Biostatistics (2005), **6**, 2, *pp.* 313–332
doi:10.1093/biostatistics/kxi012

Conditional likelihood score functions for mixed models in linkage analysis

OLA HÖSSJER

Department of Mathematics, Stockholm University, S-106 91 Stockholm, Sweden
ola@math.su.se

SUMMARY

In this paper, we develop a general strategy for linkage analysis, applicable for arbitrary pedigree structures and genetic models with one major gene, polygenes and shared environmental effects. Extending work of Whittemore (1996), McPeek (1999) and Hössjer (2003d), the efficient score statistic is computed from a conditional likelihood of marker data given phenotypes. The resulting semiparametric linkage analysis is very similar to nonparametric linkage based on affected individuals. The efficient score S depends not only on identical-by-descent sharing and phenotypes, but also on a few parameters chosen by the user. We focus on (1) weak penetrance models, where the major gene has a small effect and (2) rare disease models, where the major gene has a possibly strong effect but the disease causing allele is rare. We illustrate our results for a large class of genetic models with a multivariate Gaussian liability. This class incorporates one major gene, polygenes and shared environmental effects in the liability, and allows e.g. binary, Gaussian, Poisson distributed and life-length phenotypes. A detailed simulation study is conducted for Gaussian phenotypes. The performance of the two optimal score functions S_{wpairs} and S_{normdom} are investigated. The conclusion is that (i) inclusion of polygenic effects into the score function increases overall performance for a wide range of genetic models and (ii) score functions based on the rare disease assumption are slightly more powerful.

Keywords: Conditional likelihood; Founder alleles; Linkage analysis; Mixed model; Score functions.

1. INTRODUCTION

The goal of linkage analysis is to find the position (locus) along the genome of a gene which causes or increases the risk of development of a certain inheritable disease. Disease related quantities, so called phenotypes, and DNA samples, are collected for a number of families with aggregation of the disease. The DNA samples are typed at a large number of genetic markers, distributed throughout the genome. Using information from the markers, loci are sought at which segregation of DNA is correlated with the inheritance pattern of the disease. Since each individual's phenotype is a blurred observation of the two copies of the disease gene (disease alleles) that he/she carries, DNA transmission is correlated with phenotype segregation in close vicinity of the disease locus.

If genetic model parameters (disease allele frequencies and penetrance parameters) are known, the lod score of Morton (1955) can be computed at each locus. For complex diseases, the genetic model is rarely known and alternative procedures have been proposed. One possibility is to regard the genetic model parameters as nuisance parameters and optimize the lod score with respect to them at each locus. This is the mod score approach of Risch (1984) and Clerget-Darpoux *et al.* (1986).

For binary traits, nonparametric linkage (NPL) is another method developed in situations when the genetic model parameters are unknown. The affected-pedigree-member (APM) method is the most commonly used version of NPL. A score function S is used which does not require specification of a genetic model. Instead, it quantifies the extent to which the APMs share their alleles identical-by-descent (IBD) from the same founder alleles, see e.g. Penrose (1935), Weeks and Lange (1988), Fimmers *et al.* (1989), Whittemore and Halpern (1994) and Kruglyak *et al.* (1996).

The NPL method can be extended to more general phenotypes by first noting that the lod score is equivalent to a conditional likelihood L of DNA marker data (MD) given phenotypes (Whittemore, 1996). By differentiating $\log L$ with respect to genetic model parameters a score function S is obtained. Whittemore (1996) showed that allele sharing score functions S used in NPL can be used to construct an appropriate L. The opposite route is to start from L, based on a biologically based genetic model with disease allele frequencies and penetrance parameters, and then to compute S. McPeek (1999) showed, for binary traits, arbitrary pedigree structures and affected-only phenotypes, that several allele sharing score functions S could be derived in this way. McPeek's work was generalized in Hössjer (2003d) to arbitrary monogenic models, e.g. Gaussian, Cox proportional hazards and logistic regression models. The resulting score functions S depend on IBD allele sharing in the pedigree, the observed phenotypes and a small number of parameters which have to be specified by the user. This approach was referred to as semiparametric linkage (SPL) in Hössjer (2003d).

The purpose of this paper is to extend the work of McPeek (1999) and Hössjer (2003d) to mixed genetic models with one major susceptibility locus and polygenes/shared environmental effects. In Section 2, we define basic genetic concepts. In Section 3, we provide a general formulation of the efficient score function approach. In Section 4, we define a broad class of genetic models which includes several existing models based on Gaussian, binary, life-length or Poisson phenotypes as special cases. In Section 5, we derive the efficient score functions S for weak penetrance models, where the major gene has a small effect on the disease, and rare disease models, where the disease allele is rare. A simulation study in Section 6 for the Gaussian mixed model reveals that more powerful and robust procedures are obtained when polygenic effects are included in the score function S. Some conclusions and further recommendations are given in Section 7. In Section A of supplementary data available at *Biostatistics* online, henceforth denoted HS, we show that the SPL approach is asymptotically equivalent to mod scores under the null hypothesis of no linkage. In Section B of HS, we derive an orthogonal decomposition of functions of several genotypes. This is used in Section 5.1 to derive weak penetrance optimal score functions, but we believe it to be of independent interest. For instance, the classical variance components decomposition of genetic variance (Fisher, 1918; Kempthorne, 1955) can be obtained from these expansions. Finally, the proofs are collected in Section C of HS.

2. Basic genetic concepts

Consider a pedigree with n individuals of which f are founders (without ancestors in the pedigree) and $n - f$ nonfounders. Assume that two forms of the disease gene exist—the normal allele (0) and the disease allele (1). Each individual has a pair of alleles (genotype), of which one is inherited from the father and one from the mother. The genotypes of all pedigree members can be collected into a vector $G = (G_1, \ldots, G_n) = (a_1, \ldots, a_{2n})$, where $G_k = (a_{2k-1}, a_{2k})$ is the genotype of the kth individual, with a_{2k-1} and a_{2k} the paternally and maternally transmitted alleles, respectively. We will use the convention that founders are numbered $k = 1, \ldots, f$. Since the gene of interest is unknown, we do not observe G, but rather a vector of disease phenotypes $Y = (Y_1, \ldots, Y_n)$. Here Y_k is the phenotype of the kth individual, a quantity related to the disease. This could be binary (affected/unaffected) or a quantitative variable such as insulin concentration, body mass index, etc. Some Y_k may also represent unknown phenotypes.

Alleles are transmitted from parents to children according to Mendel's law of segregation. At a certain locus t, allele transmission can be summarized through the inheritance vector, introduced by Donnelly (1983). It is defined as $v(t) = (v_1(t), \ldots, v_m(t))$, where $m = 2(n - f)$ is the number of meioses and $v_k(t)$ equals 0 or 1 depending on whether a grandpaternal or grandmaternal allele was transmitted during the kth meiosis. A priori, grandpaternal and grandmaternal alleles are equally likely to be transmitted during each meiosis, so that

$$P(v(t) = w) = 2^{-m} \tag{2.1}$$

for all binary vectors w of length m.

If τ is the disease locus, the phenotype vector Y will be correlated to $v(\tau)$ if the genetic component is strong enough. It is standard in linkage analysis to look at the conditional distribution

$$w \longrightarrow P(v(\tau) = w|Y), \tag{2.2}$$

which differs from the prior (2.1). Moreover, (2.2) is unaffected by the way we sampled the pedigree (ascertainment). The stronger the genetic component at τ is, the stronger is the discrepancy between $P(v(\tau)|Y)$ and the uniform distribution, because of co-inheritance of phenotypes and DNA at τ. Even at loci around τ, the conditional distribution of the inheritance vector given phenotypes differs from (2.1), although the amount of co-inheritance decays with the genetic distance from the disease locus. This is because of occurrence of so called crossovers—random points along the chromosome where, during meioses, segregation switches between grandmaternal and grandpaternal transmission.

In practice we do not observe the process $v(\cdot)$, but MD from all or a subset of the pedigree members give incomplete information of it. As we will see in the next section, up to a multiplicative constant the conditional probability $P_{t,\theta}(\text{MD}|Y)$ serves as an approximation of (2.2). Here, θ is the set of genetic model parameters (disease allele frequencies and penetrance parameters) and t corresponds to the hypothesis $t = \tau$.

3. Conditional likelihood and score functions

Consider a chromosome of length $t_{\max}$ centimorgans (cM) with at most one disease locus τ located along $[0, t_{\max}]$. The hypothesis testing problem is

$$H_0\!: \ \tau \text{ is located on another chromosome}$$
$$H_1\!: \ \tau \in [0, t_{\max}].$$

The unconditional likelihood requires knowledge of MD, Y and the sampling scheme. Since the latter is often (more or less) unknown, it is common in linkage analysis to consider the conditional likelihood of MD given Y. The reason is that nuisance parameters involved in the ascertainment scheme are not included in the conditional likelihood, see e.g. Ewens and Shute (1986). If $v = v(\tau)$ is the inheritance vector at the disease locus, the conditional likelihood can be written as

$$\begin{aligned} P_{t,\theta}(\text{MD}|Y) &= \sum_v P_t(\text{MD}|v) P_\theta(v|Y) \\ &= 2^m P(\text{MD}) \sum_v P_t(v|\text{MD}) P_\theta(v|Y), \end{aligned} \tag{3.1}$$

where $\sum_v P_t(\text{MD}|v) P_\theta(v|Y)$ is short for $\sum_w P_t(\text{MD}|v = w) P_\theta(v = w|Y)$. In the last equality of (3.1) we applied Bayes' rule and $P(v) = 2^{-m}$. The factor $P(\text{MD})$ depends on marker allele frequencies and genetic distances between the markers. These are assumed to be known in linkage analysis. Since $2^m P(\text{MD})$ is independent of t and θ, it is a fixed constant which we drop. Hence,

$$L(t, \theta; \text{MD}) = \sum_v P_t(v|\text{MD}) P_\theta(v|Y) = E_t(P_\theta(v|Y)|\text{MD}) \tag{3.2}$$

is our conditional likelihood. Notice that we only include MD as an argument of L, because we condition on the phenotype vector Y and consider it as fixed. Kruglyak *et al.* (1996) noticed that the conditional inheritance distribution $P_t(v|\mathrm{MD})$ could be used both for NPL and parametric linkage analysis based on lod scores.

