown in (34) were used by Kong and Cox (1997) and Nicolae (1999) as empirical likelihood models, in which ε is estimated at each locus as a way to perform linkage analysis.

Viewing the penetrance factors $P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{Y}_i|\mathbf{G}_i)$ as functions of ε, the quantities $\kappa_{ai} = \kappa_{ai}(\varepsilon)$, $\kappa_{di} = \kappa_{di}(\varepsilon)$, and $\kappa_{li} = \kappa_{li}(\varepsilon)$ in (14) will also depend on ε. Because $P_{\boldsymbol{\psi}_0}(\mathbf{Y}_i|\mathbf{G}_i)$ is independent of $\mathbf{G}_i$, it follows from (14) that $\kappa_{ai}(0) = \kappa_{di}(0) = \kappa_{li}(0) = 0$. Hence it is reasonable to assume that for small $\varepsilon \geq 0$,

$$\begin{aligned}\kappa_{ai}(\varepsilon) &= \kappa'_{ai}(0)\varepsilon + o(\varepsilon),\\ \kappa_{li}(\varepsilon) &= \kappa'_{li}(0)\varepsilon + o(\varepsilon),\\ \kappa_{di}(\varepsilon) &= \kappa'_{di}(0)\varepsilon + \tfrac{1}{2}\kappa''_{di}(0)\varepsilon^2 + o(\varepsilon^2).\end{aligned} \tag{35}$$

Given a pedigree member $i \in \mathcal{P}_{\text{known}}$, I define a score function

$$T_i(\mathbf{v}) = \kappa'_{di}(0)\text{HBD}_i, \tag{36}$$

with HBD_i as defined before (23). Further, given two members i_1 and i_2 of $\mathcal{P}_{\text{known}}$, I define a "pair score function" $T_{i_1i_2} = T_{i_1i_2}(v)$ according to

$$T_{i_1i_2} = \begin{cases}\kappa'_{ai_1}(0)\kappa'_{ai_2}(0)\text{IBD}_{i_1i_2} & \\ \quad + \kappa'_{di_1}(0)\kappa'_{di_2}(0)\mathcal{N}_{\{\text{IBD}_{i_1i_2}=2\}}, & i_1, i_2 \in \mathcal{R},\\ \kappa'_{ai_1}(0)\kappa'_{li_2}(0)\text{IBD}_{i_1i_2}, & i_1 \in \mathcal{R}, i_2 \in \bar{\mathcal{R}},\\ \kappa'_{li_1}(0)\kappa'_{ai_2}(0)\text{IBD}_{i_1i_2}, & i_1 \in \bar{\mathcal{R}}, i_2 \in \mathcal{R},\\ \kappa''_{di}(0)\mathcal{N}_{\{i_1=i_2\}} & \\ \quad + \kappa'_{li_1}(0)\kappa'_{li_2}(0)\text{IBD}_{i_1i_2}/2, & i_1, i_2 \in \bar{\mathcal{R}},\end{cases} \tag{37}$$

with $\text{IBD}_{i_1i_2}$ as defined in Example 1. Note the similarity between T_i and $T_{i_1i_2}$ and $\tilde{T}_i$ and $\tilde{T}_{i_1i_2}$ in (23) and (24).

The following corollary of Proposition 1 can now be stated.

Corollary 1. Consider a one-parameter family, $\{P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{Y}_i|\mathbf{G}_i)\}_{\varepsilon\geq 0}$ of genetic models satisfying (35). Then, for small ε, the distribution of the inheritance vector can be expanded as

$$\begin{aligned}P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y}) = 2^{-m}\exp\Bigl(&\varepsilon\bigl(S_1(\mathbf{v}) - E_0(S_1)\bigr)\\ &+ \varepsilon^2\bigl(S_2(\mathbf{v}) - E_0(S_2) - \text{var}_0(S_1)\bigr)/2 + o(\varepsilon^2)\Bigr),\end{aligned} \tag{38}$$

where

$$\begin{aligned}S_1(\mathbf{v}) &= \sum_{i\in\mathcal{P}_{\text{known}}} T_i,\\ S_2(\mathbf{v}) &= \sum_{i_1,i_2\in\mathcal{P}_{\text{known}}} T_{i_1i_2},\end{aligned} \tag{39}$$

$E_0(S_l)$ and $\text{var}_0(S_l)$ denote expectation and variance of $S_l(\mathbf{v})$ taken under $\varepsilon = 0$ [i.e., $P_{\boldsymbol{\psi}_0}(\mathbf{v}|\mathbf{Y}) \equiv 2^{-m}$], and T_i and $T_{i_1i_2}$ are as defined in (36) and (37).

For inbred pedigrees, $\kappa'_{di}(0) \neq 0$ for at least some pedigree member. Therefore, S_1 does not vanish, and the likelihood score function is $S = S_1 - E_0(S_1)$ in (34), with $\rho = 1$. For outbred pedigrees, $\bar{\mathcal{R}} = \varnothing$. Hence $S_1 \equiv 0$ and $\rho = 2$ in (34), yielding the likelihood score function $S = S_2 - E_0(S_2)$. Further, $T_{i_1i_2}$ simplifies to

$$T_{i_1i_2}(\mathbf{v}) = \kappa'_{ai_1}(0)\kappa'_{ai_2}(0)\text{IBD}_{i_1i_2} + \kappa'_{di_1}(0)\kappa'_{di_2}(0)\mathcal{N}_{\{\text{IBD}_{i_1i_2}=2\}} \tag{40}$$

for outbred pedigrees. The score function S_2 based on (40) was essentially derived by Commenges (1994) for outbred pedigrees and an empirical additive likelihood model involving stochastic founder allele effects but not disease allele frequencies.

For strong genetic models (i.e., large $\varepsilon > 0$), higher-order cumulants of ξ and larger exponents of $B'_{jk}(0)$ will dominate. This corresponds to simultaneous allele sharing of more than two individuals. In principle, $T_{\mathcal{Q}}$ can be defined for subsets $\mathcal{Q} \subset \mathcal{P}_{\text{known}}$ with more than two individuals, although the complexity of the expressions quickly increases with $|\mathcal{Q}|$.

Example 7 (Gaussian phenotypes). The penetrance parameters in Example 3 are $\boldsymbol{\psi} = (m_0, m_1, m_2, \boldsymbol{\beta}, \sigma^2)$, where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_s)^T$ is the vector of regression parameters. Let m_0^*, m_1^*, and m_2^* be the mean parameters of a fixed reference model. When $\boldsymbol{\beta} = \mathbf{0}$, the reference model has mean phenotype $E(Y_i) = q^2m_0^* + 2pqm_1^* + p^2m_2^* =: m$ and total variance $V(Y_i) = \sigma_g^2 + \sigma^2$, where $\sigma_g^2 = q^2(m_0^* - m)^2 + 2pq(m_1^* - m)^2 + p^2(m_2^* - m)^2$ is the genetic variance. This can be split into additive and dominant variance components as $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$, where

$$\begin{aligned}\sigma_a^2 &= 2pq\bigl(p(m_2^* - m_1^*) + q(m_1^* - m_0^*)\bigr)^2,\\ \sigma_d^2 &= (pq)^2(m_2^* - 2m_1^* + m_0^*)^2.\end{aligned} \tag{41}$$

Introduce a one-parameter trajectory of penetrance parameters in a direction toward the reference model,

$$\begin{aligned}\boldsymbol{\psi}_\varepsilon = \bigl(&m + \varepsilon\sigma(m_0^* - m)/\sigma_g, m + \varepsilon\sigma(m_1^* - m)/\sigma_g,\\ &m + \varepsilon\sigma(m_2^* - m)/\sigma_g, \boldsymbol{\beta}, \sigma^2\bigr),\end{aligned}$$

for $\varepsilon \geq 0$. Note that $E(Y_i) = m$ independently of ε whereas the heritability (i.e., the ratio between genetic and total variance) is $\varepsilon^2/(\varepsilon^2 + 1)$ and hence grows with ε.

Let $r_i = (Y_i - m - \sum_{l=1}^{s} x_{il}\beta_l)/\sigma$ be the ith standardized residual. It is proved in Appendix E that the likelihood score function for an outbred pedigree is

$$S = S_2 - E_0(S_2) = 2\sum_{i_1<i_2} r_{i_1} r_{i_2}\big(C_{i_1i_2} - E_0\big(C_{i_1i_2}\big)\big), \tag{42}$$

where

$$C_{i_1i_2} = (1-c)\cdot \mathrm{IBD}_{i_1i_2}/2 + c\cdot \mathcal{N}_{\{\mathrm{IBD}_{i_1i_2}=2\}} \tag{43}$$

and $c = \sigma_d^2/\sigma_g^2$ is the fraction of genetic variance for the reference model due to dominance effects. This score function is discussed further in Section 8.

Because polygenic or shared environmental effects are not included, the vector $\mathbf{Y}|\mathbf{G}$ has independent, normal components, whereas $\mathbf{Y}|\mathbf{v}$ is a mixture of multivariate normals. For an outbred pedigree with all Y_i observed, the first two moments of $\mathbf{Y}|\mathbf{v}$ are

$$\begin{aligned} E_{\boldsymbol{\psi}_\varepsilon}(\mathbf{Y}|\mathbf{v}) &= m\mathcal{N} + \mathbf{X}\boldsymbol{\beta}, \\ \mathrm{cov}_{\boldsymbol{\psi}_\varepsilon}(\mathbf{Y}|\mathbf{v}) &= \big(\mathrm{cov}_{\boldsymbol{\psi}_\varepsilon}(Y_{i_1}, Y_{i_2})|\mathbf{v}\big)_{i_1,i_2=1}^{n} \\ &= \sigma^2(\mathbf{I} + \varepsilon^2\mathbf{C}), \end{aligned} \tag{44}$$

where $\mathcal{N} = (1,\dots,1)^T$ is an n-dimensional column vector, $\mathbf{I}$ is the $n\times n$ identity matrix, $\mathbf{C} = (C_{i_1i_2})$ and $\mathbf{X} = (x_{il};\, 1\le i\le n,\ 1\le l\le s)$ the design matrix. It is common in linkage analysis to assume multivariate normality of $\mathbf{Y}|\mathbf{v}$ based on (44) (see, e.g., Almasy and Blangero 1998). I have shown (Hössjer 2001) that the multivariate normal assumption is locally correct, in the sense that the same likelihood score function (42) is obtained as for the exact calculation based on Corollary 1.

