
Mathematical Statistics
Stockholm University

# Conditional Likelihood Score Functions in Linkage Analysis

Ola Hössjer

Research Report 2003:10

ISSN 1650-0377

## Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

## Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2003:10**,
http://www.math.su.se/matstat

# Conditional Likelihood Score Functions in Linkage Analysis

Ola Hössjer*

May 2003

**Abstract**

The purpose of linkage analysis is to map the position of a major gene contributing to or causing a certain disease. In this paper, we develop a general strategy for choosing score functions in linkage analysis. A conditional likelihood for marker data given phenotypes is constructed. Then a trajectory of genetic models $\{\theta\}$ is defined, with 'null' value $\theta_0$ representing no genetic effect of the major gene. A likelihood score function is computed at $\theta_0$, treating alleles of pedigree founders as missing data.

Our methodology generalizes previous work by Whittemore (1996) and Hössjer (2001). It handles incomplete marker data and more or less arbitrary kinds of phenotypes/genetic models (including polygenes and shared environmental effects) and pedigree structures (with or without inbreeding). It is semiparametric in the sense that the trajectory of genetic models typically has fewer degrees of freedom than the total number of genetic model parameters. The 'fixed' parameters have to be chosen by the investigator from e.g. robustness considerations or populations based parameter estimates.

Detailed examples are given for the Gaussian mixed model with polygenic effects. Two score functions are proposed, $S_{\text{wpairs}}$ and $S_{\text{normdom}}$. Their performance is investigated in a simulation study. One conclusion is that inclusion of polygenic effects in the score function increases overall performance for a wide range of genetic models.

KEY WORDS: Founder alleles, conditional likelihood, linkage analysis, missing data, score functions.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ola@math.su.se. Financial support from the Swedish Research Council, contract nr. 626-2002-6286.

# 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 for development of a certain inheritable disease. Disease related quantities (so called phenotypes) and DNA samples (so called genetic markers) are collected for a number of families with aggregation of the disease. Then positions of the genome are sought for where segregation of DNA is correlated with the inheritance pattern of the disease. The reason is that phenotypes represent blurred observations of disease gene expressions (disease alleles), so DNA transmission at the locus of the disease gene is correlated with phenotype segregation. At other positions DNA transmission will either be less correlated or independent of the phenotype inheritance pattern depending on whether the position belongs to the same chromsome as the disease gene or not.

Linkage analysis can be viewed as a combined hypothesis testing and estimation problem. The hypothesis to test is whether the disease gene is located on the chromsome(s) of interest. If so, one constructs a confidence region for the disease locus. If the statistical model for the disease is known, the standard procedure is to use parametric linkage analysis based on log likelihood ratios, so called lod scores (Morton, 1955). For complex diseases, the statistical model is rarely known, and then nonparametric linkage analysis (NPL) is used. See Ott (1999) and Sham (1998) for overviews.

The most widely used statistical model is based on binary phenotypes (affected/unaffected). Recently, statistical methodology in linkage analysis has been developed to deal with phenotypes more generally. For quantitative (continuous) phenotyhpes, the classical regression method was developed by Haseman and Elston (1972). During the last decade, more powerful variance components, regression and score function methods have been proposed, see Feingold (2002) for a recent overview. A general methodology for choosing score functions in linkage analysis was defined by Whittemore (1996). Based on a parametric model for the conditional distribution of marker data given phenotypes, she constructed an efficient score statistic. The analogous profile likelihood approach, where model parameters are estimated at each locus, was considered by Kong and Cox (1997) and is also closely related to mod scores (Risch, 1984 and Clerget-Darpoux et. al., 1986).

Both Whittemore (1996) and Kong and Cox (1997) modelled likelihoods in an empirical way. In McPeek (1999) and Hössjer (2001) the Whittemore score function approach was applied with disease alleles as hidden data. Then the likelihood includes penetrance and disease allele frequency parameters and very general models are allowed for. This includes general pedigree structures

and binary, quantitative, survival and generalized linear model phenotypes. Although the number of parameters is enlarged with disease alleles as hidden data, the unknown parameters of the resulting score function are often fewer. The score function approach was referred to as semiparametric linkage in Hössjer (2001). The nonparametric part comes from those model parameters that are differentiated in the score function, and the parametric part from those parameters that are kept fixed. The latter ones must be supplied by the user, e.g. from estimates using population data.