In linkage analysis, t is the structural parameter of interest, whereas θ contains nuisance parameters. The lod score was originally formulated as the base ten logarithm of a likelihood ratio, but it is a linear transformation of, and hence equivalent to, $Z(t) = Z(t;\theta) = \log L(t,\theta)$, as noted by Whittemore (1996). We regard $Z(t)$ as a local test statistic for testing H_0 against the simple alternative hypothesis $\tau = t$. When testing H_0 against H_1, the global test statistic

$$Z_{\max} = \sup_{t\in\Omega} Z(t) \tag{3.3}$$

is used instead, where $\Omega = [0, t_{\max}]$ or, more generally, Ω may consist of several chromosomes. The null hypothesis is rejected when $Z_{\max}$ exceeds a given threshold T, which depends on the chosen significance level, Ω and θ.

When θ is unknown, the profile conditional likelihood $Z(t;\hat{\theta}(t)) = \sup_\theta Z(t;\theta)$ can be computed at each t and H_0 is rejected when $\max_t Z(t;\hat{\theta}(t))$ exceeds a given threshold. This procedure is equivalent to mod scores. The mod score is computationally demanding, especially for larger pedigrees, although a faster version is obtained by replacing the original $L(t,\theta)$, based on penetrance and disease allele parameters, by a simplified one with only one parameter, see Whittemore (1996) and Kong and Cox (1997).

For quantitative phenotypes, it is common to use variance components techniques, see e.g. Amos (1994) and Almasy and Blangero (1998). These can be interpreted as an approximate version of the mod score, where the original $L(t,\theta)$, based on penetrance and disease allele parameters, has been replaced by a simplified one, which is a ratio of two multivariate normal densities.

In this paper, we will not maximize with respect to θ at each locus. Instead, we take a local approach and assume that $\{\theta_\varepsilon\}$ is a one-dimensional trajectory of genetic model parameters such that θ_0 corresponds to no genetic effect at the disease locus, i.e. $P_{\theta_0}(v|Y) = 2^{-m}$. In other words, under θ_0 the prior distribution of v equals the posterior distribution $v|Y$. Our objective is to compute a likelihood score function by differentiating the log conditional likelihood w.r.t. ε at each locus t.

Assume first complete MD. It corresponds to an infinitely dense set of markers genotyped for all pedigree members so that $P(v = v(t)|\mathrm{MD})$ may be determined unambiguously at all loci. We let $\mathrm{MD}^{\mathrm{compl}}$ denote complete MD. The corresponding conditional likelihood is

$$L(t,\theta;\mathrm{MD}^{\mathrm{compl}}) = P_\theta(v = v(t)|Y). \tag{3.4}$$

Define the score function

$$S(v) = \left.\frac{\mathrm{d}^\rho \log P_{\theta_\varepsilon}(v|Y)}{\mathrm{d}\varepsilon^\rho}\right|_{\varepsilon=0}, \tag{3.5}$$

where ρ is the smallest positive integer such that the right-hand side of (3.5) is nonzero for at least one v. When $\rho \geqslant 2$ the estimation problem is singular (with zero Fisher information) for ε at $\varepsilon = 0$. By means of the reparametrization

$$\epsilon = \varepsilon^\rho/\rho!, \tag{3.6}$$

S is interpreted as a (conditional) likelihood score function for ϵ at $\epsilon = 0$. Originally, Whittemore (1996) used $\rho = 1$ in her definition of S, but $\rho = 2$ is also possible for weak penetrance models, see McPeek (1999) and Hössjer (2003d).

Our goal is to test $\varepsilon = 0$ against $\varepsilon \neq 0$ at each locus t, which can also be interpreted as testing H_0 against $\tau = t$ at each t. Depending on the application, there may or may not be a sign constraint on ε. In any case, the sign of ε is not of interest, and for this reason we use ϵ instead of ε as parameter when formulating a test statistic, also when ρ is even. The Fisher information for ϵ is $I^{\text{compl}} = E_{\theta_0}(S^2(v)|Y) = 2^{-m}\sum_w S^2(w)$ at $\epsilon = 0$, with the sum ranging over all binary vectors w of length m. The square root of the likelihood score statistic for testing $\epsilon = 0$ against $\epsilon \neq 0$ is $W^{\text{compl}}(t) = S(v(t))/\sqrt{I^{\text{compl}}}$ at locus t.

For incomplete MD, the observed conditional likelihood $L(t, \theta; \text{MD}) = E(L(t, \theta; \text{MD}^{\text{compl}})|\text{MD})$ is obtained by averaging the complete conditional likelihood, treating MD$^{\text{compl}}$ as hidden data. Then

$$W(t) = \frac{[\mathrm{d}^\rho \log L(t, \theta_\varepsilon; \text{MD})/\mathrm{d}\varepsilon^\rho]_{\varepsilon=0}}{\sqrt{I(t)}} = \frac{\sqrt{I^{\text{compl}}}}{\sqrt{I(t)}} E\left(W^{\text{compl}}(t)|\text{MD}\right), \tag{3.7}$$

where $I(t) = E_{\theta_0}(E^2(S(v(t))|\text{MD})|Y)$ is the Fisher information. It depends on t in a way that reflects positioning and informativity of markers. The outer expectation is taken over variations in MD and is typically computationally involved. It can be calculated exactly (Whittemore and Halpern, 1994), approximated as $I(t) \approx I^{\text{compl}}$ so that the first factor of the right-hand side of (3.7) vanishes (Kruglyak *et al.*, 1996) or approximated by a multiple imputation Monte Carlo algorithm (Clayton, 2001). Yet, another method of handling incomplete MD was suggested by Kong and Cox (1997).

By definition of $W(t)$ we have

$$\begin{aligned} E_{\theta_0}(W(t)) &= 0, \\ V_{\theta_0}(W(t)) &= 1, \end{aligned} \tag{3.8}$$

where expectation is with respect to variations in MD.

So far, we have only discussed a single family. The extension to N pedigrees with mutually independent phenotype/MD is straightforward. We allow the pedigree structures to vary arbitrarily and index quantities for the ith family with i. The overall conditional likelihood $L(t, \theta) = \prod_{i=1}^N L_i(t, \theta)$ is then simply the product of the familywise conditional likelihoods, and the total Fisher information is $I(t) = \sum_{i=1}^N I_i(t)$ at locus t. From this it follows that the linkage score function for N families can be written as

$$W(t) = \frac{[\mathrm{d}^\rho \log L(t, \theta_\varepsilon; \text{MD})/\mathrm{d}\varepsilon^\rho]_{\varepsilon=0}}{\sqrt{I(t)}} = \sum_{i=1}^N \gamma_i(t) W_i(t), \tag{3.9}$$

where $W_i(t)$ is the score (3.7) for family i and $\gamma_i(t) = \sqrt{I_i(t)/\sum_{j=1}^N I_j(t)}$ are the locally optimal weights (McPeek, 1999). Since $\sum_1^N \gamma_i^2(t) = 1$, (3.8) also holds for the total linkage score with N families. Whittemore (1996) noticed that (3.9) includes APM methods as special case.

The local SPL test statistic at locus t is defined as

$$Z(t) = \begin{cases} W(t)^2, & \text{no sign constraint on } \epsilon, \\ W(t), & \epsilon \geqslant 0, \end{cases} \tag{3.10}$$

and the corresponding global test statistic for testing H_0 against H_1 is obtained by inserting $Z(t)$ in (3.10) into (3.3). Two separate definitions of $Z(t)$ are needed, because $\epsilon = 0$ is at the boundary of the parameter space when $\epsilon \geqslant 0$. This is always the case when ρ is even and sometimes, depending on the application, when ρ is odd (in that case $\varepsilon \geqslant 0$ is imposed). A derivation of (3.10) is given in Section A of HS. There, it is also shown that the SPL and mod score approaches are asymptotically equivalent under the null hypothesis of no linkage when the genetic model parameters are restricted to a one-dimensional trajectory $\{\theta_\varepsilon\}$.

For quantitative phenotypes, the SPL approach is also asymptotically equivalent to variance component techniques when the parameters of the latter model are varied along a one-dimensional trajectory. This is briefly discussed in Example 6 of Section 5.

4. GENETIC MODELS

To begin with, it is mathematically more convenient to work with

$$P_\theta(Y|v) = 2^m P_\theta(v|Y) P_\theta(Y) \tag{4.1}$$

than with $P_\theta(v|Y)$. The reason is that

$$P_\theta(Y|v) = \sum_G P_\psi(Y|G) P_p(G|v) \tag{4.2}$$

can be expanded by summing over all possible genotype configurations in the pedigree. In (4.2), we split $\theta = (p, \psi)$ into p, the frequency (probability) of the disease allele, and ψ, the penetrance parameter(s). The latter describe the relationship between phenotypes and genotypes.