Example 8 (Survival analysis). Continuing Example 5, I define the one-parameter family $\boldsymbol{\psi}_\varepsilon = (\alpha_{0\varepsilon}, \alpha_{1\varepsilon}, \alpha_{2\varepsilon}, \beta_2, \dots, \beta_p)$ of genetic models with $\alpha_{i\varepsilon} = \varepsilon\alpha_i^*/\sigma_g$. Here α_0^*, α_1^*, and α_2^* are given constants of a reference model and $\sigma_g^2 = q^2(\alpha_0^*)^2 + 2pq(\alpha_1^*)^2 + p^2(\alpha_2^*)^2$. I split $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$ into additive and dominance components, with σ_a^2 and σ_d^2 defined by replacing α_i^* with m_i^* in (41). For an outbred pedigree with $\mathcal{P}_{\text{known}} = \mathcal{P}$, the likelihood score function (also see Commenges 1994) is of the form (42), with $r_i = \mathcal{N}_{\{Y_i=1\}} - \Lambda_i$ a so-called martingale residual and $c = \sigma_d^2/\sigma_g^2$ the fraction of genetic variance for the reference model due to dominance effects. Appendix E gives a derivation.

6.2 Linear Expansions

In this section, I expand the inheritance distribution linearly for small ε as

$$P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y}) = 2^{-m}\big(1 + \varepsilon^\rho S(v)/\rho! + o(\varepsilon^\rho)\big). \tag{45}$$

This is slightly different from (34), where the logarithm of $P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y})$ was used. The linear expansion is, for example, more convenient to use for binary phenotypes; refer to Example 6. As in Section 6.1, I arrive at (45) by introducing a one-parameter family of penetrance factors, $\{P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{Y}_i|\mathbf{G}_i)\}_{0\le\varepsilon\le\varepsilon_{\max}}$. To this end, I let $P^*(\mathbf{Y}_i|\mathbf{G}_i)$ denote the penetrance factors of a fixed reference model with mean penetrance $\tilde{\mu}_i$ [cf. (21)]. Then introduce

$$P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{Y}_i|\mathbf{G}_i) = \tilde{\mu}_i + \varepsilon\big(P^*(\mathbf{Y}_i|\mathbf{G}_i) - \tilde{\mu}_i\big); \tag{46}$$

that is, $P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{Y}_i|\mathbf{G}_i)$ is a linear function of ε. Note that an upper bound $\varepsilon_{\max}$ of ε is needed to ensure that $0 \le P_\varepsilon(\mathbf{Y}_i|\mathbf{G}_i) \le 1$ for all $i \in \mathcal{P}_{\text{known}}$.

Because of linearity, the penetrance variation coefficients $\tilde{\kappa}_{ai} = \tilde{\kappa}_{ai}(\varepsilon)$, $\tilde{\kappa}_{di} = \tilde{\kappa}_{di}(\varepsilon)$, and $\tilde{\kappa}_{li} = \tilde{\kappa}_{li}(\varepsilon)$ in (22) satisfy

$$\begin{aligned} \tilde{\kappa}_{ai}(\varepsilon) &= \varepsilon\kappa_{ai}^*, \\ \tilde{\kappa}_{di}(\varepsilon) &= \varepsilon\kappa_{di}^*, \\ \tilde{\kappa}_{li}(\varepsilon) &= \varepsilon\kappa_{li}^*, \end{aligned} \tag{47}$$

where κ_{ai}^*, κ_{di}^*, and κ_{li}^* are the corresponding coefficients of the reference model P^*. This is more restrictive than the exponential expansion (35).

The following local expansion of the inheritance distribution can now be derived.

Corollary 2. Consider the family (46) of genetic models. Then the distribution of the inheritance vector can be expanded as

$$P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y}) \propto 1 + \sum_{l=1}^{|\mathcal{P}_{\text{known}}|} \varepsilon^l \tilde{S}_l(\mathbf{v}). \tag{48}$$

The proportionality constant depends on ε but not on $\mathbf{v}$. Here $\{\tilde{S}_l\}_{l=1}^{|\mathcal{P}_{\text{known}}|}$ are defined as in Proposition 2, with κ_{ai}^*, κ_{di}^*, and κ_{li}^* replacing $\tilde{\kappa}_{ai}$, $\tilde{\kappa}_{di}$, and $\tilde{\kappa}_{li}$. For small $\varepsilon \ge 0$,

$$\begin{aligned} P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y}) = 2^{-m}\big(1 &+ \varepsilon\big(\tilde{S}_1(\mathbf{v}) - E_0(\tilde{S}_1)\big) \\ &+ \varepsilon^2\big(\tilde{S}_2(\mathbf{v}) - E_0(\tilde{S}_2) \\ &- E_0(\tilde{S}_1)\big(\tilde{S}_1 - E_0(\tilde{S}_1)\big)\big) + o(\varepsilon^2)\big), \end{aligned} \tag{49}$$

where E_0 denotes expectation taken under the uniform inheritance distribution (6).

The strict linear expansion (47) is needed only for establishing (48). A weaker expansion related to (35) suffices to prove (49), at the cost of more laborious notation.

From Corollary 2, I conclude that lower-order score functions $\tilde{S}_l$, involving simultaneous IBD sharing of few individuals, dominate the inheritance distribution when the genetic component is weak (i.e., ε is small). In fact, for inbred pedigrees, (45) holds with $\rho = 1$ and $S = \tilde{S}_1 - E_0(\tilde{S}_1)$, provided that the dominance components κ_{di}^* are not 0. For outbred pedigrees, the first-order term $\tilde{S}_1$ vanishes. Then $\rho = 2$ in (45) with $S = 2(\tilde{S}_2 - E_0(\tilde{S}_2))$. Higher-order $\tilde{S}_l$'s have more influence when the genetic component is strong (i.e., ε is large), corresponding to simultaneous allele sharing of many individuals.

Example 9 (Binary phenotypes, continued). In Example 6, consider a fixed reference model with penetrance probabilites f_0^*, f_1^*, and f_2^* and prevalence K_p. Then define a 1-df family of genetic models $\boldsymbol{\psi}_\varepsilon = (f_{0\varepsilon}, f_{1\varepsilon}, f_{2\varepsilon})$, where $f_{i\varepsilon} = K_p + \varepsilon K_p(f_i^* - K_p)/\tilde{\sigma}_g$. To be precise, I replaced ε by $\varepsilon K_p/\sigma_g$ in (46) to get a natural interpretation of ε. Here $\tilde{\sigma}_g^2 = q^2 \times (f_0^* - K_p)^2 + 2pq(f_1^* - K_p)^2 + p^2(f_2^* - K_p)^2$ is the total genetic variance of the reference model. The prevalence is K_p for all ε, and $1 + \varepsilon^2$ can be interpreted as the recurrence risk ratio $P(Y_{i_2} = 1|Y_{i_1} = 1)/P(Y_{i_2} = 1)$ of a monozygotic twin pair i_1i_2 (Risch 1990a). The local score function follows immediately by applying Corollary 2 with $\tilde{S}_1$ and $\tilde{S}_2$ as in (31) and (32) and $\tilde{\sigma}_a^2$ and $\tilde{\sigma}_d^2$ as the additive and dominant genetic variances of the reference model. For outbred pedigrees, $\rho = 2$, and the optimal score function $S = 2(\tilde{S}_2 - E_0(\tilde{S}_2))$ can be written in the form of (42), with $r_i = Y_i - K_p$ and $c = \tilde{\sigma}_d^2/\tilde{\sigma}_g^2$ the fraction of total genetic variance due to dominance effects for the reference model. For pedigrees with $\mathcal{P}_{\text{known}}$ containing only affected individuals, the local optimality of $\tilde{S}_1$ and $\tilde{S}_2$ has been derived by McPeek (1999).

7. RARE DISEASES

In Section 6 I kept the disease allele frequency fixed while varying the penentrance parameters $\boldsymbol{\psi}$. I now do the opposite,—keep $\boldsymbol{\psi}$ fixed and let $p \to 0$. This is of interest for rare diseases, with at most one disease allele present in the pedigree with high probability.

Put

$$\bar{S}_1(\mathbf{v}) = \sum_{j=1}^{2f} \prod_{i \in \mathcal{R}_j(\mathbf{v})} \frac{P(\mathbf{Y}_i|(10))}{P(\mathbf{Y}_i|(00))} \prod_{i \in \bar{\mathcal{R}}_j(\mathbf{v})} \frac{P(\mathbf{Y}_i|(11))}{P(\mathbf{Y}_i|(00))}, \qquad (50)$$

with $\mathcal{R}_j(\mathbf{v})$ and $\bar{\mathcal{R}}_j(\mathbf{v})$ as defined in (17). The jth term of (50) measures the relative risk if the jth founder allele is changed from 0 to 1, while all other founder alleles equal 0. Note that no (11) genotypes occur in $\mathcal{R}_j$, because it is very unlikely that two founders have a disease allele when p is small.

The following proposition states that a centered version of $\bar{S}_1$ is a pointwise optimal score function as $p \to 0$.

Proposition 3. As $p \to 0$,

$$P_p(\mathbf{v}|\mathbf{Y}) = 2^{-m}\big(1 + p\big(\bar{S}_1(\mathbf{v}) - E_0(\bar{S}_1)\big) + o(p)\big), \qquad (51)$$

with $\bar{S}_1$ as defined in (50) and E_0 denoting expectation with respect to the uniform inheritance distribution (6).

Example 10 (Binary phenotypes, continued). Consider the binary model (29). For an outpred pedigree, (50) becomes

$$\bar{S}_1(\mathbf{v}) = \sum_{j=1}^{2f} \left(\frac{1-f_1}{1-f_0}\right)^{|\mathcal{R}_j \cap \mathcal{P}_0|} \left(\frac{f_1}{f_0}\right)^{|\mathcal{R}_j \cap \mathcal{P}_1|}.$$

For a pedigree with $\mathcal{P}_{\text{known}}$ consisting of affecteds only ($\mathcal{P}_1 = \mathcal{P}_{\text{known}}$), $\bar{S}_1$ further reduces to

$$S_{\text{robdom}}(\mathbf{v}) = \sum_{j=1}^{2f} \left(\frac{f_1}{f_0}\right)^{|\mathcal{R}_j|}. \qquad (52)$$

This robust dominant score function was introduced by McPeek (1999), who found that $f_1/f_0 = 7$ gave high power in simulations for a variety of additive and dominant models with different disease allele frequencies and phenocopy rates f_0.