The purpose of this paper is to extend the work in Hössjer (2001) in several ways: 1) We show how to include the sampling mechanism (ascertainment). It turns out that conditioning on phenotypes can be formulated as a conditional likelihood, so that ascertainment parameters need not be estimated. 2) We extend the perfect marker data approach in Hössjer (2001) to incomplete marker data. In this way, the score functions are natural generalizations of the NPL multipoint score functions proposed by Kruglyak et al. (1996). 3) Shared environmental and polygenic effects are allowed for. 4) We introduce the score statistic of the full likelihood (including disease alleles) to make formulas more transparent. 5) Vector valued score functions are considered.

We apply the results to Gaussian mixed models and obtain score functions that are either new or generalizations of existing ones. Our methodology can also be applied to other phenotypes of exponential family, life length type etc.

The paper is organized as follows: Basic principles of linkage analysis are introduced in Section 2. The conditional likelihood and score functions are defined in Section 3. Then, we show how to compute the score function explicitly in Section 4 for weak penetrance and rare disase models. In Section 5 we generalize the setup to multidimensional trajectories and vector valued score functions. A simulation study is presented in Section 6, followed by conclusions and gathering of more technical results in the appendix.

# 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, so called alleles, exist - the normal (0) and 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 $G = (G_1, \ldots, G_n) = (a_1, \ldots, a_{2n})$. Here $G_k = (a_{2k-1}, a_{2k})$ is the genotype of the $k^{\text{th}}$ individual, with $a_{2k-1}$ and

$a_{2k}$ the paternally and maternally transmitted alleles respectively. Since the gene of interest is unknown, we don't observe $G$, but rather a vector of disease phenotypes $Y = (Y_1, \ldots, Y_n)$. Here $Y_k$ is the phenotype of the $k^{\text{th}}$ individual, a quantity related to the disease. This could be affection status (affected/unaffected) or a quantitative variable such as insuline concentration, body mass index etc.

Alleles are transmitted from parents to children according to Mendel's law of segregation. This means that when a sperm or ova cell is formed (so called meiosis) in a parent, only one of the two alleles of a given gene is passed on. The probability is 0.5 that either the grandpaternal or grandmaternal allele is transmitted. There are $m = 2(n - f)$ meioses in the pedigree, since each nonfounder receives two alleles, one from each parent. At a certain locus $t$, allele transmission can be summarized through the inheritance vector $v(t) = (v_1(t), \ldots, v_m(t))$, where $v_k(t)$ equals zero or one depending on whether a grandpaternal or grandmaternal allele was transmitted during the $k^{\text{th}}$ meiosis. A priori, without phenotype information we have

$$P(v(t) = w) = 2^{-m} \tag{1}$$

for all possible binary vectors $w$ of length $m$. This reflects Mendelian segregation and the fact that allele transmissions are independent for different meioses.

The purpose of linkage analysis is to test whether the disease gene is located on the genomic region of interest, and if so, to produce a confidence region for it's location. By determining the DNA content of pedigree members, using so called genetic markers, it is possible to estimate $v(t)$ at all loci $t$ of interest. Each marker is a small segment of DNA of known position which can have different forms (alleles) for different individuals. Marker alleles are registered at a number of positions along the chromosome for all pedigree members that are genotyped. If $t$ is close to a highly polymorphic marker (one with many possible alleles) and sufficiently many pedigree members are genotyped, $v(t)$ is essentially determined by data from that marker. In other cases, information about $v(t)$ is determined from several surrounding markers.