We will introduce a class of genetic models with penetrance factor of the form

$$P_\psi(Y|G) = \int P_{\psi_Y}(Y|x) P_{\psi_X}(x|G)\,\mathrm{d}x \tag{4.3}$$

by integrating with respect to a vector $X = (X_1, \ldots, X_n)$ of liabilities. This vector contains influences from the major gene G as well as polygenes and environmental components. Given X, the components of Y are conditionally independent

$$P_{\psi_Y}(Y|X) = \prod_{k=1}^{n} P(Y_k|X_k, z_k), \tag{4.4}$$

where z_k is a (possibly empty) set of observed covariates for k and ψ_Y is the (possibly empty) set of penetrance parameters involved in (4.4). In many cases Y_k is just a deterministic function of X_k. If k has unknown phenotype we put $P(Y_k|X_k, z_k) = 1$.

Conditional on G, we assume that the liability vector is multivariate normal,

$$X|G \in N(\mu(G) + \beta\Lambda, \sigma^2\Sigma), \tag{4.5}$$

where $s \geqslant 0$ is the number of covariates, Λ is the $s \times n$ design matrix of observed covariates and β is a $1 \times s$ vector of regression coefficients. The vector $\mu(G)$ depends on the major gene G whereas the stochastic variation is due to polygenic and environmental effects. Therefore, neither the conditional variance $\sigma^2 = \mathrm{Var}(X_k|G)$ nor the conditional correlation matrix $\Sigma = \mathrm{Corr}(X|G)$ depends on G.

In order to describe $\mu(G)$ and Σ, we need some more definitions. Assume there are numbers m_0, m_1 and m_2 such that $X_k|G_k \in N(m_{|G_k|} + (\beta\Lambda)_k, \sigma^2)$, with $|G_k| = a_{2k-1} + a_{2k}$ the number of disease alleles of G_k. Then put

$$\mu(G) = (m_{|G_1|}, \ldots, m_{|G_n|}). \tag{4.6}$$

For instance, if large values of the liability indicate disease, a natural constraint is $m_0 \leqslant m_1 \leqslant m_2$. Let $\mathrm{IBD}_{kl} = \mathrm{IBD}_{kl}(w)$ be the number of alleles that two individuals k and l ($1 \leqslant k, l \leqslant n$) share identical by descent for inheritance vector w. The coefficient of relationship between k and l is defined as $r_{kl} = E(\mathrm{IBD}_{kl}(w))/2$. We also put $\delta_{kl} = P(\mathrm{IBD}_{kl}(w) = 2)$, where expectation and probability is taken w.r.t. a uniform distribution (2.1). Following Fisher (1918) and Kempthorne (1955), the correlation matrix with polygenic and shared environmental effects is

$$\Sigma = (1 - h_a^2 - h_d^2 - h_s^2)I_n + h_a^2 R + h_d^2 \Delta + h_s^2 S, \tag{4.7}$$

where I_n is an identity matrix of order n, $R = (r_{kl})$ and $\Delta = (\delta_{kl})$. Further, h_a^2 and h_d^2 are the additive and dominant polygenic heritabilities, respectively. These are the fractions of total environmental and polygenic liability variance (σ^2) due to additive and dominant genetic effects, respectively. The matrix $S = (s_{kl})$ models a shared environment. For instance, each entry s_{kl} can be put to zero or one depending on whether k and l share the same household or not. The parameter h_s^2 is the fraction of σ^2 due to shared environment. The part of the penetrance vector ψ involved in (4.5) is $\psi_X = (m_0, m_1, m_2, \sigma^2, h_a^2, h_d^2, h_s^2, \beta)$, with the first three components due to the major gene, and the next five caused by polygenic/environmental effects.

Summarizing, (4.3)–(4.5) is a penetrance model with multivariate Gaussian liability X and penetrance parameters $\psi = (\psi_X, \psi_Y)$. A similar class of models has been suggested in the geostatistical literature by Diggle *et al*. (1999).

EXAMPLE 1 (GAUSSIAN MIXED MODEL) When liabilities are observed, we put $Y = X$. This is the Gaussian mixed model of Ott (1979). The name refers to Y being a mixture of multivariate normal distributions $Y|G \in N(\mu(G) + \beta\Lambda, \sigma^2\Sigma)$.

EXAMPLE 2 (LIABILITY THRESHOLD MODEL) When the phenotypes Y_k are binary ($Y_k = 1$ affected, $Y_k = 0$ unaffected) but show no simple Mendelian inheritance pattern, it is common to model the distribution of Y_k as a function of an underlying quantitative variable X_k involving alleles from several loci. The liability threshold model was originally introduced by Pearson and Lee (1901). With T a given threshold, the phenotypes are defined as $Y_k = 1_{\{X_k \geqslant T\}}$ and $\psi_Y = \{T\}$. For identifiability we assume $m = 0$ and $\sigma^2 = 1$, where

$$m = E(X_k) - (\beta\Lambda)_k = q^2 m_0 + 2pqm_1 + p^2 m_2$$

and $q = 1 - p$. Usually this model does not include a covariate, but we may include a single covariate containing the liability class or age of each individual. For a recent review of binary liability models with various extensions, see Todorov and Suarez (2002).

EXAMPLE 3 (LOGISTIC REGRESSION) As in the previous example, we consider a binary trait with $Y_k = 1$ and 0 corresponding to an affected or unaffected individual. We also include a design vector $\Lambda = (t_1, \ldots, t_n)$, where t_k is either the time of examination or time of onset of k and β is a nonnegative regression parameter. Then assume

$$\begin{aligned} P(Y_k|X_k) &= F(X_k)^{Y_k}(1 - F(X_k))^{1-Y_k}, \quad \text{if } t_k \text{ is age of examination,} \\ P(Y_k|X_k) &= \beta f(X_k), \quad \text{if } Y_k = 1 \text{ and } t_k \text{ is age of onset,} \end{aligned}$$

where $F(x) = \mathrm{e}^x/(1 + \mathrm{e}^x)$ is the logistic distribution function and $f(x) = F'(x)$ the corresponding density. In this case $\psi_Y = \{\beta\}$ (so that β appears in ψ_X and ψ_Y). A similar model has been considered by Bonney (1986) and Elston and George (1989), but these authors use a Markov rather than Gaussian liability model for Y.

EXAMPLE 4 (SURVIVAL ANALYSIS) In Example 3, an alternative is to use a Cox proportional hazards model with hazard rate $\lambda(t; X_k) = \lambda_0(t)\exp(X_k)$, baseline hazard λ_0 and distribution function $F(t; X_k) = 1 - \exp(-\int_0^t \lambda(u; X_k)\,\mathrm{d}u)$. In this case there are no covariates ($s = 0$) in (4.5). Instead, we include t_k as covariate in (4.4) ($z_k = t_k$) and put

$$\begin{aligned} P(Y_k|X_k, t_k) &= F(t_k; X_k)^{Y_k}(1 - F(t_k; X_k))^{1-Y_k}, \quad \text{if } t_k \text{ is age of examination,} \\ P(Y_k|X_k, t_k) &= f(t_k; X_k), \quad \text{if } Y_k = 1 \text{ and } t_k \text{ is age of onset,} \end{aligned}$$

so that $\psi_Y = \{\lambda_0\}$. For identifiability we put $m = 0$. See Thomas and Gauderman (1996) for more details.

Another possibility is that Y_k and X_k are related through a generalized linear model (McCullagh and Nelder, 1989), i.e. $h(E(Y_k)) = X_k$ for some link function h. An example is Poisson distributed data $Y_k \in \mathrm{Po}(\exp(X_k))$.

5. CHOOSING SCORE FUNCTIONS

To start with, we establish the following simple but very useful result.

PROPOSITION 1 Let $\bar{S}(v) = \mathrm{d}^\rho \log P_{\theta_\varepsilon}(Y|v)/\mathrm{d}\varepsilon^\rho|_{\varepsilon=0}$ be the score function of $P_{\theta_\varepsilon}(Y|v)$ at $\varepsilon = 0$. Then S is the centered version of $\bar{S}$, i.e.

$$S(v) = \bar{S}(v) - C,$$

where C is a centering constant, ensuring that $E_{\theta_0}(S(v)|Y) = 2^{-m}\sum_w S(w) = 0$.

Proposition 1 implies that it suffices to consider score functions of $P_\theta(Y|v)$. In formula (4.2), $P_\theta(Y|v)$ is defined by summing over all possible genotype vectors G. This is equivalent to summing over all founder allele vectors $a = (a_1, \ldots, a_{2f})$. In fact, $J(w) = (j_1(w), \ldots, j_{2n}(w))$, the gene-identity state of the pedigree (Thompson, 1974), is a function of the inheritance vector w, such that $j_k(w) \in \{1, \ldots, 2f\}$ is the number of the founder allele that has been transmitted to allele number k. Since

$$G = G(a, v) = a_{J(v)} = \left(a_{j_{1(v)}}, \ldots, a_{j_{2n(v)}}\right),$$

we obtain

$$P_\theta(Y|v) = \sum_a P_\psi(Y|a, v)P_p(a) = E_p(P_\psi(Y|a, v)), \tag{5.1}$$

where the sum ranges over all 2^{2f} possible founder allele vectors a and the last expectation is w.r.t. a. We assumed in (5.1) that a and v are independent (no segregation distortion), so that $P_p(G|v) = P_p(a)$. Under random mating, the components of a are independent,

$$P_p(a) = p^{|a|}q^{2f-|a|}, \tag{5.2}$$

where $|a| = \sum_1^{2f} a_j$. Viewing a as hidden data, $P_\psi(Y|a, v)$ is the complete likelihood corresponding to $P_\theta(Y|v)$. Formulas (5.1) and (5.2) will be used in the next two subsections for deriving score functions S.