Example 11 (Gaussian phenotypes, continued). Consider now the Gaussian phenotypes from Example 3. For simplicity, I set all regression coefficients β_l to 0. Then $P(Y_i|(10))/P(Y_i|(00)) = K^{r_i - \log(K)/2}$, where $r_i = (Y_i - m_0)/\sigma$ is a standardized version of Y_i and $K = \exp((m_1 - m_0)/\sigma)$ measures the strength of the genetic component. In fact, the heritability is $2(\log(K))^2 p + o(p)$ for small p. For an outbred pedigree, $\bar{S}_1(\mathbf{v})$ in (50) reduces to

$$S_{\text{normdom}}(\mathbf{v}) := \sum_{j=1}^{2f} K^{\sum_{i \in \mathcal{R}_j}(r_i - \log(K)/2)}, \qquad (53)$$

which is a dominance score function for normal phenotypes and rare disease alleles.

For recessive binary models ($f_0 = f_1$) and inbred pedigrees, only the second product in each term of (50) is included. This indicates that the inbred individuals ($= \bar{\mathcal{R}}$) are most important in this case. In fact, Feingold and Siegmund (1997) reported that inbred individuals are more powerful in detecting linkage than sib pairs for recessive traits and very rare disease alleles. On the other hand, $\bar{S}_1 \equiv 0$ for recessive traits and outbred pedigrees. Then higher-order terms are needed in the expansion (51).

8. A SEMIPARAMETRIC WAY TO CHOOSE SCORE FUNCTIONS

I now describe a general strategy for doing linkage analysis. Given a genetic model, with arbitrary phenotypes and covariates, I compute the locally optimal score function according to Corollaries 1 and 2 or Proposition 3 and then use it for linkage analysis as described in Section 2. The score function obtained ususally contains one or several unknown parameters, which must be chosen in some way from existing/previous data or other experience. Therefore, the proposed procedure is semiparametric. Existing software for multipoint nonparametric linkage, such as the Genehunter program, can be easily used in this framework. Only the appropriate score function S and family weights γ_i have to be added to the program [cf. (3)]. The semiparametric procedure can be described in more detail as follows:

1. Choose a parametric model.
2. Derive a locally optimal likelihood score function S from Corollary 1 by letting $\varepsilon \to 0$ [$S = S_1 - E_0(S_1)$ or $S = S_2 - E_0(S_2)$, depending on whether the pedigree is inbred or not], or Proposition 3 by letting $p \to 0$ [$S = \bar{S}_1 - E_0(\bar{S}_1)$]. The locally optimal score functions $\tilde{S}_1$ and $\tilde{S}_2$ are equivalent to S_1 and S_2. This is because the linear and exponential expansions are equivalent as $\varepsilon \to 0$.
3. Normalize the score function to have 0 mean and unit variance under H_0, that is, $S \leftarrow S/\sqrt{I}$, where $I = \text{var}_0(S)$.
4. Compute the family score $Z(t)$ at all loci t of interest according to (3).
5. Based on previous experience or data, choose any unknown quantities remaining in the score function. (The number of these quantities is often much smaller than the original number of genetic model parameters.)
6. Repeat steps 1–6 for all pedigrees and define a total linkage score function according to (1), with weights $\gamma_i \propto \sqrt{I_i}$, I_i representing the value of I in step 3 for the ith pedigree.

The quantity I in step 3 can be interpreted as a Fisher information. For instance, for weak penetrances, I introduced the likelihood function $L(\varepsilon; \mathbf{v})$ in Section 6.1. Then $I = E_0(S^2(\mathbf{v}))$

is the second moment of the likelihood score function S evaluated at $\varepsilon = 0$. This gives the usual definition of Fisher information when $\rho = 1$ and a generalized definition when $\rho > 1$. A reparameterization $\varepsilon \leftarrow \varepsilon^{\rho}/\rho!$ gives the usual definition of Fisher information for all ρ. The choice of weights $\gamma_i \propto \sqrt{I_i}$ in step 6 corresponds to normalizing the $\varepsilon = 0$ score function of the total likelihood, $\prod_{i=1}^{N} L_i(\varepsilon)$, by the square root of the total Fisher information, $\sum_{i=1}^{N} I_i$. This is optimal in a pointwise testing (McPeek 1999) or estimation (Hössjer 2003a) sense. Note that unknown multiplicative constants appearing in S are authomatically removed by the normalization in step 3 and the normalization of weights in (1).

Example 12 (Local specificity models). I have shown that the likelihood score function (42) is locally optimal in a weak penetrance sense for Gaussian, binary and survival analysis models, provided that the pedigree is outbred. The unknown parameters of $S = S_2 - E_0(S_2)$ are the fraction c of genetic variance due to dominance effects and possible parameters of the residuals r_i. It generally involves little loss of information to put $c = 0$, especially when p is small. The corresponding (noncentered) score function $S_2/2$,

$$S_{\text{WPC}} = \sum_{i_1 < i_2} r_{i_1} r_{i_2} \text{IBD}_{i_1 i_2}, \tag{54}$$

is the weighted pairwise correlation (WPC) statistic previously derived by Commenges (1994) using different criteria. For binary models, when unaffecteds are included in $\mathcal{P}_{\text{known}}$, S_{WPC} contains the prevalence K_p as unknown parameter. It can be easily estimated from population data. If $\mathcal{P}_{\text{known}}$ contains only affecteds, then $(1 - K_p)$ enters as a multiplicative constant of S_{WPC} and can be removed. The resulting equivalent score function is S_{pairs}. For Gaussian models, r_i contains a multiplicative constant, σ^{-1}, that can be removed. The only essential parameters remaining in r_i are the "mean genotype effect" m and regression parameters $\boldsymbol{\beta}$, which can be estimated from population data. Note that each term of (42) has the form $r_{i_1} r_{i_2}(\text{IBD}_{i_1 i_2} - E_0(\text{IBD}_{i_1 i_2}))$ when $c = 0$. Risch and Zhang (1995, 1996) found that extremely concordant ($r_{i_1} r_{i_2}$ large) and discordant ($-r_{i_1} r_{i_2}$ large) sib pairs were most informative for linkage for Gaussian models. This is intitively plasible, because concordant (discordant) pairs of relatives on average have a larger (lower) degree of allele sharing under H_1 than expected under H_0. Hence they contribute positively to the overall linkage score around the disease locus under H_1, which increases the power. In survival analysis, r_i contains the Cox regression parameters $(\beta_2, \ldots, \beta_p)$ and the baseline hazard.

Example 13 (Rare disease alleles). In Example 10, (the centered version of) S_{robdom} gave the locally optimal score functions for binary phenotypes with affecteds only. It includes one unknown quantity, f_1/f_0, to be specified by the user.

The Gaussian phenotype generalization S_{normdom} in Example 11 contains K, m_0, and σ as the unknown quantities. Here K is the important design parameter, reflecting the strength of the genetic component. For rare diseases, m_0 and σ^2 can be well approximated by $E(Y_i)$ and the total variance, $\text{var}(Y_i) = \sigma_g^2 + \sigma^2 =: \sigma_t^2$. Both of these can easily be estimated from population data.

9. A SIMULATION STUDY

Here I present a simulation study for the Gaussian model. I first define some additional concepts. The noncentrality parameter (NCP) is the expected value (conditional on all phenotypes), $E(Z(\tau))$, of the linkage score (1) at the disease locus. It is closely related to the power for detecting linkage, especially when the power is large (see Prop. 2 and Feingold et. al. 1993; Sham, Zhao, and Curtis 1997; Nilsson 1999). But NCP is analytically more convenient, and it does not require specification of significance level and threshold. A noncentrality parameter of about 4 corresponds to significant linkage for a genomewide scan with perfect marker data (Lander and Kruglyak 1995), although the exact value depends on the set of pedigrees and phenotypes, the score function, and the genetic model (Änquist and Hössjer 2003; Hössjer 2003b). Under perfect marker information,

$$\text{NCP} = E(S) = \sum_{w} S(\mathbf{w}) P(\mathbf{v} = \mathbf{w}|\mathbf{Y}) \tag{55}$$

for one pedigree, provided that the score function S has been standardized to have mean 0 and variance 1 under H_0, as in step 3 of the preceding section. For a sample of N pedigrees,

$$\text{NCP} = \sqrt{N} \cdot \frac{\sum_{i=1}^{N} \gamma_i \text{NCP}_i / N}{\sqrt{\sum_{i=1}^{N} \gamma_i^2 / N}} \tag{56}$$

under perfect marker information, with NCP_i the noncentrality parameter of the ith pedigree.

Assume that pedigrees are sampled from a large population according to a probability distribution $dP(\mathbf{Y})$. (In general, $\mathbf{Y}$ includes information about which individuals that have known phenotypes, the phenotype values for these *and* the pedigree structure. For simplicity, I retain the notation $\mathbf{Y}$, see Hössjer 2003a for a more general notation.) Then the second factor of (56) converges to the asymptotic noncentrality parameter, (ANCP),

$$\text{ANCP} = \frac{\int \gamma(\mathbf{Y}) \text{NCP}(\mathbf{Y})\, dP(\mathbf{Y})}{\sqrt{\int \gamma^2(\mathbf{Y})\, dP(\mathbf{Y})}}, \tag{57}$$

as $N \to \infty$. (The square of ANCP is called the "asymptotic signal-to-noise ratio" in Hössjer 2003a.) The sampling measure P could be defined through some ascertainment scheme based on probands. Another possiblity is using random sampling, $P = P_{\text{rnd}}$. This means that pedigree structures are sampled according to some distribution. Then, conditional on pedigree structure, the subset of individuals with known phenotypes is sampled according to a distribution specific for that pedigree structure. Finally, conditional on pedigree structure and the subset of individuals with known phenotype, $\mathbf{Y}$ is sampled according to $\sum P(\mathbf{G}) P(\mathbf{Y}|\mathbf{G})$, that is, by combining (8) with the denominator of (7). I consider a sampling scheme where a fraction, α $(0 < \alpha \leq 1)$ of the most informative and randomly sampled pedigrees is retained according to

$$dP(\mathbf{Y}) \propto dP_{\text{rnd}}(\mathbf{Y}) \mathcal{N}_{\{\gamma(\mathbf{Y}) \geq c\}}, \tag{58}$$

where c is chosen so that $\int_{\gamma(\mathbf{Y}) \geq c} dP_{\text{rnd}}(\mathbf{Y}) = \alpha$ and the proportionality constant is chosen so that P becomes a probability measure. This means that the fraction α of pedigrees with largest weights is retained. If $\alpha = 1$, then P is identical to random sampling.