If $\tau$ is the disease locus, the phenotype vector $Y$ will be correlated to $v(\tau)$ if the genetic component is strong enough. This means that the conditional distribution

$$w \longrightarrow P(v(t) = w|Y) \tag{2}$$

differs from the prior (1) at $t = \tau$. The stronger the genetic component at $\tau$ is, the stronger is the discrepancy between $P(v(\tau)|Y)$ and the uniform distribution. Even at loci $t$ around $\tau$ (2) will differ from (1). This

is because of occurence of so called crossovers - random points along the chromosome where, during meioses, segregation switches between grandmaternal and grandpaternal transmission. As a result, allele transmissions at nearby loci are correlated, and this correlation decays with genetic distance between the loci. Such a distance is defined in terms of the average number of crossovers. If two loci are at distance $x$ centiMorgans (cM), there is on average $0.01x$ crossovers between the two loci during each meiosis.

The purpose of statistical linkage analysis is the following: A) To detect if there is any $\tau$, i.e. any locus $t$ such that (2) differs significantly from (1). B) Given that there is a $\tau$, estimate it's position. Notice that A) is a hypothesis testing problem whereas B) is an estimation problem.

# 3 Conditional Likelihood and Score Functions

In this section we extend the score function approach of Whittemore (1996) to i) include ascertainment ii) allow for different pedigree structures in families. Consider a chromosome of length $T$ cM with at most one disese locus $\tau$ located along $[0, T]$. At our disposal we have marker data (MD), phenotype data ($Y$) and the fact that the pedigree has been acertained (asc).

The distribution $P_\theta(Y)$ depends on genetic model parameters $\theta$. The probability $P_\pi(\text{asc}|Y)$ of ascertainment given phenotypes depends on additional sampling parameters $\pi$. The statistical parameters are $(t, \theta, \pi)$, where $t$ corresponds to the hypothesis that $\tau$ is positioned at $t$. The parameter of interest is $t$, whereas $\theta$ and $\pi$ are nuiscance parameters. The hypothesis testing problem for a single chromosome is

$$\begin{aligned} H_0: &\quad \tau \text{ is located on another chromosome} \\ H_1: &\quad \tau \in [0, T]. \end{aligned}$$

The probability of observed data is

$$P_{t,\theta,\pi}(\text{MD}, Y, \text{asc}) = P_{t,\theta}(\text{MD}|Y) P_{\theta,\pi}(Y, \text{asc}). \tag{3}$$

Here we assumed that marker data and ascertainment are conditionally independent given phenotypes. In general, it is difficult to model the ascertainment procedure. Therefore, it is common to use the conditional likelihood $P_{t,\theta}(\text{MD}|Y)$ for inference, cf. e.g. Winter (1980), Ewens and Shute (1986) and Elston (1989).

Let $v = v(\tau)$ be the inheritance vector at the disease locus. Then

$$\begin{aligned} P_{t,\theta}(\text{MD}|Y) &= \textstyle\sum_v P_t(\text{MD}|v) P_\theta(v|Y) \\ &= 2^m P(\text{MD}) \textstyle\sum_v P_t(v|\text{MD}) P_\theta(v|Y), \end{aligned} \tag{4}$$

where $\sum_v P_t(\mathrm{MD}|v)P_\theta(v|Y)$ is short for for $\sum_w P_t(\mathrm{MD}|v=w)P_\theta(v=w|Y)$. In the last equality of (4) we applied Bayes' rule and $P(v)=2^{-m}$. The factor $P(\mathrm{MD})$ depends on marker allele frequencies and genetic distances between the markers. These are assumed to be known in linkage analysis. Since $2^m P(\mathrm{MD})$ is independent of $t$ and $\theta$ it is a fixed constant which we drop. Hence

$$L(t,\theta;\mathrm{MD}) = \sum_v P_t(v|\mathrm{MD})P_\theta(v|Y) = E_t\left(P_\theta(v|Y)|\mathrm{MD}\right) \tag{5}$$

is our conditional likelihood. Notice that we only included MD as argument of $L$. The reason is that we condition on the phenotype vector $Y$ and consider it as fixed. Kruglyak et al. (1996) noticed that the conditional inhertiance distribution $P_t(v|\mathrm{MD})$ could be used both in nonparametric and parametric linkage analysis. Assuming no crossover interference, they applied a Hidden Markov algorithm for its computation, originally proposed in the genetics context by Lander and Green (1987).