5.1 *Local penetrance models*

Assume the disease allele frequency p is fixed whereas the penetrance parameters ψ_ε vary with ε so that $P_{\psi_0}(Y|G) = P_{\psi_0}(Y)$ is independent of G. This implies no genetic effect at the disease locus when $\varepsilon = 0$. In more detail, we consider penetrance functions of the form

$$P_\psi(Y|G) = f(Y; \mu), \tag{5.3}$$

with $\mu = \mu(G)$ as in (4.6). The Gaussian liability class of models (4.3) is included in (5.3). For brevity, we write $\psi = (m_0, m_1, m_2)$, since only these three penetrance parameters depend on ε according to

$$\psi_\varepsilon = (m^*, m^*, m^*) + \varepsilon(u(0), u(1), u(2)). \tag{5.4}$$

Hence, it is only μ that depends on ε in (5.3). Define $\sigma_g^2 = \mathrm{Var}(u(|G_k|))$. Then, in (4.5), the variance of the liability is $\mathrm{Var}(X_k) = \varepsilon^2\sigma_g^2 + \sigma^2$. The first term $(\varepsilon^2\sigma_g^2)$ is genetic variance at the main locus and the second term (σ^2) is variance due to polygenic and shared environmental effects. We may further split $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$ into additive and dominant variance components, where

$$\begin{aligned}\sigma_a^2 &= 2pq(p(u(2) - u(1)) + q(u(1) - u(0)))^2,\\ \sigma_d^2 &= (pq)^2(u(2) - 2u(1) + u(0))^2.\end{aligned} \tag{5.5}$$

Let $c = \sigma_d^2/\sigma_g^2$ be the fraction of dominance variance at the main locus, put $\mu_0 = (m^*, \ldots, m^*)$ and introduce weights

$$\begin{aligned}\omega_k &= \omega_k(Y) = \sigma_g \, \partial f(Y; \mu)/\partial \mu_k|_{\mu=\mu_0} / f(Y; \mu_0),\\ \omega_{kl} &= \omega_{kl}(Y) = \sigma_g^2 \partial^2 f(Y; \mu)/\partial \mu_k \partial \mu_l|_{\mu=\mu_0} / f(Y; \mu_0),\end{aligned} \tag{5.6}$$

assigned to individuals and pairs of individuals. Then, the following result holds.

THEOREM 1 Consider a weak penetrance model (4.6), (5.3), (5.4) and assume random mating (5.2). Then, for an inbred pedigree, $\rho = 1$ and the score function S in (3.5) satisfy

$$S(v) = \sqrt{c} \sum_{k=1}^{n} \omega_k \text{HBD}_k - C, \tag{5.7}$$

provided $c \neq 0$ and at least one $\omega_k \neq 0$. Here $\text{HBD}_k = \text{HBD}_k(v) = 1_{\{j_{2k-1}(v) = j_{2k}(v)\}}$ is the homozygosity of descent indicator of k and C is a centering constant. For an outbred pedigree, $\rho = 2$ and

$$S(v) = 2 \sum_{1 \leqslant k < l \leqslant n} \omega_{kl} \big((1 - c)\text{IBD}_{kl}/2 + c 1_{\{\text{IBD}_{kl}=2\}}\big) - C, \tag{5.8}$$

provided ω_{kl} is nonzero for at least one pair kl and, again, C is a centering constant.

Notice that ρ in Theorem 1 depends on the pedigree structure. On the other hand, the weights ω_k and ω_{kl} depend on the phenotypes and the genetic model. They determine how various individuals or pairs of individuals should be weighted in the optimal score function.

EXAMPLE 5 (MONOGENIC DISEASES) When there are no polygenic or shared environmental components, we assume

$$P(Y|G) = \prod_{k=1}^{n} P(Y_k|G_k). \tag{5.9}$$

This includes the Gaussian regime class of models with Σ diagonal. McPeek (1999) derived (5.7)–(5.8) for binary traits and pedigrees whose members are affected or have unknown phenotype. In this case (5.7) simplifies to a score function which counts the number of affected individuals that are HBD, i.e. ω_k are identical for all affected individuals. For outbred pedigrees, (5.8) becomes a linear combination of the two score functions S_{pairs}, which sums the number of alleles shared IBD over all pairs of affected individuals (Whittemore and Halpern, 1994) and $S_{\text{g-prs}}$, which sums all pairs of affected individuals that have both alleles IBD. McPeek's results were generalized to arbitrary genetic models in Hössjer (2003d), where it was shown that

$$\omega_{kl} = \omega_k \omega_l, \quad k \neq l. \tag{5.10}$$

For instance, for the Gaussian mixed model of Example 1 without polygenic effects, (5.8) equals the weighted pairwise correlation statistic S_{WPC} of Commenges (1994) when the main locus is additive, i.e. $c = 0$.

Example 6 (Gaussian Mixed Model, Outbred Pedigree) Consider the Gaussian mixed model of Example 1. Assume $\sigma_g^2 = \sigma^2$. It is shown in HS that

$$\begin{aligned} \omega_k &= (r\Sigma^{-1})_k, \\ \omega_{kl} &= (r\Sigma^{-1})_k (r\Sigma^{-1})_l - \Sigma_{kl}^{-1}, \end{aligned} \tag{5.11}$$

where $r = (Y - \mu_0 - \beta\Lambda)/\sigma = (Y - E(Y))/\sigma$ is the standardized vector of residuals and Σ_{kl}^{-1} is the (k, l)th component of Σ^{-1}. If Y_k is unknown, we put $\omega_k = \omega_{kl} = 0$.

For an outbred pedigree $\rho = 2$, hence $\epsilon = \varepsilon^2/2$ can be written as

$$\epsilon = \frac{h^2}{2(1 - h^2)},$$

where $h^2 = \mathrm{Var}(m_{|G_k|})/\mathrm{Var}(Y_k)$ is the heritability of the phenotype at the main locus; when ε is small, $\epsilon \approx h^2/2$.

We denote the score function obtained when inserting (5.11) into (5.8) by S_{wpairs}. The name reflects that S_{wpairs} is a weighted sum of pairwise IBD sharing. The unknown parameters of S_{wpairs} are $(m^*, \sigma^2, c, h_a^2, h_d^2, h_s^2, \beta)$. For additive models we put $c = h_d^2 = 0$, and in absence of polygenic and shared environmental effects we reduce the parameter vector further by letting $h_a^2 = h_s^2 = 0$. This special case of S_{wpairs} is the weighted pairwise correlation statistic S_{WPC}. It contains σ^{-2} as a multiplicative constant which can be dropped. The only remaining parameters to estimate for S_{WPC} are (m^*, β).

One may approximate the exact distribution (4.2) of $Y|v$ by a multivariate normal one. The VC techniques are based on this approximation. Tang and Siegmund (2001), Putter *et al.* (2002) and Wang and Huang (2002) have shown, for nuclear families, that S_{wpairs} is the score obtained with the multivariate normal approximation. Hence, the two approaches are locally equivalent for small ε.

For an inbred pedigree $\rho = 1$. If $q^2u(0) + 2pqu(1) + p^2u(2) = 0$, it follows that

$$E(Y_k) = m^* + F_k\sigma_d\varepsilon + (\beta\Lambda)_k,$$

where $F_k = E(\text{HBD}_k)$ is the inbreeding coefficient of k. This means that $\epsilon = \varepsilon$ is proportional to the change in phenotype mean for all inbred individuals compared to the null model ψ_0. The score function obtained by inserting (5.11) into (5.7) has $(m^*, h_a^2, h_d^2, h_s^2, \beta)$ as parameters that need to be estimated or put to prior values. If we ignore polygenic dominance and shared environmental effects and there are no covariates, we only need to choose (m^*, h_a^2) a priori.

Inbred pedigrees are often used for recessive models. If $u(0) \leqslant u(1) \leqslant u(2)$, it is natural to put the constraint $\varepsilon \geqslant 0$ in (5.4) in order to maintain monotonicity of the three mean parameters $m_0 \leqslant m_1 \leqslant m_2$. Then θ_0 is at the boundary of the parameter space and $Z(t) = W(t)$ is a natural test statistic at locus t. It is also possible to replace the $\varepsilon = 0$ model (m^*, m^*, m^*) in (5.4) by an additive model $(m_0, (m_0 + m_2)/2, m_2)$ with $m_2 > m_0$. The argument leading to (5.7) carries over to this case. If any deviation from additivity is of interest, we put no sign constraint on ε and use $Z(t) = W(t)^2$ as test statistic at locus t.

Example 7 (Gaussian Liability Models) It is shown in HS that for the Gaussian liability model (4.3),

$$\begin{aligned} \omega_k &= \frac{\sigma_g}{\sigma} \int (x\Sigma^{-1})_k P(x|Y)\,\mathrm{d}x, \\ \omega_{kl} &= \frac{\sigma_g^2}{\sigma^2} \int ((x\Sigma^{-1})_k (x\Sigma^{-1})_l - \Sigma_{kl}^{-1}) P(x|Y)\,\mathrm{d}x, \end{aligned} \tag{5.12}$$

where $P(x|Y) \propto P(Y|\sigma x + \mu_0 + \beta\Lambda)P(x)$ is the posterior density of $(X - \mu_0 - \beta\Lambda)/\sigma$ and $P(x)$ is the density of an $N(0, \Sigma)$-distribution.

5.2 *Rare disease models*

In this subsection, we keep the penetrance parameter ψ fixed whereas $p_\varepsilon = \varepsilon$ is a function of ε.

PROPOSITION 2 Assume random mating (5.2) and let e_j and 0 be binary vectors of length $2f$ with e_j having a one in the jth position and zeros elsewhere and 0 having zeros everywhere. Then $\rho = 1$ and

$$S(\upsilon) = \sum_{j=1}^{2f} \frac{P(Y|e_j, \upsilon)}{P(Y|0, \upsilon)} - C \qquad (5.13)$$

whenever the right-hand side is a nonconstant function of υ. The constant C is chosen so that $E_{\theta_0}(S(\upsilon)|Y) = 0$.