Four pedigree structures were used: a sib pair (SP), sib trio (Strio), sib quartet (Squart), and first-cousin pedigree (Cous) with eight pedigree members; two grandparents, their two children with spouses, and two first cousins in the third generation. For all four pedigrees, the two members of the first generation had unknown phenotypes.

I did not include covariates in the Gaussian model, and I assumed, for simplicity, that $m = E(Y_i)$ and $\sigma_t^2 = \text{var}(Y_i)$ had been estimated from population data. Because the ANCP is invariant with respect to location and scale transformations of the phenotypes, the four penetrance parameters $(m_0, m_1, m_2, \text{and } \sigma^2)$ can, without loss of generality, be reduced to two,

$$\text{Disp} = (m_2 - m_0)/\sigma$$

and

$$\text{Dom} = (2m_1 - m_0 - m_2)/(m_2 - m_0).$$

The displacement, Disp, quantifies the strength, and Dom quantifies the degree of dominance of the genetic component. Under the mild restriction that m_i's are nondecreasing, Disp ≥ 0 and $-1 \leq \text{Dom} \leq 1$, with Dom taking values -1, for recessive models, 0 for additive models, and 1 for dominant models.

Figures 1–5 show the results of the simulation study. In these figures, plots of $50 \cdot$ ANCP correspond to a noncentrality parameter for a sample of 2,500 pedigrees. However, the reader can easily rescale to interpret the results for other sample sizes.

Four score functions were used: S_{WPC}; S_{normdom} with $K = 1.5$; the Haseman–Elston score function,

$$S_{\text{HE}}(\mathbf{v}) = \sum_{i_1 < i_2} \left(2\sigma_t^2 - (Y_{i_1} - Y_{i_2})^2\right)\text{IBD}_{i_1 i_2}; \tag{59}$$

and the optimal score function $S(\mathbf{v}) = P(\mathbf{v}|\mathbf{Y})$. Phenotypes were centered, $\mathbf{Y}_i \leftarrow \mathbf{Y}_i - \mathbf{m}$, for S_{WPC} and standardized, $\mathbf{Y}_i \leftarrow (\mathbf{Y}_i - \mathbf{m})/\sigma_t$, for S_{normdom}. S_{HE} is the score function analog of the classical linkage method for QTL mapping due to Haseman and Elston (1972), where squared trait differences $(Y_{i1} - Y_{i2})^2$ are regressed against $\text{IBD}_{i_1 i_2}$. The constant $2\sigma_t^2$ can be removed in (59) without affecting the ANCP. Then S_{HE} becomes a score test for the regression slope being 0. However, the finite-sample behavior of NCP is often a bit better when the $2\sigma_t^2$ term is included. The optimal score function S_{optimal} maximizes (55) with respect to S subject to the constraints of 0 mean and unit variance under H_0 (Hössjer 2003a). It is ideal (parametric) in the sense that it requires knowledge of the genetic model, and is a good benchmark to use for comparisons with other score functions.

The two winners are S_{WPC} and S_{normdom}. These two score functions have fairly good performance for a wide range of models. S_{WPC} is better than S_{normdom} for large p, whereas the opposite is true for small p and small sampling fractions α. S_{HE} has comparable or better performance than the other two score functions only for strong genetic models and large p. I also investigated the version of S_{WPC} with optimal $c = \sigma_d^2/\sigma_g^2$ instead of $c = 0$. This score function never had more than a few percent higher NCP than S_{WPC}, even for

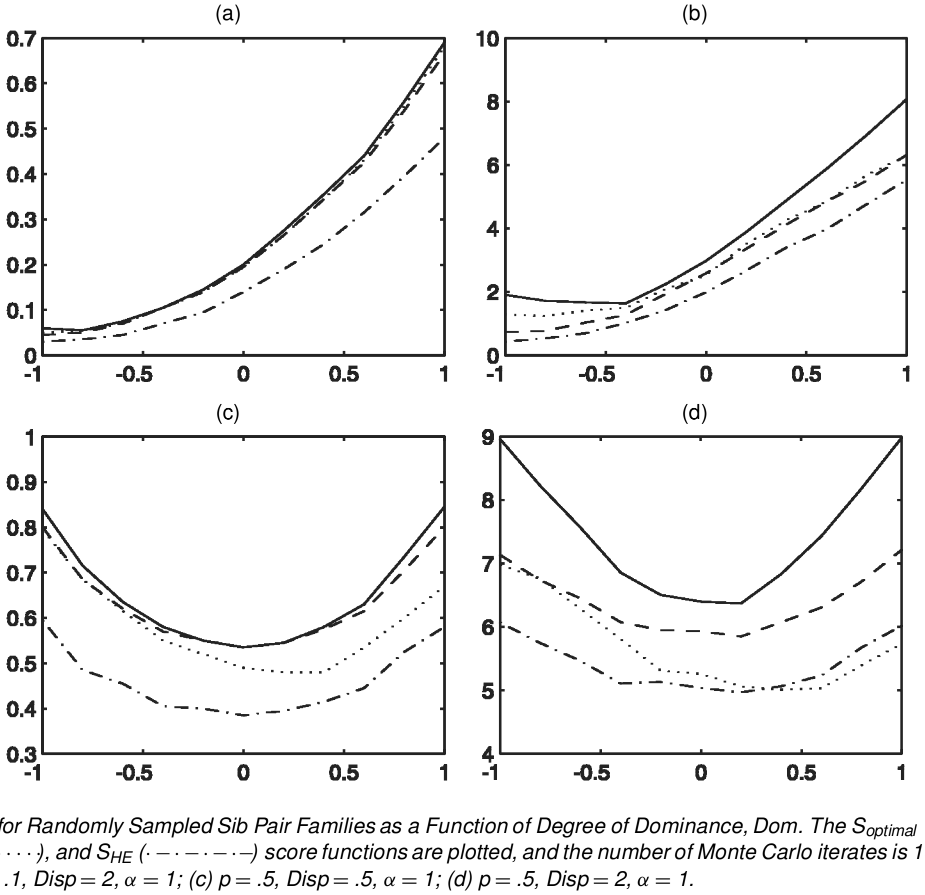

Figure 1. $50 \cdot$ ANCP for Randomly Sampled Sib Pair Families as a Function of Degree of Dominance, Dom. The $S_{optimal}$ (——), S_{WPC} (– – – –), $S_{normdom}$ with $K = 1.5$ (· · · · ·), and S_{HE} (· – · – · – · –) score functions are plotted, and the number of Monte Carlo iterates is 10,000 in (57). (a) $p = .1$, Disp $= .5$, $\alpha = 1$; (b) $p = .1$, Disp $= 2$, $\alpha = 1$; (c) $p = .5$, Disp $= .5$, $\alpha = 1$; (d) $p = .5$, Disp $= 2$, $\alpha = 1$.

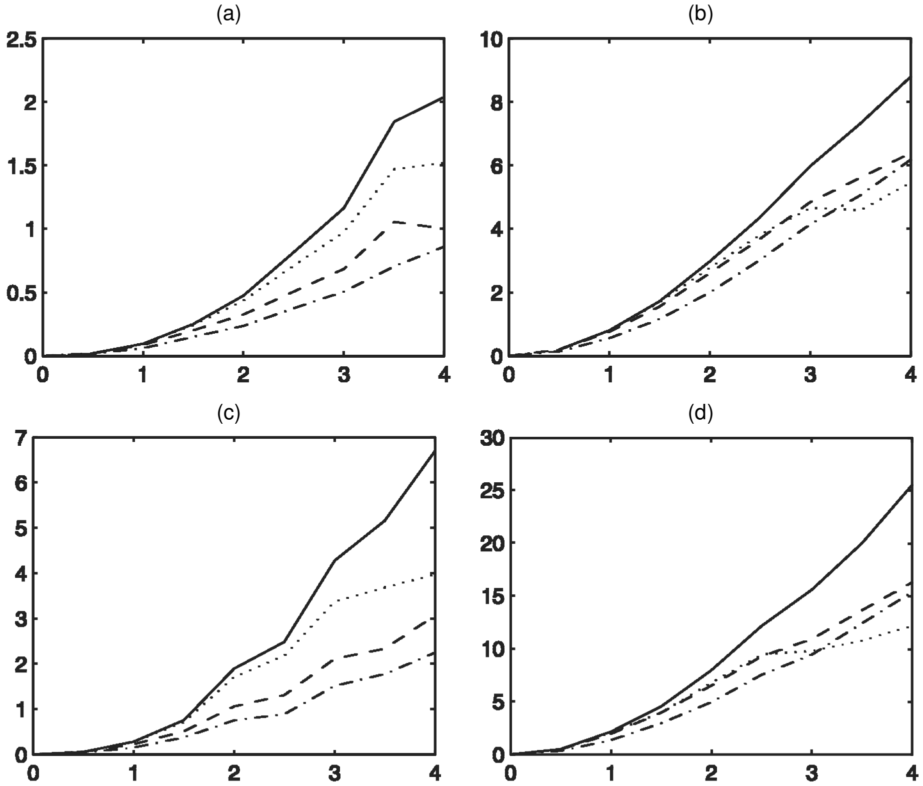

Figure 2. 50 · ANCP for Randomly Sampled Sib Pair and Sib Quartet Families as a Function of Displacement Disp. The $S_{optimal}$ (——), S_{WPC} (– – – –), $S_{normdom}$ with K = 1.5 (· · · · ·), and S_{HE} (· – · – · – ·–) score functions are plotted, and the number of Monte Carlo iterates in (57) is 10,000 (a and b) and 2,000 (c and d). (a) p = .01, Dom = 0, α = 1, SP; (b) p = .1, Dom = 0, α = 1, SP; (c) p = .01, Dom = 0, α = 1, Squart; (d) p = .1, Dom = 0, α = 1, Squart.