One approach for testing $H_0$ against $H_1$ and then estimating $\tau$ is to compute a profile conditional likelihood $\sup_\theta L(t,\theta)$. Its maximum is then used as test statistic and its argmax as an estimator of $\tau$. In this paper, we will take a local approach instead. Assume that $\{\theta_\varepsilon\}$ is a one-dimensional trajectory of genetic model parameters such that $\theta_0$ corresponds to no genetic effect at the disease locus, i.e. $P_{\theta_0}(v|Y)=2^{-m}$. In other words, under $\theta_0$ 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. $\varepsilon$ at each locus $t$. In order to do so, it is convenient to start with perfect marker data.

Under perfect marker data, we can unambigously determine $\{v(t)\}$ at all loci[1]. We let $\mathrm{MD}^{\mathrm{perf}} = \{v(t);\, 0 \le t \le l\}$ denote perfect marker data. The corresponding conditional likelihood is

$$L(t,\theta;\mathrm{MD}^{\mathrm{perf}}) = P_\theta(v=v(t)|Y). \tag{6}$$

Notice that $v(t)$ is treated as a constant in (6) whereas $v$ is random. The reason is that $P_t(v|\mathrm{MD}^{\mathrm{perf}})$ has a one point distribution with point mass at $v(t)$ so the sum in (5) just contains one term. Define the score function

$$S(v) = \left.\frac{d^\rho \log P_{\theta_\varepsilon}(v|Y)}{d\varepsilon^\rho}\right|_{\varepsilon=0}, \tag{7}$$

[1]To be precise, $v(t)$ can only be determined up to uncertainty of switching the paternal and maternal alleles of each founder, assuming that the ordering of alleles in a genotype is unimportant. However, all likelihoods are invariant w.r.t. this uncertainty. Therefore, it is no loss of generality to assume that $\{v(t)\}$ is known for perfect marker data. Cf. Kruglyak et al. (1996) and Hössjer (2003a) for more details.

where $\rho$ is the smallest positive integer such that the right-hand side of (7) is nonzero for at least one $v$. When $\rho \geq 2$ the estimation problem is singular (with zero Fisher information) for $\varepsilon$ at $\varepsilon = 0$. By means of the reparametrization

$$\epsilon = \varepsilon^\rho / \rho!, \tag{8}$$

$S$ is interpreted as a (conditional) likelihood score function for $\epsilon$ at $\epsilon = 0$. Depending on the application, there may or may not be a sign constraint on $\varepsilon$. In any case, we are only interested in testing $\varepsilon = 0$ against $\varepsilon \neq 0$ at each locus $t$ - the sign of $\varepsilon$ is not of interest. For this reason, we may use $\epsilon$ instead of $\varepsilon$ as parameter when formulating a test statistic, even when $\rho$ is even. The Fisher information for $\epsilon$ is $I^{\text{perf}} = E_{\theta_0}(S^2(v)) = 2^{-m} \sum_w S^2(w)$ 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 $W^{\text{perf}}(t) = S(v(t))/\sqrt{I^{\text{perf}}}$ at locus $t$.

For non-perfect marker data, the observed conditional likelihood $L(t, \theta; \text{MD}) = E_t(L^{\text{perf}}(t, \theta; \text{MD}^{\text{perf}})|\text{MD})$ is obtained by averaging $L^{\text{perf}}$, treating $\text{MD}^{\text{perf}}$ as hidden data, cf. (5). The square root of the likelihood score test now becomes

$$W(t) = \frac{[d^\rho \log L(t, \theta_\varepsilon; \text{MD})/d\varepsilon^\rho]_{\varepsilon=0}}{\sqrt{I(t)}} = \frac{\sqrt{I^{\text{perf}}}}{\sqrt{I(t)}} E\left(W^{\text{perf}}(t)|\text{MD}\right) \tag{9}$$

where $I(t) = E_{\theta_0} E^2(S(v(t))|\text{MD})$ is the Fisher information. It depends on $t$ in a way that reflects positioning and informativity of markers. The outer expectation in the definition of $I(t)$ is w.r.t. variations in MD and typically involves simulations for computation. By definition of $W(t)$ we have

$$\begin{array}{rcl} E_{\theta_0}(W(t)) & = & 0 \\ V_{\theta_0}(W(t)) & = & 1, \end{array} \tag{10}$$

where both expectations are w.r.t. variations in marker data. The first equation of (10) is a general property of likelihood score statistics (under mild regularity conditions) and the second follows because of the normalization with $\sqrt{I(t)}$.