Notice that θ_0 is at the boundary of the parameter space because of the constraint $p \geqslant 0$ on the disease allele frequency. Hence, $Z(t) = W(t)$ is the appropriate test statistic to use.

McPeek (1999) derived (5.13) for binary traits and affected pedigree members. The resulting score function S she referred to as S_{robdom}, since it had good and robust performance over a wide range of dominant models. McPeek's result was extended in Hössjer (2003d) to the monogenic model (5.9), whilst (5.13) is a further extension to include polygenic and shared environmental effects.

EXAMPLE 8 (GAUSSIAN MIXED MODELS FOR RARE DISEASES) In Example 1, assume that the pedigree is outbred. Define $b_j = b_j(\upsilon) = (b_{j1}, \ldots, b_{jn})$, where b_{jk} is 1 iff individual k receives the jth founder allele via one of its parents (either $j_{2k-1}(\upsilon)$ or $j_{2k}(\upsilon)$ equals j). Put $K = \exp((m_1 - m_0)/\sigma)$, and let $r = (Y - m_0 - \beta\Lambda)/\sigma$ be a standardized residual vector in the absence of disease alleles ($m = m_0$). Then, inserting $Y|e_j, \upsilon \sim N(m_0 1 + (m_1 - m_0)b_j + \beta\Lambda, \sigma^2\Sigma)$ into (5.13) we arrive at

$$S_{\text{normdom}}(\upsilon) = \sum_{j=1}^{2f} K^{b_j \Sigma^{-1}(r - 0.5\log(K) b_j)'} - C, \qquad (5.14)$$

where C is a centering constant. We use the score function name $S = S_{\text{normdom}}$ introduced in Hössjer (2003d) for the special case $h_a^2 = h_d^2 = h_s^2 = 0$ of no polygenic effects. Notice that m_2 does not enter into S_{normdom}, because for rare disease alleles it is very unlikely that there is more than one disease allele among the founders. Since the pedigree is assumed to have no loops, the disease allele can appear at most once in each individual. The unknown parameters of S_{normdom} are $(K, m_0, \sigma^2, h_a^2, h_d^2, h_s^2)$. Of these, K is most important, since it measures the strength of the major genetic component. For rare disease alleles one has $E(Y_k) \approx m_0 + (\beta\Lambda)_k$ and $V(Y_k) \approx \sigma^2$. This motivates why $m_0 + (\beta\Lambda)_k$ and σ are used for standardizing phenotypes.

6. A SIMULATION STUDY

In this section we investigate various score functions for the Gaussian mixed model of Example 1. For simplicity we do not include covariates and put $\beta = 0$ in (4.5) (with $Y = X$). We assume that the phenotype mean $E(Y_k) = m$ and total variance $V(Y_k) = \sigma_t^2 = \text{Var}(m_{|G_k|}) + \sigma^2$ have been estimated from population data. Here $\text{Var}(m_{|G_k|})$ is the total genetic variance of the major gene while σ^2, defined in (4.5) with $Y = X$, is the sum of all environmental and polygenic variance components. For simplicity, we

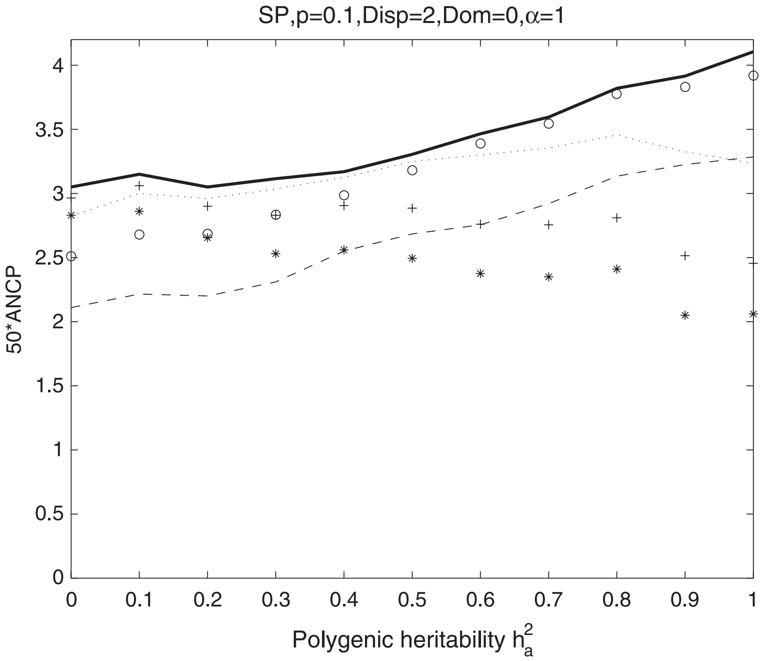

Fig. 1. $50 \cdot$ ANCP, where ANCP is the asymptotic noncentrality parameter, as function of true h_a^2 for optimal (thick solid line), Haseman–Elston (dashed line) and S_{normdom} score functions with $k = 1.5$ and different choices of assumed h_a^2: $h_a^2 = 0$ ($*$), $h_a^2 = 0.2$ (+), $h_a^2 = 0.5$ (dotted line) and $h_a^2 = 0.8$ (o). The number of Monte Carlo iterates is 5000.

assume there are no dominant polygenic or shared environmental effects, i.e. $h_d^2 = h_s^2 = 0$ in (4.7). The four essential unknown genetic model characteristics are then p, h_a^2 and

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

The displacement Disp quantifies the strength and Dom the degree of dominance of the main locus genetic component. Under the mild restriction that m_i are nondecreasing we have Disp $\geqslant 0$ and $-1 \leqslant$ Dom $\leqslant 1$, with Dom taking values -1, 0 and 1 for recessive, additive and dominant models, respectively.

We only consider outbred pedigrees, hence the linkage score function is $Z(t) = W(t)$, i.e. the second row of (3.10) is used. As performance criterion we use the noncentrality parameter, NCP $= E(Z(\tau)|Y)$, the expected value w.r.t. MD and conditional on phenotypes of the linkage score function at the disease locus. This criterion is related to the power $P_{H_1}(Z_{\max} \geqslant T)$ to detect linkage (Feingold *et al.*, 1993), but does not require specification of a threshold T, genome region Ω or significance level $P_{H_0}(Z_{\max} \geqslant T)$. For a genomewide scan, an NCP of about 4 corresponds to significant linkage, although the exact value depends on the collection of pedigrees, the score function, marker informativeness and the genetic model (Lander and Kruglyak, 1995; Ängquist and Hössjer, 2004a).

Assuming complete MD, one has

$$\text{NCP} = \sum_w S(w) P_\theta(v = w|Y) \Big/ \sqrt{2^{-m} \sum_w S(w)^2}, \tag{6.1}$$

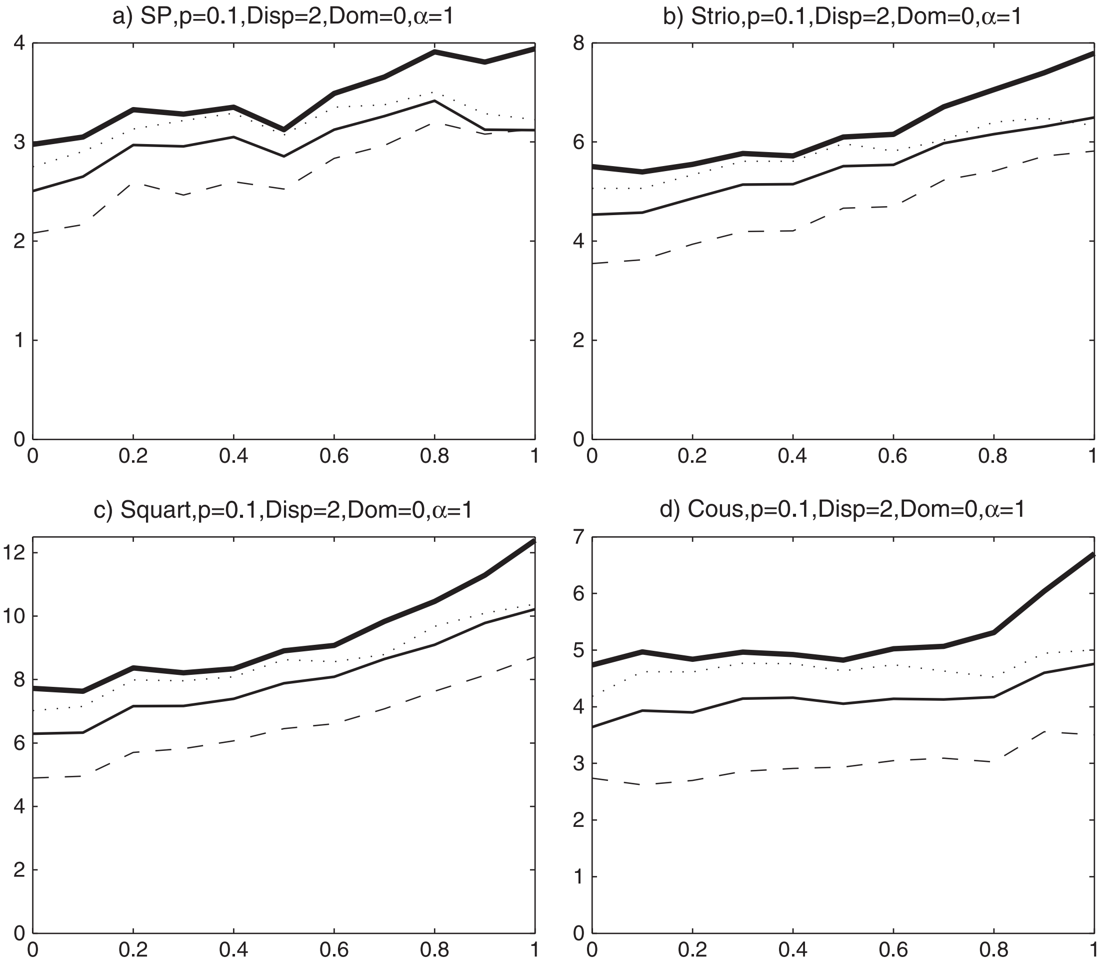

Fig. 2. $50 \cdot$ ANCP as function of true h_a^2 for different score functions: optimal (thick solid line), S_{normdom} with assumed $h_a^2 = 0.5$ and $k = 1.5$ (dotted line), S_{wpairs} with assumed $h_a^2 = 0.5$ (thin solid line) and Haseman–Elston (dashed line). The four subplots correspond to different pedigree structures. The number of Monte Carlo iterates is 5000 (a,b) and 2000 (c,d).