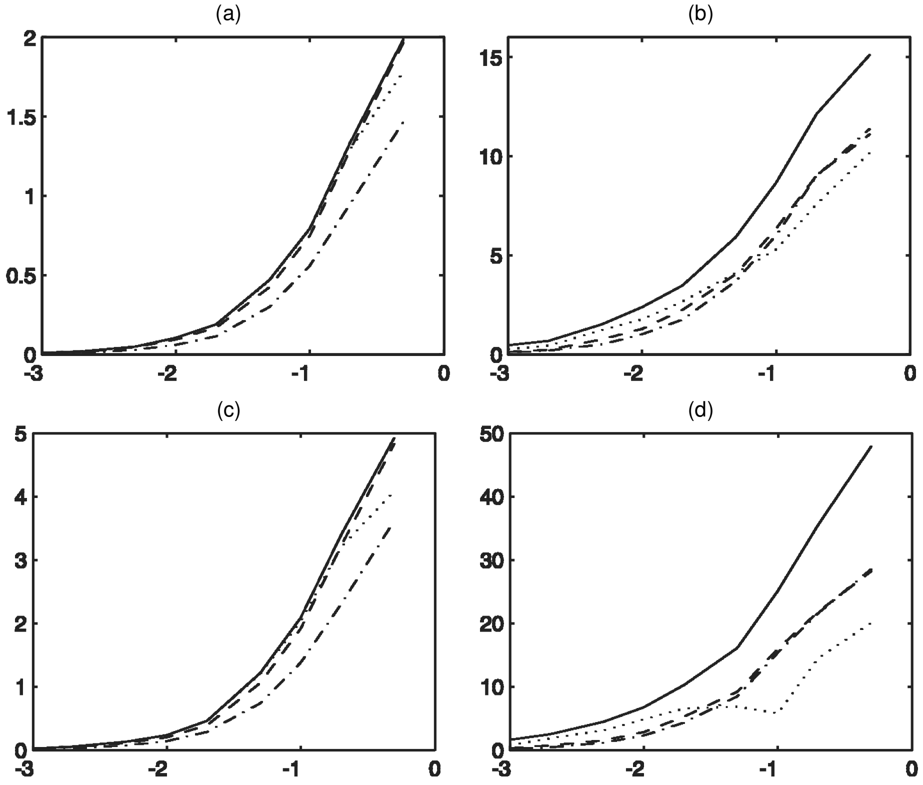

Figure 3. 50 · ANCP for Randomly Sampled Sib Pair and Sib Quartet Families as a Function of Tenth Logarithm of Disease Allele Frequency $\log_{10}(p)$. The $S_{optimal}$ (——), S_{WPC} (– – – –), $S_{normdom}$ with K = 1.5 (· · · · ·), and S_{HE} (· – · – · – ·–) score functions are plotted, and the number of Monte Carlo iterates in (57) is 10,000 (a and b) and 2,000 (c and d). (a) Disp = 1, Dom = 0, α = 1, SP; (b) Disp = 4, Dom = 0, α = 1, SP; (c) Disp = 1, Dom = 0, α = 1, Squart; (d) Disp = 4, Dom = 0, α = 1, Squart.

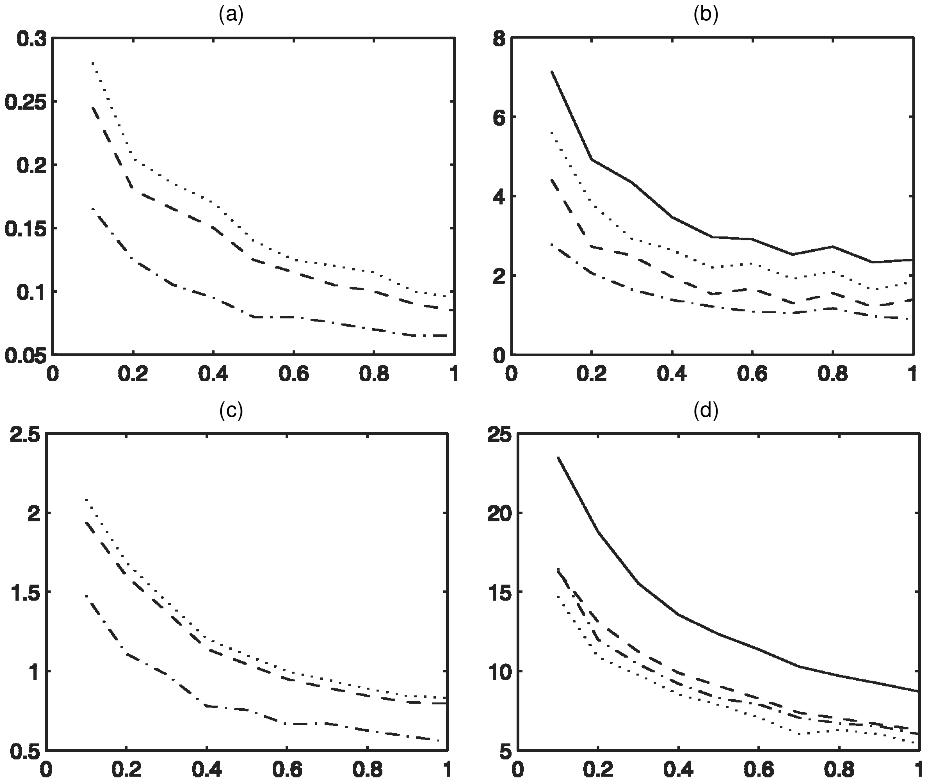

Figure 4. 50 · ANCP for Sib Pair Families as a Function of Sampling Fraction α. The $S_{optimal}$ (——), S_{WPC} (— — — —), $S_{normdom}$ with $K = 1.5$ (· · · · ·), and S_{HE} (· — · — · —) score functions are plotted. The number of Monte Carlo iterates in (57) is 10,000, of which the fraction α most informative ones are retained. In (a) and (c) the solid line is omitted, because it essentially coincides with the dotted line. (a) p = .01, Disp = 1, Dom = 0; (b) p = .01, Disp = 4, Dom = 0; (c) p = .1, Disp = 1, Dom = 0; (d) p = .1, Disp = 4, Dom = 0.

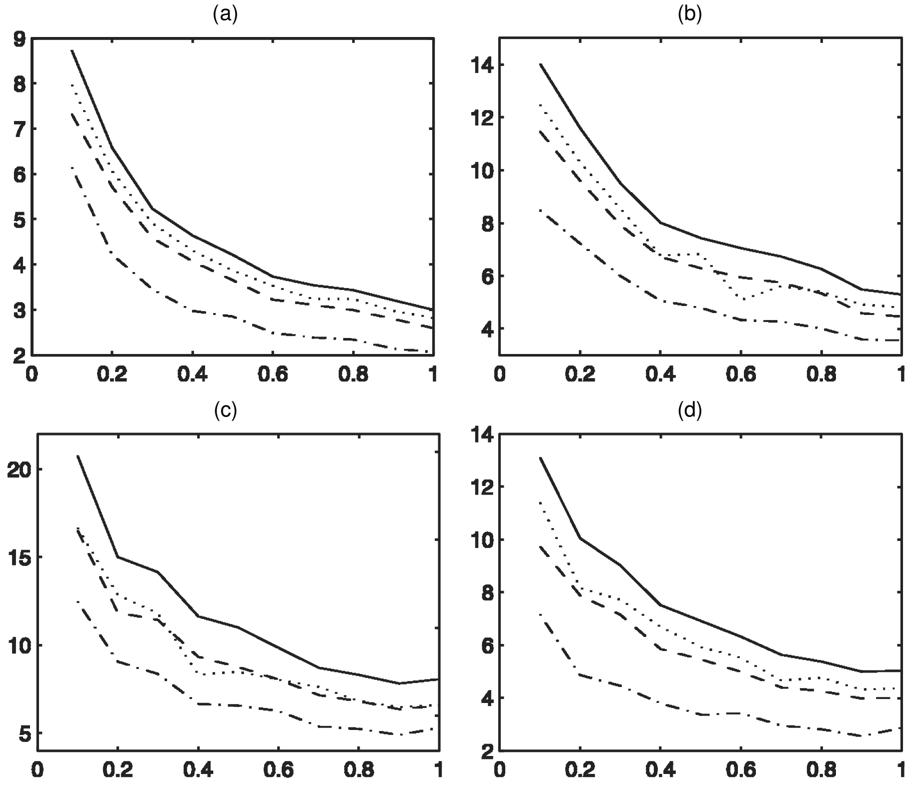

Figure 5. 50 · ANCP for Sib Pair, Sib Trio, Sib Quartet, and First-Cousin Families as a Function of Sampling Fraction α. The $S_{optimal}$ (——), S_{WPC} (— — — —), $S_{normdom}$ with $K = 1.5$ (· · · · ·), and S_{HE} (· — · — · —) score functions are plotted. The number of Monte Carlo iterates is 10,000 (a and d), 5,000 (b), and 2,000 (c), of which the fraction α most informative ones are retained. (a) p = .1, Disp = 2, Dom = 0, SP; (b) p = .1, Disp = 2, Dom = 0, Strio; (c) p = .1, Disp = 2, Dom = 0, Squart; (d) p = .1, Disp = 2, Dom = 0, Cous.

purely recessive (Dom $= -1$) or dominant (Dom $= 1$) models. I also tried S_{normdom} with other values of K in the range of 2 to 10. As a rule of thumb, larger values of K gave much less robust prestanda. Especially for large p and strong genetic models, the performance of S_{normdom} drastically decreased with increasing K. The same was true for the optimal version of S_{normdom} with $K = \exp((m_1 - m_0)/\sigma)$ and standardization $Y_i \leftarrow (Y_i - m_0)/\sigma$. It was best only among the S_{normdom} score functions for very small disease allele frequencies.

10. OUTLOOK

10.1 Oligogenic, Polygenic, and Environmental Effects

In (8) I assumed conditionally independent phenotypes given the genotypes. This assumption is not appropriate if shared environmental effects or trait loci from other chromosomes are contributing to the disease.

Two-locus IBD probabilities for affected sib pairs have been considered by other authors. Cordell, Todd, Bennett, Kawagucki, and Farrall (1995) used variance components techniques of James (1971) and Risch (1990a), whereas Dudoit and Speed (1999) and Bengtsson (2001) used methods that couple monotonicity of penetrances and IBD probabilities. Further results for additive models of multilocus penetrance have been derived by Feingold and Siegmund (1997) and Teng and Siegmund (1997).

It is possible to derive $P(\mathbf{v}|\mathbf{Y}) \propto \sum_{\mathbf{v}'} P(\mathbf{Y}|\mathbf{v}, \mathbf{v}')$ for oligogenic models with two major unlinked genes. Here $\mathbf{v}$ and $\mathbf{v}'$ are the inheritance vectors at the two disease loci and the contribution from $\mathbf{v}'$ is summed, because markers from the first chromosome give no information about $\mathbf{v}'$ (see Hössjer 2001 for more details).