So far, we have only discussed one single family. The extension to $N$ pedigrees with mutually independent phenotype/marker data is straightforward. We allow the pedigree structures to vary arbitrarily and index quantities for for the $i^{\text{th}}$ 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. Further, 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

$$W(t) = \frac{[d^\rho \log L(t, \theta_\varepsilon; \mathrm{MD})/d\varepsilon^\rho]_{\varepsilon=0}}{\sqrt{I(t)}} = \sum_{i=1}^{N} \gamma_i(t) W_i(t), \tag{11}$$

where $W_i(t)$ is the score (9) for one family and $\gamma_i(t) = \sqrt{I_i(t)/\sum_{j=1}^{N} I_j(t)}$ are the weights assigned to different families. Since $\sum_1^N \gamma_i^2(t) = 1$, (10) holds for the total linkage score with $N$ families.

Using results of Rotnitzky et. al. (2000), we may expand the conditional log likelihood ratio

$$\begin{array}{rcl} \log(L(t, \theta_\varepsilon)/L(t, \theta_0)) & \approx & \frac{1}{\rho!}\sqrt{I(t)} W(t) \varepsilon^\rho - \frac{1}{2(\rho!)^2} I(t) \varepsilon^{2\rho} \\ & = & \sqrt{I(t)} W(t) \epsilon - \frac{1}{2} I(t) \epsilon^2 \end{array} \tag{12}$$

for all $\varepsilon$ at distance $O(N^{-1/(2\rho)})$ from $\varepsilon = 0$. This requires that the conditional likelihood $L(t, \theta_\varepsilon)$ is sufficiently regular as a function of $\varepsilon$. Notice that all terms with powers $\varepsilon^{\rho+1}, ..., \varepsilon^{2\rho-1}$ are asymptotically neglible in (12) (this follows because they have mean zero).

Let $\hat{\theta}(t)$ be the profile likelihood estimate of $\theta$ at $t$, i.e. the maximizer of $L(t, \theta_\varepsilon)$ w.r.t. $\varepsilon$. It follows from (12) that

$$\log \frac{L(t, \hat{\theta}(t))}{L(t, \theta_0)} \approx \begin{cases} 0.5 W(t)^2, & \text{no sign constraint on } \epsilon \\ 0.5 \max(0, W(t))^2, & \epsilon \geq 0, \end{cases} \tag{13}$$

under $H_0$. Notice that $\epsilon = 0$ is at the boundary of the parameter space when $\epsilon \geq 0$. This is always the case when $\rho$ is even and sometimes, depending on the application, when $\rho$ is odd (in that case $\varepsilon \geq 0$ is imposed). The right hand side of (13) is a score test for the parameter $\epsilon$. The normalization by expected rather than observed Fisher information is advantageous because $\varepsilon = 0$ is a singular point when $\varepsilon$ is regarded as the parameter (Bottai, 2003).

Motivated by (13), we use

$$Z(t) = \begin{cases} W(t)^2, & \text{no sign constraint on } \epsilon \\ W(t), & \epsilon \geq 0. \end{cases} \tag{14}$$

as as local test statistic for testing $H_0$ against the simple alternative hypothesis $\tau = t$. That is, we replaced $\max(0, W(t))^2$ by $W(t)$ when $\epsilon \geq 0$. Since only large positive values of $W(t)$ are of interest in this case, the two statistics are essentially equivalent.