for one pedigree and any centered score function S. Here m is the number of meioses of the pedigree and $\upsilon = \upsilon(\tau)$. For a collection of N pedigrees, the NCP grows at rate $\sqrt{N}$, since

$$\mathrm{NCP} = \sqrt{N}\,\frac{\sum_{i=1}^{N} \gamma_i \mathrm{NCP}_i / N}{\sqrt{\sum_{i=1}^{N} \gamma_i^2 / N}}, \tag{6.2}$$

with NCP_i the NCP and γ_i the weight of the ith pedigree. We choose γ_i as in the denominator of (6.1). For the locally optimal score function (3.5), this weighting scheme is equivalent to (3.9).

If pedigrees (including their phenotypes) are drawn from a population, the second factor of (6.2) converges to $\mathrm{ANCP} = \int \gamma(Y)\mathrm{NCP}(Y)\,\mathrm{d}P(Y) / \sqrt{\int \gamma^2(Y)\,\mathrm{d}P(Y)}$ as N grows, where $\mathrm{d}P(Y)$ is the sampling

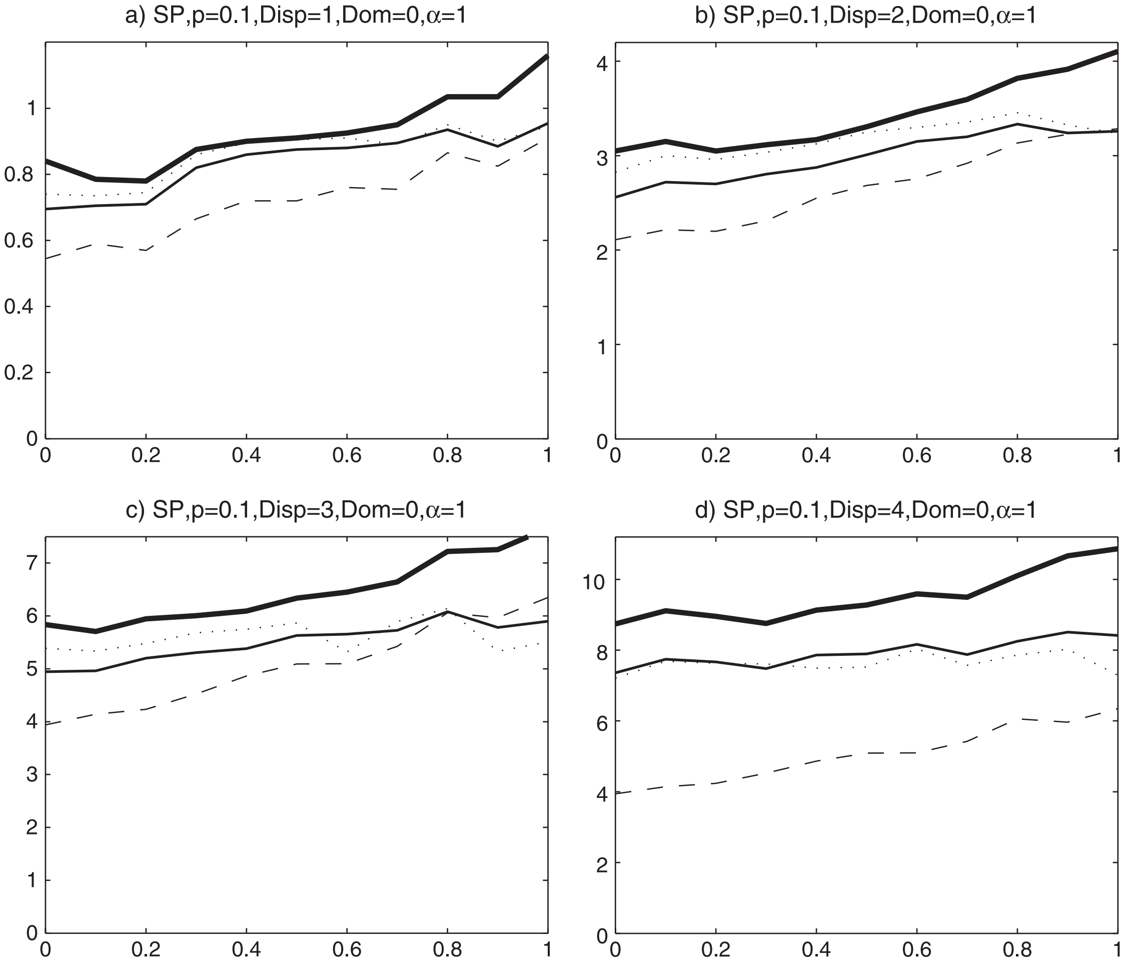

Fig. 3. $50 \cdot$ ANCP as function of true h_a^2 for different score functions: optimal (thick solid line), S_{normdom} with assumed $h_a^2 = 0.5$ and $k = 1.5$ (dotted line), S_{wpairs} with assumed $h_a^2 = 0.5$ (thin solid line) and Haseman–Elston (dashed line). The four subplots correspond to different strengths of the penetrance parameters (Disp). The number of Monte Carlo iterates is 5000.

distribution of pedigrees including their phenotype vectors Y (Hössjer, 2003b,d). Hence,

$$\text{NCP} \approx \sqrt{N}\,\text{ANCP}$$

for large N. When sampling pedigrees, we consider one fixed pedigree structure with certain pedigree members having unknown phenotypes. For the remaining pedigree members, the phenotype vector Y is drawn from the fraction α $(0 < \alpha \leqslant 1)$ of randomly sampled Y $(P_\theta(Y) = \sum_G P_\psi(Y|G)P_p(G))$ with largest weights $\gamma(Y)$. For the locally optimal score function (3.5), this means that a fraction α of the most informative pedigrees are considered, because the weights γ_i are then proportional to $\sqrt{I_i^{\text{compl}}}$, the square roots of the Fisher informations.

Four score functions were included in the simulations, S_{wpairs}, S_{normdom}, S_{HE} and S_{optimal}. Since m and σ_t are assumed to be known, we use the residual vector $r = (Y - m)/\sigma_t$ in the definition of S_{wpairs}