Another approach is to retain just one inheritance vector v and build in conditional dependence into $P(\mathbf{Y}|\mathbf{G})$ via some given penetrance function that includes polygenic and/or shared environmental effects. Note that this penetrance function is defined for $\mathbf{Y}|\mathbf{G}$ rather than the components $\mathbf{Y}_i|\mathbf{G}_i$ (see Hössjer 2003c for details).

Tang and Siegmund (2001), Putter, Sandkuijl, and van Houwelingen (2002), and Wang and Huang (2002) recently defined score functions for quantitative traits. These authors incorporate polygenic and shared environmental effects by modeling $\mathbf{Y}|\mathbf{v}$ as a multivariate normal vector with dependent components due to effects from polygenes and the gene at the main locus.

10.2 Nonparametric Approach

When little is known about the genetic model, it is plausible that power can be increased by maximizing the local score functions over several parameters simultaneously with the linkage analysis. This possiblity was mentioned by Whittemore (1996). Let $S(\mathbf{v}; \theta^*)$ represent a local likelihood score function derived from a trajectory θ^*, with $Z(t; \theta^*)$ the corresponding total linkage score (1). Then

$$\sup_{\theta^*} Z(t; \theta^*) \qquad (60)$$

is a nonparametric alternative to the semiparametric procedure of Section 8, where the trajectory was kept fixed. Of course, the increased number of degrees of freedom must be adjusted for when computing p values for the maximal linkage score. For the Gaussian model, the score function S in (42) could be used. The trajectory θ^* depends on one parameter c, which can be maximized over. However, because $c = 0$ is often a good approximation, it is not certain whether the improved model fitting justifies the increased degrees of freedom.

10.3 Other Extensions

It is possible to generalize the genetic model in various other ways. If the assumption of random mating is dropped, Proposition 1 remains still valid, although the variables $\{\xi_i\}$ are no longer independent. This will affect the local penetrance expansion of Corollary 1 for outbred pedigrees. Multiallelic models can be treated in principle, using U statistics theory to extend Lemma A.1 in Appendix A. Finally, various kinds of environmental parameters (such as the regression coefficents of Example 7) can be easily incorporated into the general framework.

APPENDIX A: DECOMPOSITION OF BIVARIATE FUNCTIONS

Here I demonstrate how to expand a bivariate function $h(a_j, a_k)$ as a linear combination of the random variables $\{\xi_j\}$ and $\{\xi_{jk}\}$ introduced in (11) and (12). It is convenient to introduce two bivariate forms, $\langle\cdot,\cdot\rangle_l : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$, $l = 1, 2$, according to

$$\langle \mathbf{x}, \mathbf{y}\rangle_1 = q^2 x_1 y_1 + 2pq x_2 y_2 + p^2 x_3 y_3$$

and

$$\langle \mathbf{x}, \mathbf{y}\rangle_2 = q x_1 y_1 + p x_3 y_3,$$

where $\langle\cdot,\cdot\rangle_1$ introduces a norm and $\langle\cdot,\cdot\rangle_2$ a seminorm on $\mathbb{R}^3$. Then it can be seen that

$$\mathbf{e}_1 = (1, 1, 1),$$

$$\mathbf{e}_2 = \frac{1}{\sqrt{2pq}}(-2p, q - p, 2q),$$

and

$$\mathbf{e}_3 = (1/q - 1, -1, 1/p - 1),$$

and

$$\bar{\mathbf{e}}_1 = (1, 0, 1),$$

and

$$\bar{\mathbf{e}}_2 = \left(-\sqrt{p/q}, 0, \sqrt{q/p}\right)$$

are mutually orthogonal unit vectors with respect to $\langle\cdot,\cdot\rangle_1$ and $\langle\cdot,\cdot\rangle_2$. I can now formulate the following result.

Lemma 1. Consider a bivariate function $h : \{0, 1\} \times \{0, 1\} \to \mathbb{R}$, which is symmetric ($h(0, 1) = h(1, 0)$). Then

$$h(a_j, a_k) = \begin{cases} \langle \mathbf{h}, \mathbf{e}_1\rangle_1 + \frac{1}{\sqrt{2}}\langle \mathbf{h}, \mathbf{e}_2\rangle_1(\xi_j + \xi_k) & \\ \quad + \langle \mathbf{h}, \mathbf{e}_3\rangle_1 \xi_{jk} & \text{if } j \neq k \\ \langle \mathbf{h}, \bar{\mathbf{e}}_1\rangle_2 + \langle \mathbf{h}, \bar{\mathbf{e}}_2\rangle_2 \xi_j & \text{if } j = k, \end{cases}$$

where $\mathbf{h} = (h(0, 0), h(0, 1), h(1, 1))$.

Proof. The result can be derived from U statistics theory (see, e.g., Lee 1990), but I sketch a direct elementary proof here in the case where $j \neq k$ (the case $j = k$ is similar). Define

$$r(a_j, a_k) = h(a_j, a_k) - \left(\langle \mathbf{h}, \mathbf{e}_1\rangle_1 + \frac{1}{\sqrt{2}}\langle \mathbf{h}, \mathbf{e}_2\rangle_1(\xi_j + \xi_k) + \langle \mathbf{h}, \mathbf{e}_3\rangle_1 \xi_{jk}\right)$$

and $\mathbf{r} = (r(0,0), r(0,1), r(1,1))$. Then, it follows from (13) that

$$E(r(a_j, a_k)) = 0 \iff \langle \mathbf{r}, \mathbf{e}_1 \rangle_1 = 0,$$
$$E(\xi_j r(a_j, a_k)) = 0 \iff \langle \mathbf{r}, \mathbf{e}_2 \rangle_1 = 0,$$

and

$$E(\xi_{jk} r(a_j, a_k)) = 0 \iff \langle \mathbf{r}, \mathbf{e}_3 \rangle_1 = 0.$$

Because $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ are linearly independent, $\mathbf{r} = \mathbf{0}$, and this proves the assertion.

APPENDIX B: PROOF OF PROPOSITION 1

It is convenient to introduce $h_i(0,0) = \log P(Y_i|(00))$, $h_i(1,0) = h_i(0,1) = \log P(Y_i|(10))$, and $h_i(1,1) = \log P(Y_i|(11))$. Then write $\log P(Y_i|\mathbf{G}_i) = h(a_{j_i}, a_{k_i})$. This representation uses the fact that given $\mathbf{v}$, the logarithm of the penetrance factor is a function of two founder alleles, a_{j_i} and a_{k_i}. Let $\mu_i = q^2 h_i(0,0) + 2pq h_i(0,1) + p^2 h_i(1,1)$ and $\mu_{li} = q h_i(0,0) + p h_i(1,1)$ be the mean penetrance effects for an individual who (given v) has alleles IBD or not. Note that κ_{ai}, κ_{di}, and κ_{li} in (14) can be written in terms of $h_i(\cdot,\cdot)$ as well. Further, $h_i(\cdot,\cdot) \equiv 0$ for pedigree members with unknown phenotypes.

It follows from (8) and Lemma A.1 in Appendix A that

$$\begin{aligned} P(\mathbf{Y}|\mathbf{v}) &= E\left(\exp\left(\sum_{i=1}^{n} h_i\big(a_{j_i}, a_{k_i}\big)\right)\right) \\ &= E\left(\exp\left(\sum_{i\in\mathcal{R}} \big(\mu_i + \kappa_{ai}\big(\xi_{j_i} + \xi_{k_i}\big) + \kappa_{di}\xi_{j_i k_i}\big) \right.\right. \\ &\qquad \left.\left. + \sum_{i\in\bar{\mathcal{R}}} \big(\mu_{li} + \kappa_{li}\xi_{j_i}\big)\right)\right) \\ &\propto E\left(\exp\left(\sum_{i\in\mathcal{R}} \big(\kappa_{ai}\big(\xi_{j_i} + \xi_{k_i}\big) + \kappa_{di}\xi_{j_i k_i}\big) \right.\right. \\ &\qquad \left.\left. + \sum_{i\in\bar{\mathcal{R}}} \big(\kappa_{di} + \kappa_{li}\xi_{j_i}\big)\right)\right). \end{aligned} \tag{B.1}$$

I divided by $\exp(\sum_{i\in\mathcal{P}_{\text{known}}} \mu_i)$, which is independent of v, in the last step of (B.1), and also used the fact that $\mu_{li} - \mu_i = \kappa_{di}$. The proposition now follows by combining (7) and (B.1) with the definition of $M_\xi(B)$ and rearranging terms in (B.1).

APPENDIX C: PROOF OF PROPOSITION 2

Introduce the bivariate function $\tilde{h}_i(0,0) = P(Y_i|(00))/\tilde{\mu}_i - 1$, $\tilde{h}_i(0,1) = \tilde{h}_i(1,0) = P(Y_i|(10))/\tilde{\mu}_i - 1$, and $\tilde{h}_i(1,1) = P(Y_i|(11))/\tilde{\mu}_i - 1$. Further, let $\tilde{\mu}_{li} = qP(Y_i|(00)) + pP(Y_i|(11))$ be the average penetrance value of an individual homozygous by descent. Then it follows from (8), Lemma A.1, and the definitions of $\tilde{\mu}_i$, $\tilde{\kappa}_{di}$, $\tilde{\kappa}_{ai}$, and $\tilde{\kappa}_{li}$ that

$$\begin{aligned} P(\mathbf{Y}|\mathbf{v}) &= \prod_{i=1}^{n} \tilde{\mu}_i \cdot E\left(\prod_{i=1}^{n} \big(1 + \tilde{h}_i\big(a_{j_i}, a_{k_i}\big)\big)\right) \\ &\propto E\left(\prod_{i\in\mathcal{R}} \big(1 + \tilde{\kappa}_{ai}\big(\xi_{j_i} + \xi_{k_i}\big) + \tilde{\kappa}_{di}\xi_{j_i k_i}\big)\right. \\ &\qquad \left. \times \prod_{i\in\bar{\mathcal{R}}} \big(1 + \tilde{\kappa}_{di} + \tilde{\kappa}_{li}\xi_{j_i}\big)\right). \end{aligned} \tag{C.1}$$

The last step of (C.1) used the fact that $\tilde{\mu}_{li}/\tilde{\mu}_i - 1 = \tilde{\kappa}_{di}$.