When testing $H_0$ against $H_1$, we use the global test statistic

$$Z_{\max} = \sup_{0 \le t \le T} Z(t).$$

The null hypothesis is rejected when $Z_{\max}$ exceeds a given threshold, which depends on the chosen significance level and $T$. (Or more generally, the size of the genomic region tested, which might consist of several chromosomes.) This is discussed e.g. by Feingold et al. (1993), Lander and Kruglyak (1995) and Ängquist and Hössjer (2003). By a central limit thorem argument, $W$ is approximated by a Gaussian process for large $N$. This process has mean zero and unit variance and is stationary for perfect marker data. When $N$ is small and/or there are large pedigrees in the data set, it is important to correct for non-Gaussianity as discussed by Ängquist and Hössjer (2003).

Given $H_1$, confidence regions for $\tau$ can be constructed as the set of loci $t$ for which $Z_{\max} - Z(t)$ is smaller than a given value. See Kruglyak and Lander (1995), Dupuis and Siegmund (1999) and Hössjer (2003b) for details.

# 4 Choosing Score Functions

In this section we discuss computation of the score function $S$ in (7) for a given pedigree. To begin with, it is mathematically more convenient to work with

$$P_\theta(Y|v) = 2^m P_\theta(v|Y) P_\theta(Y) \tag{15}$$

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

$$P_\theta(Y|v) = \sum_G P_\psi(Y|G) P_p(G|v) \tag{16}$$

can be expanded by summing over all possible genotype configurations in the pedigree. In (16), we split $\theta = (p, \psi)$ into $p$, the frequency (probability) of the disease allele, and $\psi$, the penetrance parameter(s). The latter describe the relationship between phenotypes and genotypes. The following result implies that it suffices to consider score functions of $P_\theta(Y|v)$:

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

$$S(v) = \bar{S}(v) - E_{\theta_0}(\bar{S}),$$

*where* $E_{\theta_0}(\bar{S}) = 2^{-m} \sum_w \bar{S}(w)$.

Notice that $P_p(G|v)$ appears in (16). The genotype configuration $G$ in the pedigree is uniquely determined by all founder alleles and the inheritance vector. This is so, since the inheritance vector specifies how the $2f$ founder allels are transmitted through the pedigree. If the founders are numbered as the first $f$ individuals in the pedigree then $a = (a_1, \ldots, a_{2f})$ is a binary vector of length $2f$ containing the founder alleles. The ones correspond to disease alleles and the zeros to normal alleles. We let $J(w) = (j_1(w), \ldots, j_{2n}(w))$ be the gene-identity-state (Thompson, 1974) of the pedigree for the inheritance vector $w$. This means that $j_k(w) \in \{1, \ldots, 2f\}$ gives the number of the founder alleles that have been transmitted to allele number $k$. Then

$$G = G(a, v) = a_{J(v)} = (a_{j_1(v)}, \ldots, a_{j_{2n}(v)}).$$

If $a$ and $v$ are independent (no segregation distortion) we have $P_p(G|v) = P_p(a)$. Then (16) can be written as

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

where the sum ranges over all $2^{2f}$ possible founder allele vectors $a$ and the last expectation is w.r.t. $a$. Under random mating, the components of $a$ are independent, so that

$$P_p(a) = p^{|a|}q^{2f-|a|}, \tag{17}$$

where $q = 1 - p$ and $|a| = \sum_1^{2f} a_j$. Viewing $a$ as hidden data, $P_\psi(Y|a, v)$ is the complete likelihood corresponding to $P_\theta(Y|v)$.

**Example 1 (Gaussian mixed model.)** The canonical example of this paper is a Gaussian mixed model. The genetic influence is a mixture of one major gene and a large number of so called polygenes, which are unlinked to the major gene (i.e. located on other chromosomes). The goal of linkage analysis is to map the major gene, but take the polygenes into account in order to increase statistical efficiency. See Lynch and Walsh (1998) for detailed treatment of such models.