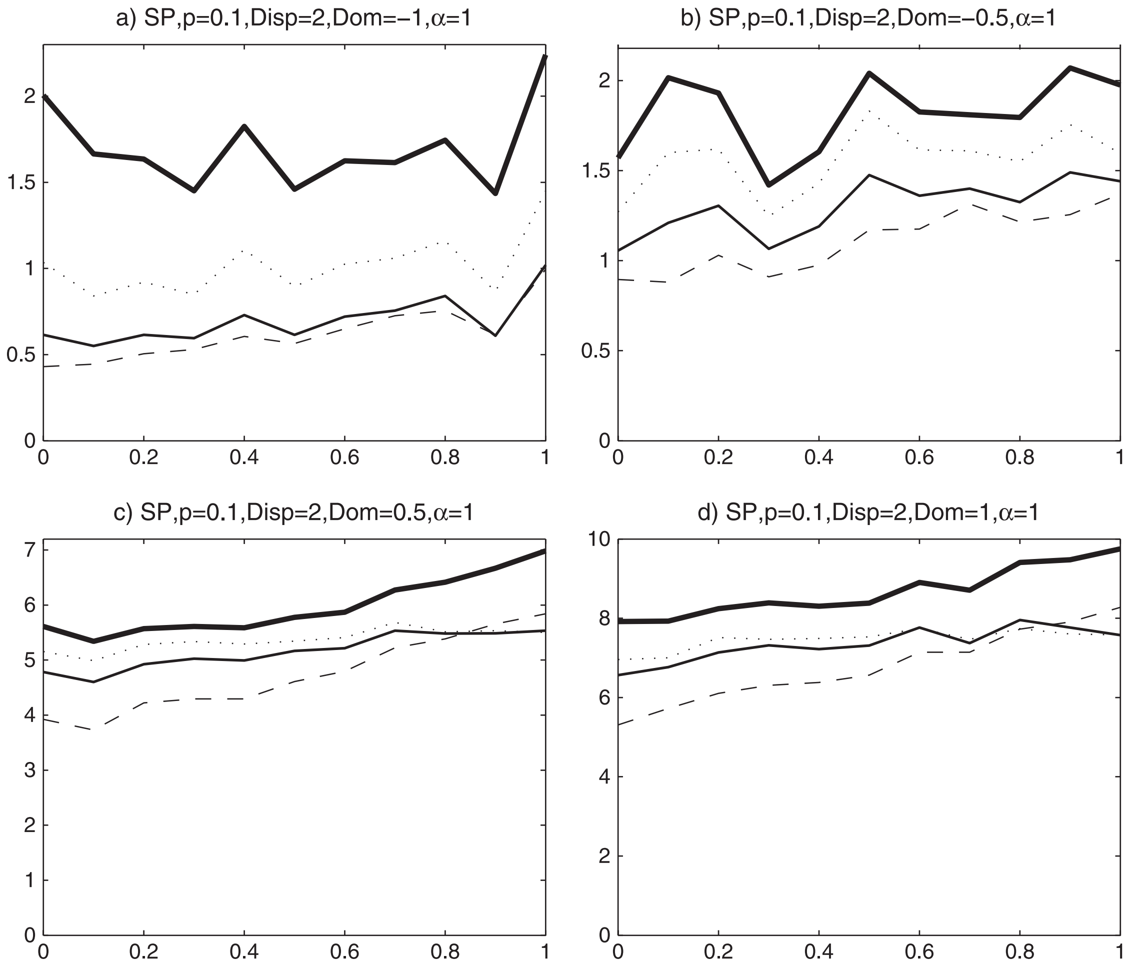

Fig. 4. $50 \cdot$ ANCP as function of true h_a^2 for different score functions: optimal (thick solid line), S_{normdom} with assumed $h_a^2 = 0.5$ and $k = 1.5$ (dotted line), S_{wpairs} with assumed $h_a^2 = 0.5$ (thin solid line) and Haseman–Elston (dashed line). The four subplots correspond to different degrees of dominance (Dom). The number of Monte Carlo iterates is 5000.

in (5.11) and S_{normdom} in (5.14). We also put $c = h_d^2 = h_s^2 = 0$ in the definition of S_{wpairs} and $h_d^2 = h_s^2 = 0$ and $k = 1.5$ in the definition of S_{normdom}. The value $k = 1.5$ yields good and robust performance for a wide range of genetic models. As a score function analogue of the classical Haseman–Elston regression method for quantitative traits we included

$$S_{\text{HE}}(v) = \sum_{k<l} (2\sigma_t^2 - (Y_k - Y_l)^2)\text{IBD}_{kl} - C,$$

see Haseman and Elston (1972) and Hössjer (2003d). Finally, as a benchmark, we also included the optimal (in terms of NCP) score function S_{optimal}, which is the centered version of $P(v|Y)$ (Hössjer, 2003b).

In Figures 1–6 we have plotted 50 × ANCP for complete MD, all four score functions and various genetic models (Disp, Dom, h_a^2, p), pedigrees and sampling fractions α. This corresponds to an NCP of a sample with $N = 2500$. We assume, for simplicity of interpretation, that all families in the populations

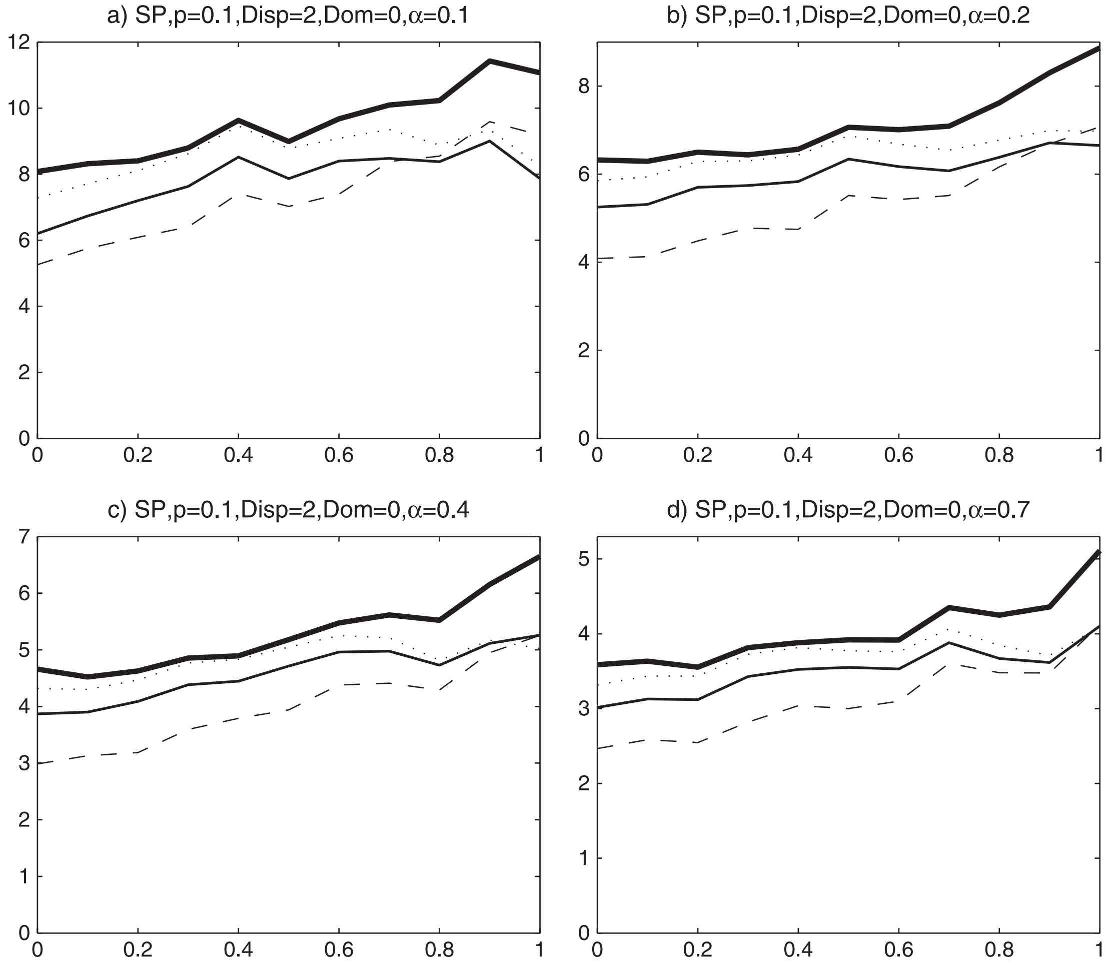

Fig. 5. $50 \cdot$ ANCP as function of true h_a^2 for different score functions: optimal (thick solid line), S_{normdom} with assumed $h_a^2 = 0.5$ and $k = 1.5$ (dotted line), S_{wpairs} with assumed $h_a^2 = 0.5$ (thin solid line) and Haseman–Elston (dashed line). The four subplots correspond to different sampling fractions α. The number of Monte Carlo iterates is 5000.

have the same pedigree structure. We have included four pedigree structures in the simulations—sib pair (SP), sib trio (Strio), sib quartet (Squart) and first cousin (Cous) families. In all cases, all individuals except the two parents of the first generation have known phenotypes.

Notice that ANCP for S_{optimal} is a measure of informativity for the particular combination of pedigree structure, genetic model parameters and sampling fraction. Figures 1–6 show that informativity in general increases with increased polygenic heritability h_a^2, the explanation for which is that deconvolution (recovering G from Y) is easier for dependent errors than for independent ones. Risch and Zhang (1995) noticed, for sib pairs, that residual correlation increases and decreases informativity of discordant and concordant sib pairs, respectively. From our simulation results it is evident that sib correlation, in most cases, *on average*, increases informativity.

Figure 1 shows the effect of varying the assumed h_a^2 of S_{normdom}. It turns out that a value of h_a^2 around 0.5 gives the best overall performance when the true h_a^2 varies between 0 and 1. The same conclusions