Now (C.1) can be expanded as a sum of $4^{|\mathcal{R}|}3^{|\bar{\mathcal{R}}|}$ terms. For any nonempty $\mathcal{Q}$, let $T_\mathcal{Q}$ be the sum of those terms for which the nonunity factors $\tilde{\kappa}_{ai}\xi_{j_i}$, $\tilde{\kappa}_{ai}\xi_{k_i}$, $\tilde{\kappa}_{di}\xi_{j_i k_i}$, $\tilde{\kappa}_{di}$, and $\tilde{\kappa}_{li}\xi_{j_i}$ are taken precisely from the individuals in $\mathcal{Q}$. It is easy to see that this definition of $T_\mathcal{Q}$ coincides with (23), (24), and (26). The latter formula is given only for outbred pedigrees, and the exact definition of u_i is

$$u_i = \begin{cases} 0 & \text{if a grandpaternal factor } \tilde{\kappa}_{ai}\xi_{j_i} \text{ is picked} \\ 1 & \text{if a grandmaternal factor } \tilde{\kappa}_{ai}\xi_{k_i} \text{ is picked} \\ 2 & \text{if a joint factor } \tilde{\kappa}_{di}\xi_{j_i k_i} \text{ is picked.} \end{cases}$$

Note also that the first moments of ξ_j and ξ_{jk} are 0; this explains why only HBD individuals are included in (23).

By the definition of $T_\mathcal{Q}$, $P(\mathbf{Y}|\mathbf{v}) \propto 1 + \sum_\mathcal{Q} T_\mathcal{Q}$. In conjunction with (7) and (28), this proves (27).

APPENDIX D: SCORE FUNCTIONS FOR BINARY PHENOTYPES OF ORDER 1 AND 2

Here I derive the expressions (31) and (32) for $\tilde{S}_1$ and $\tilde{S}_2$. Define

$$\tilde{\kappa}_a = \sqrt{pq}\big(p(f_2 - f_1) + q(f_1 - f_0)\big)/K_p,$$
$$\tilde{\kappa}_d = pq(f_2 - 2f_1 + f_0)/K_p,$$

and

$$\tilde{\kappa}_l = \sqrt{pq}(f_2 - f_0)/K_p.$$

Then it follows from (22) that $\tilde{\kappa}_{ai} = \tilde{\kappa}_a$ when $i \in \mathcal{P}_1$ and $\tilde{\kappa}_{ai} = -R\tilde{\kappa}_a$ when $i \in \mathcal{P}_0$. Hence $\tilde{\kappa}_{ai}$ attains only two values in the pedigree. Similarly, $\tilde{\kappa}_{di}$ and $\tilde{\kappa}_{li}$ are the same as $\tilde{\kappa}_d$ and $\tilde{\kappa}_l$ when $i \in \mathcal{P}_1$ and differ by a factor $-R$ when $i \in \mathcal{P}_0$. From this, I immediately derive (31) from (23) and (28). Noticing that $\tilde{\kappa}_a^2 = \tilde{\sigma}_a^2/(2K_p^2)$ and $\tilde{\kappa}_d^2 = \tilde{\sigma}_d^2/K_p^2$, I also deduce (32) from (24) and (28).

APPENDIX E: PROOFS FROM SECTIONS 6 AND 7

Proof of Corollary 1

Let $B(\varepsilon)$ be the value of $\mathbf{B}$ in (16) corresponding to the penetrance vector $\boldsymbol{\psi}_\varepsilon$. Because the cumulants of $\{\xi_j, \xi_{jk}\}$ up to order 2 are given by (13), it follows that $\log(M(B(\varepsilon))) = .5\sum_{j\le k} B_{jk}(\varepsilon)^2 + o(\varepsilon^2)$. Differentiating (16) with respect to ε and using (35), note that

$$B'_{jk}(0) = \begin{cases} \displaystyle\sum_{i\in\mathcal{R}_{jk}} \kappa'_{di}(0), & j < k \\ \displaystyle\sum_{i\in\mathcal{R}_j} \kappa'_{ai}(0) + \sum_{i\in\bar{\mathcal{R}}_j} \kappa'_{li}(0), & j = k. \end{cases}$$

Interchanging summation with respect to (j,k) and (i_1, i_2), observe that S_2 in (39) can be written as $S_2(\mathbf{v}) = \sum_{i\in\bar{\mathcal{R}}} \kappa''_{di}(0) + \sum_{j\le k} B'_{jk}(0)^2$. This, in conjunction with (35) and Proposition 1, gives $P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y}) \propto \exp(\varepsilon S_1(\mathbf{v}) + \varepsilon^2 S_2(\mathbf{v})/2 + o(\varepsilon^2))$. I finally arrive at (38) by using $\sum_\mathbf{w} P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v} = \mathbf{w}|\mathbf{Y}) = 1$, which holds for each $\varepsilon \ge 0$.

Deriving Local Score Functions for Gaussian Phenotypes

Let $u_i = m_i^* - m$. Because $\log P(\mathbf{Y}_i|\mathbf{G}_i) = \text{constant} - .5(r_i - \varepsilon\mathbf{u}_{|\mathbf{G}_i|}/\sigma_g)^2$, the additive, dominant, and loop components become

$$\begin{aligned} \kappa_{ai}(\varepsilon) &= \varepsilon r_i \sqrt{pq}\big(p(m_2^* - m_1^*) + q(m_1^* - m_0^*)\big)/\sigma_g \\ &\quad - \varepsilon^2 \sqrt{pq}\big(p(u_2^2 - u_1^2) + q(u_1^2 - u_0^2)\big)/(2\sigma_g^2), \\ \kappa_{di}(\varepsilon) &= \varepsilon r_i pq(m_2^* - 2m_1^* + m_0^*)/\sigma_g \\ &\quad - \varepsilon^2 pq(u_2^2 - 2u_1^2 + u_0^2)/(2\sigma_g^2), \\ \kappa_{li}(\varepsilon) &= \varepsilon r_i \sqrt{pq}(m_2^* - m_0^*)/\sigma_g - \varepsilon^2\sqrt{pq}(u_2^2 - u_0^2)/(2\sigma_g^2), \end{aligned} \tag{E.1}$$

for $i \in \mathcal{P}_{\text{known}}$. Expanding (E.1) locally around $\varepsilon = 0$, (35) holds, with

$$\kappa'_{ai}(0) = r_i\sqrt{pq}\big(p(m_2^* - m_1^*) + q(m_1^* - m_0^*)\big)/\sigma_g,$$
$$\kappa'_{di}(0) = r_i pq(m_2^* - 2m_1^* + m_0^*)/\sigma_g,$$
$$\kappa''_{di}(0) = -pq(u_2^2 - 2u_1^2 + u_0^2)/\sigma_g^2,$$

and

$$\kappa'_{li}(0) = r_i\sqrt{pq}(m_2^* - m_0^*)/\sigma_g.$$

Inserting these expressions into (37) yields

$$\begin{aligned} T_{i_1i_2} &= r_{i_1}r_{i_2}\big((1-c)\cdot \text{IBD}_{i_1i_2}/2 + c\cdot \mathcal{N}_{\{\text{IBD}_{i_1i_2}=2\}}\big) \\ &= r_{i_1}r_{i_2}C_{i_1i_2}, \end{aligned} \tag{E.2}$$

when $\bar{\mathcal{R}} = \varnothing$ and $c = \sigma_d^2/\sigma_g^2$. Because I am assuming an outbred pedigree, the likelihood score function is $S = S_2 - E_0(S_2)$, with $S_2 = \sum_{i_1i_2} T_{i_1i_2}$ as defined in (39). The matrix $(T_{i_1i_2})$ symmetric, and its diagonal entries are independent of $\mathbf{v}$. Hence the diagonal can be absorbed into the centering constant of S_2, and $S = S_2 - E_0(S_2)$ can be written with S_2 as in (42).

Proof of Corollary 2

Equation (48) follows from Proposition 2 and the expansions in (47). Then, using $\sum_{\mathbf{w}} P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v} = \mathbf{w}|\mathbf{Y}) = 1$ gives

$$P_{\boldsymbol{\psi}_\varepsilon}(\mathbf{v}|\mathbf{Y}) = 2^{-m}\frac{1 + \sum_{l=1}^{|\mathcal{P}_{\text{known}}|}\varepsilon^l \tilde{S}_l(v)}{1 + \sum_{l=1}^{|\mathcal{P}_{\text{known}}|}\varepsilon^l E_0(\tilde{S}_l)},$$

and Taylor expansion of this formula up to order 2 gives (49).

Score Function in Example 8

I use (40) and (39) to derive the score function. By differentiating (18) and the analogous expression for κ_{di} with respect to ε, I obtain $\kappa'_{ai}(0) = \kappa_a^{*,II} r_i/\sigma_g$ and $\kappa'_{di}(0) = \kappa_d^{*,II} r_i/\sigma_g$, where $\kappa_a^{*,II} = \sqrt{pq}(p(\alpha_2^* - \alpha_1^*) + q(\alpha_1^* - \alpha_0^*))$ and $\kappa_d^{*,II} = pq(\alpha_2^* - 2\alpha_1^* + \alpha_0^*)$. Notice further that $\sigma_a^2 = 2(\kappa_a^{*,II})^2$ and $\sigma_d^2 = (\kappa_d^{*,II})^2$. This gives (42) on noticing that $\{C_{i_1i_2}\}$ is symmetric and that its diagonal elements are independent of $\mathbf{v}$.