The major genes have a disecrete influence on the phenotype, whereas polygenes and environmental factors give continuous phenotypic variation. If $G_k$ is the major gene's genotype for the $k^{\text{th}}$ individual in the peidgree, we assume $Y_k|G_k \in N(m_{|G_k|}, \sigma^2)$. Here $|G_k| = a_{2k-1} + a_{2k}$ is the number of disease alleles of $G_k$ and $m_0, m_1, m_2$ and $\sigma^2$ are given parameters. If large values of the phenotype indicates disease a natural constraint is $m_0 \leq m_1 \leq m_2$. In vector form we have

$$Y|G \in N(\mu, \sigma^2\Sigma),$$

where $G$ is the set of genotypes at the main locus for all pedigree members and $\mu = (m_{|G_1|}, \ldots, m_{|G_n|})$. Correlations between individuals are caused by polygenic or shared environmental effects and are contained in the $n \times n$ matrix $\Sigma$. For simplicity, we assume there are no shared environmental effects.

Two alleles $k$ and $l$ $(1 \leq k, l \leq 2n)$ at a given locus are identical by descent (IBD) if they originate from the same founder allele, i.e. $j_k(w) = j_l(w)$, where $w$ is the inheritance vector describing segregation at the locus of interest. The number of alleles that two individuals $k$ and $l$ $(1 \leq k, l \leq n)$ share IBD is denoted $\text{IBD}_{kl} = \text{IBD}_{kl}(w)$. It is either 0, 1 or 2. If there is no inbreeding (loops) in the pedigree, it is defined as

$$\text{IBD}_{kl} = 1_{\{j_{2k-1}=j_{2l-1}\}} + 1_{\{j_{2k-1}=j_{2l}\}} + 1_{\{j_{2k}=j_{2l-1}\}} + 1_{\{j_{2k}=j_{2l}\}}. \tag{18}$$

Put $r_{kl} = E(\text{IBD}_{kl}(w))/2$ and $\delta_{kl} = P(\text{IBD}_{kl}(w) = 2)$, where expectation and probability is taken w.r.t. a uniform prior (1) on the inheritance vector $w$. Here $r_{kl}$ is referred to as the coefficient of relationship between $k$ and $l$. With these definitions, the correlation matrix with polygenic effects becomes

$$\Sigma = (1 - h_a^2 - h_d^2)I_n + h_a^2 R + h_d^2 \Delta,$$

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. These are the fractions of total environmental and polygenic variance $(\sigma^2)$ due to additive and dominant genetic effects respectively. The penetrance parameter of a Gaussian mixed model can be written as

$$\psi = (m_0, m_1, m_2, \sigma^2, h_a^2, h_d^2),$$

with the first three components due to the major gene and the last three caused by polygenic/environmental effects. □

## 4.1 Local Penetrance Models

In this subsection, we assume the disease allele frequency $p$ is fixed whereas the penetrance parameters $\psi_\varepsilon$ vary with $\varepsilon$ so that $P_{\psi_0}(Y|a, v) = P_{\psi_0}(Y)$ is independent of $a$ and $v$. This implies no genetic effect at the disease locus when $\varepsilon = 0$. We let

$$\tilde{S}_k(a, v) = \left. \frac{d^k \log P_{\psi_\varepsilon}(Y|a, v)}{d\varepsilon^k} \right|_{\varepsilon=0},$$

represent the likelihood score functions of various orders $k = 1, 2, \ldots$ at $\varepsilon = 0$ for the full data set (including founder alleles).

**Proposition 2** *For a weak penetrance model, the conditional likelihood score function is given by*

$$S(v) = \begin{cases} E(\tilde{S}_1(a,v)) - C_1, & \text{if } \rho = 1, \\ E(\tilde{S}_1^2(a,v)) + E(\tilde{S}_2(a,v)) - C_2, & \text{if } \rho = 2. \end{cases} \tag{19}$$

*Expectations in (19) are w.r.t. $a$ and $C_1$ and $C_2$ are centering constants chosen so that $E_{\theta_0}(S) = 2^{-m} \sum_w S(w) = 0$.*