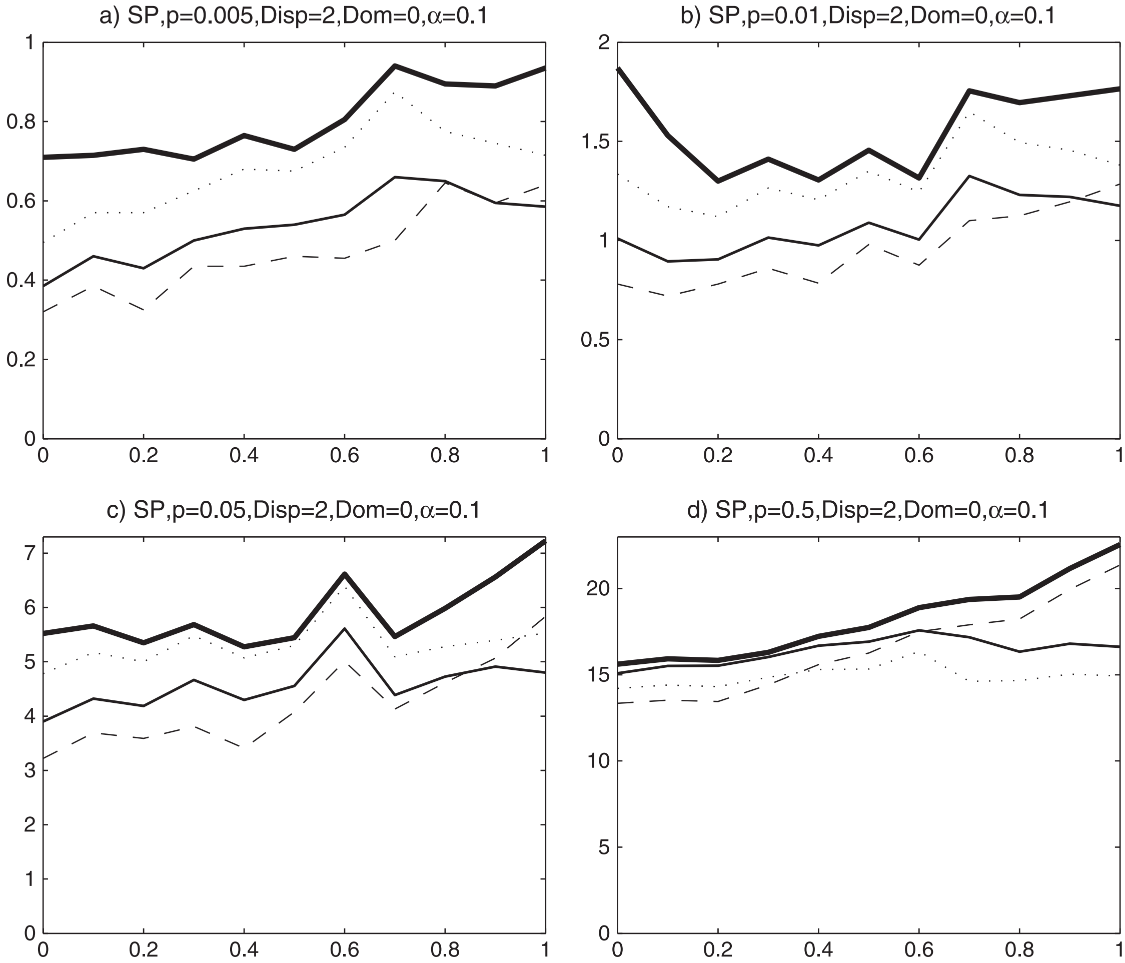

Fig. 6. $50 \cdot$ ANCP as function of true h_a^2 for different score functions: optimal (thick solid line), S_{normdom} with assumed $h_a^2 = 0.5$ and $k = 1.5$ (dotted line), S_{wpairs} with assumed $h_a^2 = 0.5$ (thin solid line) and Haseman–Elston (dashed line). The four subplots correspond to different disease allele frequencies p. The number of Monte Carlo iterates is 5000.

can be drawn for other pedigree structures, genetic models and sampling fractions, both for S_{normdom} and S_{wpairs}. For this reason, we have compared S_{normdom} and S_{wpairs}, with assumed $h_a^2 = 0.5$ with S_{HE} and S_{optimal}. Of the three nonideal score functions, S_{normdom} is best with S_{wpairs} almost as good. The Haseman–Elston score function is competitive for large h_a^2 and large disease allele frequencies p. The optimal version of S_{wpairs}, with $c = \sigma_d^2/\sigma_g^2$, was also included in the simulations (results not shown here). Its performance differed marginally from the $c = 0$ version of S_{wpairs}, even when Dom assumed values -1 or 1, i.e. when the degree of dominance at the main locus was maximal.

7. Conclusions

In this paper, we have presented a general semiparametric framework for choosing score functions in linkage analysis based on local expansions of conditional likelihoods, extending previous work of Whittemore

(1996), McPeek (1999) and Hössjer (2003d). In particular, we have derived score functions for weak penetrance and rare disease models. The methodology is applicable for arbitrary pedigree structures and genetic models for which one major gene, polygenes and shared environmental effects are built into the penetrance function. In particular, we have introduced a wide class of Gaussian liability mixed models, which unifies and extends several models considered in the literature.

The SPL method is asymptotically equivalent to mod scores (and for Gaussian phenotypes also to variance components techniques) under the null hypothesis of no linkage. However, SPL scores are faster to compute than mod scores, since no parameter optimization is needed at each locus. Once the score function S is specified, existing multipoint NPL software such as Genehunter (Kruglyak *et al*., 1996), Allegro (Gudbjartsson *et al*., 2000) or Meurlin (Abecasis *et al*., 2002) can be utilized. These computational savings are important when genomewide p-values are calculated by Monte Carlo, using e.g. the importance sampling algorithm of Ängquist and Hössjer (2004b). Another advantage of SPL is that analytical formulas for genomewide p-values (Feingold *et al*., 1993; Lander and Kruglyak, 1995; Ängquist and Hössjer, 2004a) and confidence regions (Kruglyak and Lander, 1995; Hössjer, 2003a) can easily be adapted from NPL by changing the score function S.

Our simulations for the Gaussian mixed model indicate that incorporating polygenic effects into the score functions leads to improved and robust performance over a wide range of parameters. The rare disease score function S_{normdom} was most powerful, closely followed by the weak penetrance score function S_{wpairs}.

We believe that the semiparametric approach, with a strategic choice of the fixed parameter(s) often leads to procedures with good performance and robustness toward parameter misspecification. When the chosen parameter(s) is not too misspecified, the decreased number of degrees of freedom compared to a fully nonparametric approach can be worthwile. However, more investigations are needed to compare the two approaches in terms of power.

The orthogonal decomposition of functions of genotypes in Section B of HS is of independent interest. It incorporates the classical decomposition of genetic variance into additive and dominant variance components. When applied to local penetrance models, it yields score functions involving HBD sharing for inbred pedigrees and pairwise IBD sharing for outbred pedigrees.

One way of reducing the number of unknown parameters of the score function is to consider a trajectory $\{\theta_\varepsilon\}$, where ε has more than one degree of freedom. The resulting multiparameter score function S is vector-valued, and a linkage score test for weak penetrance models has been derived in Hössjer (2003c).

ACKNOWLEDGMENTS

The author wishes to thank Yudi Pawitan for pointing out valuable references and two referees and the editor for several valuable comments that improved the paper. This research was sponsored by the Swedish Research Council, under contract number 626-2002-6286.

REFERENCES

ABECASIS, G. R., CHERNY, S. S., COOKSON, W. O. AND CARDON, L. R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.

ALMASY, L. AND BLANGERO, J. (1998). Multipoint quantitative trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62**, 1198–1211.

AMOS, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics* **54**, 535–543.

ÄNGQUIST, L. AND HÖSSJER, O. (2004a). Improving the calculation of statistical significance in genome-wide scans. *Biostatistics* (in press).

Ängquist, L. and Hössjer, O. (2004b). Using importance sampling to improve simulation in linkage analysis. *Statistical Applications of Genetics and Molecular Biology* **3**, Article 5.

Bonney, G. D. (1986). Regressive logistic models for familial disease and other binary traits. *Biometrics* **42**, 611–625.

Clayton, D. (2001). Tests for genetic linkage and association with incomplete data. Invited talk, Easter North American Region of Biometrics Society.

Clerget-Darpoux, F., Bonaïti-Pellié, C. and Hochez, J. (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics* **42**, 393–399.

Commenges, D. (1994). Robust genetic linkage analysis based on a score test of homogeneity: the weighted pairwise correlation statistic. *Genetic Epidemiology* **11**, 189–200.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1999). Model-based geostatistics. *Applied Statistics* **47**, 299–350.

Donnelly, K. (1983). The probability that some related individuals share some section of the genome identical by descent. *Theoretical Population Biology* **23**, 34–64.

Elston, R. C. and George, V. T. (1989). Age of onset, age at examination and other covariates in the analysis of family data. *Genetic Epidemiology* **6**, 217–220.

Ewens, W. J. and Shute, N. C. E. (1986). A resolution of the ascertainment problem. I. Theory. *Theoretical Population Biology* **30**, 388–412.

Feingold, E., Brown, P. 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.

Fimmers, R., Seuchter, S. A., Neugebauer, M., Knapp, M. and Baur, M. P. (1989). Identity-by-descent analysis using complete high-resolutions. In Elston, R. C., Spence, M. A., Hodge, S. E. and MacCluer, J. W. (eds), *Multipoint Mapping and Linkage Based on Affected Pedigree Members*, Genetic Analysis Workshop 6. New York: Liss, pp. 123–128.

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

Gudbjartsson, D. F., Jonasson, K., Frigge, M. L. and Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics* **25**, 12–13.

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. (2003a). Assessing accuracy in linkage analysis by means of confidence regions. *Genetic Epidemiology* **25**, 59–72.

Hössjer, O. (2003b). Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. *Annals of Statistics* **31**, 1075–1109.

Hössjer, O. (2003c). Conditional likelihood score functions in linkage analysis. Report 2003:10. Mathematical Statistics, Stockholm University.

Hössjer, O. (2003d). Determining inheritance distributions via stochastic penetrances. *Journal of the American Statistical Association* **98**, 1035–1051.

Kempthorne, O. (1955). *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.

Kruglyak, L. and Lander, E. (1995). High resolution genetic mapping of complex traits. *American Journal of Human Genetics* **56**, 1212–1223.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, **11**, 241–247.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

McPeek, S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology* **16**, 225–249.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics* **61**, 277–318.

Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigree analysis. *American Journal of Human Genetics* **31**, 161–175.

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

Penrose, L. S. (1935). The detection of autosomal linkage in data which consists of brothers and sisters of unspecified parentage. *Annals of Eugenics* **6**, 133–138.

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.

Risch, N. and Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**, 1584–1589.

Tang, H.-K. and Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics* **2**, 147–162.

Thomas, D. C. and Gauderman, W. J. (1996). Gibbs sampling methods in genetics. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds), *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Thompson, E. A. (1974). Gene identities and multiple relationships. *Biometrics* **30**, 667–680.

Todorov, A. A. and Suarez, B. K. (2002). Liability model. In Elston, R., Olson, J. and Palmer, L. (eds), *Biostatistical Genetics and Genetic Epidemiology*. Chichester, UK: Wiley, pp. 430–435.

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.

Weeks, D. and Lange, L. (1988). The affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics* **42**, 315–326.

Whittemore, A. (1996). Genome scanning for linkage: an overview. *Biometrics* **59**, 704–716.

Whittemore, A. and Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics* **50**, 118–127.

[*Received October 8, 2003; first revision June 16, 2004; second revision October 20, 2004; third revision November 29, 2004; accepted for publication December 9, 2004*]