Proof of Proposition 3

I apply (15). Note first that

$$\kappa_{di} = p\big(h_i(1,1) - 2h_i(0,1) + h_i(0,0)\big) + o(p) \tag{E.3}$$

as $p \to 0$, with $h(\cdot,\cdot)$ as defined in Appendix B. Thus $\kappa_{di}/\sqrt{p} \to 0$, whereas $\kappa_{ai}/\sqrt{p} \to h_i(0,1) - h_i(0,0)$ and $\kappa_{li}/\sqrt{p} \to h_i(1,1) - h_i(0,0)$ as $p \to 0$. Hence, by (16),

$$\begin{aligned} B_{jj}/\sqrt{p} \to &\sum_{i\in\mathcal{R}_j}\big(h_i(0,1) - h_i(0,0)\big) \\ &+ \sum_{i\in\bar{\mathcal{R}}_j}\big(h_i(1,1) - h_i(0,0)\big) =: s_j \end{aligned}$$

and $B_{jk}/\sqrt{p} \to 0$ when $j < k$. After some computations, it is seen that

$$M_\xi(\mathbf{B}) = \prod_{j=1}^{2f} M_{\xi_j}(B_{jj}) + o(p), \tag{E.4}$$

which follows from the independence of $\{\xi_j\}_{j=1}^{2f}$ and the fact that the nondiagonal entries of $\mathbf{B}$ can be ignored in the limit $p \to 0$. Now the definition of M_{ξ_j} after (19) implies that

$$\begin{aligned} M_{\xi_j}(B_{jj}) &= q\exp\left(-\frac{B_{jj}}{\sqrt{p}}\frac{p}{\sqrt{q}}\right) + p\exp\left(\frac{B_{jj}}{\sqrt{p}}\sqrt{q}\right) \\ &= 1 + p\big(\exp(s_j) - s_j - 1\big) + o(p). \end{aligned} \tag{E.5}$$

Observe that

$$\begin{aligned} \sum_{j=1}^{2f} s_j &= 2\sum_{i\in\mathcal{R}}\big(h_i(0,1) - h_i(0,0)\big) \\ &\quad + \sum_{i\in\bar{\mathcal{R}}}\big(h_i(1,1) - h_i(0,0)\big) \\ &= 2\sum_{i\in\mathcal{P}_{\text{known}}}\big(h_i(0,1) - h_i(0,0)\big) \\ &\quad + \sum_{i\in\bar{\mathcal{R}}}\big(h_i(1,1) - 2h_i(0,1) + h_i(0,0)\big). \end{aligned} \tag{E.6}$$

Because the third line of (E.6) is independent $\mathbf{v}$, (15) may be combined with (E.3), (E.4), (E.5), and (E.6) to deduce that $P_p(\mathbf{v}|\mathbf{Y}) \propto 1 + c(p) + p\sum_{j=1}^{2f}\exp(s_j) + o(p)$, with $c(p) = O(p)$ independent of $\mathbf{v}$. Formula (51) finally follows by using the definition of $h_i(\cdot,\cdot)$ and $\sum_w P_p(\mathbf{v} = \mathbf{w}|\mathbf{Y}) = 1$.

[Received October 2001. Revised September 2003.]

REFERENCES

Almasy, L., and Blangero, J. (1998), “Multipoint Quantitative Trait Linkage Analysis in General Pedigrees,” *American Journal of Human Genetics*, 62, 1198–1211.

Ängquist, L., and Hössjer, O. (2003), “Improving the Calculation of Statistical Significance in Genome-Wide Scans,” Report 2003:3, Mathematical Statistics, Stockholm University.

Bengtsson, O. (2001), “Two-Locus Affected Sib-Pair Identity by Descent Probabilities,” licentiate thesis, Chalmers University of Technology, Gothenburg.

Commenges, D. (1994), “Robust Genetic Linkage Analysis Based on a Score Test of Homogeneity: The Weighted Pairwise Correlation Statistic,” *Genetic Epidemiology*, 11, 189–200.

Cordell, H. J., Todd, J. A., Bennett, S. T., Kawaguchi, Y., and Farrall, M. (1995), “Two-Locus Maximum LOD Score Analysis of Multifactorial Trait: Joint Considerations of IDDM2 and IDDM4 With IDDM1 in Type I Diabetes,” *American Journal of Human Genetics*, 57, 920–934.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall.

Dudoit, S., and Speed, T. P. (1999), “Triangle Constraints for Sib-Pair Identity by Descent Probabilities Under a General Multilocus Model for Disease Susceptibility,” in *Statistics in Genetics*, eds. M. E. Halloran and S. Geisser, New York: Springer.

——— (2000), “A Score Test for Linkage Analysis of Qualitative and Quantitative Traits Based on Identity by Descent Data From Sib-Pairs,” *Biostatistics*, 1, 1–26.

Feingold, E., Brown, P. O., and Siegmund, D. (1993), “Gaussian Models for Genetic Linkage Analysis Using Complete High-Resolution Maps of Identity by Descent,” *American Journal of Human Genetics*, 53, 234–251.

Feingold, E., and Siegmund, D. (1997), “Strategies for Mapping Heterogeneous Recessive Traits by Allele-Sharing Methods,” *American Journal of Human Genetics*, 60, 965–978.

Goldstein, D. R., Dudoit, S., and Speed, T. P. (2001), “Power and Robustness of a Score Test for Linkage Analysis of Quantitative Traits Using Identity by Descent Data for Sib Pairs,” *Genetic Epidemiology*, 20, 415–431.

Haseman, J. K., and Elston, R. C. (1972), “The Investigation of Linkage Between a Quantitative Trait and a Marker Locus,” *Behavior Genetics*, 2, 3–19.

Hössjer, O. (2001), “Determining Inheritance Distributions via Stochastic Penetrances,” Report 2001:17, Lund University, Centre for Mathematical Sciences.

——— (2003a), “Asymptotic Estimation Theory of Multipoint Linkage Analysis Under Perfect Marker Information,” *The Annals of Statistics*, 31, 1075–1109.

——— (2003b), “Assessing Accuracy in Linkage Analysis by Means of Confidence Regions,” *Genetic Epidemiology*, 25, 59–72.

——— (2003c), “Conditional Likelihood Score Functions in Linkage Analysis,” Report 2003:10, Stockholm University, Division of Mathematical Statistics.

James, J. W. (1971), “Frequencies in Relatives for an All-or-None Trait,” *The Annals of Human Genetics*, 35, 47–48.

Kempthorne, O. (1957), *An Introduction to Genetic Statistics*, New York: Wiley.

Kong, A., and Cox, N. J. (1997), “Allele-Sharing Models: Lod Scores and Accurate Linkage Tests,” *American Journal of Human Genetics*, 61, 1179–1188.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996), "Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach," *American Journal of Human Genetics*, 58, 1347–1363.

Lander, E. S., and Green, P. (1987), "Construction of Multilocus Genetic Maps in Humans," *Proceedings of the National Academy of Science U S A*, 84, 2363–2367.

Lander, E. S., and Kruglyak, L. (1995), "Genetic Dissection of Complex Traits: Guidelines for Interpreting and Reporting Linkage Results," *Nature Genetics*, 11, 241–247.

Lee, A. J. (1990), *U-Statistics, Theory and Practice*, New York: Marcel Dekker.

Liang, K.-Y., Chiu, Y.-F., and Beaty, T. H. (2001), "A Robust Identity-by-Descent Procedure Using Affected Sib Pairs: Multipoint Mapping for Complex Diseases," *Human Heredity*, 51, 64–78.

Liang, K.-Y., Huang, C.-Y., and Beaty, T. H. (2000), "A Unified Sampling Approach to Multipoint Analysis of Qualitative and Quantitative Traits in Sib Pairs," *American Journal of Human Genetics*, 66, 1631–1641.

McPeek, M. S. (1999), "Optimal Allele-Sharing Statistics for Genetic Mapping Using Affected Relatives," *Genetic Epidemiology*, 16, 225–249.

Nicolae, D. L. (1999), "Allele Sharing Models in Gene Mapping: A Likelihood Approach," Ph.D. thesis, University of Chicago, Dept. of Statistics.

Nilsson, S. (1999), "Model-Based Sampling and Weights in Affected Sib Pair Methods," licentiate thesis, Chalmers University of Technology.

Putter, H., Sandkuijl, L. A., and van Houwelingen, J. C. (2002), "Score Test for Detecting Linkage to Quantitative Traits," *Genetic Epidemiology*, 22, 345–355.

Risch, N. (1984), "Segregation Analysis Incorporating Genetic Markers, I: Single-Locus Models With an Application to Type I Diabetes," *American Journal of Human Genetics*, 36, 363–386.

——— (1990a), "Linkage Strategies for Genetically Complex Traits, I: Multilocus Models," *American Journal of Human Genetics*, 46, 222–228.

——— (1990b), "Linkage Strategies for Genetically Complex Traits, II: The Power of Affected Relative Pairs," *American Journal of Human Genetics*, 46, 229–241.

Risch, N., and Zhang, H. (1995), "Extreme Discordant Sib Pairs for Mapping Quantitative Trati Loci in Humans," *Science*, 268, 1584–1589.

——— (1996), "Mapping Quantitative Trait Loci With Extreme Discordant Sib Pairs: Sampling Considerations," *American Journal of Human Genetics*, 58, 836–843.

Sham, P. C., Purcell, S., Cherny, S. S., and Abecasis, G. R. (2002), "Powerful Regression-Based Quantitative-Trait Linkage Analysis of General Pedigrees," *American Journal of Human Genetics*, 71, 238–251.

Sham, P. C., Zhao, J. H., Cherny, S. S., and Hewitt, J. K. (2000), "Variance Components QTL Linkage Analysis of Selected and Nonnormal Samples, Conditioning on Trait Values," *Genetic Epidemiology*, 10(Suppl 1), S22–S28.

Sham, P., Zhao, J., and Curtis, D. (1997), "Optimal Weighting Scheme for Affected Sib-Pair Analysis of Sibship Data," *The Annals of Human Genetics*, 61, 61–69.

Suarez, B. K., Rice, J., and Reich, T. (1978), "The Generalized Sib Pair IBD Distribution and Its Use in Detection of Linkage," *The Annals of Human Genetics*, 44, 87–94.

Tang, H.-K., and Siegmund, D. (2001), "Mapping Quantitative Trait Loci in Oligogenic Models," *Biostatistics*, 2, 147–162.

Teng, J., and Siegmund, D. (1997), "Combining Information Within and Between Pedigrees for Mapping Complex Traits," *American Journal of Human Genetics*, 60, 979–992.

Thomas, D. C., and Gauderman, W. J. (1996), "Gibbs Sampling Methods in Genetics," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, New York: Chapman & Hall.

Thompson, E. A. (1974), "Gene Identities and Multiple Relationships," *Biometrics*, 30, 667–680.

——— (1997), "Conditional Gene Indentity in Affected Individuals," in *Genetic Mapping of Disease Genes*, eds. I.-H. Pawlowitzki, J. H. Edwards, and E. A. Thompson, San Diego, CA: Academic Press, pp. 137–146.

Wang, K., and Huang, J. (2002), "A Score-Statistic Approach for the Mapping of Quantitative-Trait Loci With Sibships of Arbitrary Size," *American Journal of Human Genetics*, 70, 412–424.

Whittemore, A. S. (1996), "Genome Scanning for Linkage: An Overview," *American Journal of Human Genetics*, 59, 704–716.

Whittemore A. S., and Halpern, J. (1994), "A Class of Tests for Linkage Using Affected Pedigree Members," *Biometrics*, 50, 118–127.