**Example 2 (Gaussian mixed model, contd.)** Consider a fixed model with first three penetrance parameters $(m_0^*, m_1^*, m_2^*)$. Under random mating, the expected value of the phenotype is $E(Y_k) = m = \sum_{i=0}^2 E(Y_k||G_k| = i) = q^2 m_0^* + 2pqm_1^* + p^2 m_2^*$. The genetic variance due to the major gene is $\sigma_g^2 = q^2(m_0^* - m)^2 + 2pq(m_1^* - m)^2 + p^2(m_2^* - m)^2$. Consider a one parameter family of penetrance vectors

$$\psi_\varepsilon = (m+(\sigma/\sigma_g)\varepsilon(m_0^*-m), m+(\sigma/\sigma_g)\varepsilon(m_1^*-m), m+(\sigma/\sigma_g)\varepsilon(m_2^*-m), \sigma^2, h_a^2, h_d^2)$$

for $\varepsilon \geq 0$. The mean $m = E(Y_k)$ is independent of $\varepsilon$ whereas the genetic variance at the main locus is $\varepsilon^2\sigma^2$. Hence the heritability at the main locus, defined as the ratio of genetic variance and total variance is $\varepsilon^2/(1+\varepsilon^2)$. Now

$$Y|G = Y|a, v \in N(\mu_\varepsilon, \sigma^2\Sigma), \tag{20}$$

where $\mu_\varepsilon = m1 + \sigma\varepsilon u$, $1 = (1, \ldots, 1)$ is a row vector of $n$ ones, $u = u(a,v) = (u_1, \ldots, u_n)$ and $u_k = u_k(a,v) = (m^*_{a_{j_{2k-1}(v)} + a_{j_{2k}(v)}} - m)/\sigma_g$.

We use Proposition 2 for computing the score function $S$. By (20) we have $\log P_{\psi_\varepsilon}(Y|a,v) = -0.5n\log(2\pi) - -0.5n\log\sigma - 0.5\log|\Sigma| - 0.5\sigma^{-2}(Y - \mu_\varepsilon)\Sigma^{-1}(Y-\mu_\varepsilon)'$, where $'$ denotes vector transposition. Differentiating twice w.r.t. $\varepsilon$ we get

$$\begin{aligned} \tilde{S}_1(a,v) &= u\Sigma^{-1}r', \\ \tilde{S}_2(a,v) &= -u\Sigma^{-1}u, \end{aligned} \tag{21}$$

where $r = (Y - m)/\sigma$ is the vector of standardized residuals. The next step is to use (19) and average (21) w.r.t. $a$. This requires computation of $E(u_k)$ and $E(u_k u_l)$, where expectation is taken over $a$ but not $v$. For an outbred pedigree, we have $E(u_k) = 0$ for all $v$ and

$$E(u_k u_l) = (1-c)\text{IBD}_{kl}/2 + c1_{\{\text{IBD}_{kl}=2\}}. \tag{22}$$

Here $c = \sigma_d^2/\sigma_g^2$ is the fraction of main locus genetic variance due to dominance effects for the reference model. It is based on splitting the total genetic variance $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$ at the main locus into additive and dominance components $\sigma_a^2 = 2pq(p(m_2^* - m_1^*) + q(m_1^* - m_0^*))^2$ and $\sigma_d^2 = (pq)^2(m_2^* - 2m_1^* + m_0^*)^2$. See the appendix for a motivation of (22). Applying Proposition 2 with $\rho = 2$ we obtain

$$S_{\text{wpairs}}(v) = \sum_{k,l}(\omega_k\omega_l - \Sigma_{kl}^{-1})\left((1-c)\text{IBD}_{kl}/2 + c1_{\{\text{IBD}_{kl}=2\}}\right) - C, \qquad (23)$$

from (21) and (22). The name $S = S_{\text{wpairs}}$ reflects that this score function is a weighted sum of pairwise IBD sharing, where $\omega_k$ is the $k^{\text{th}}$ component of $r\Sigma^{-1}$, $\Sigma_{kl}^{-1}$ the $(k,l)^{\text{th}}$ component of $\Sigma^{-1}$ and $C$ a centering constant. The diagonal terms of $S_{\text{wpairs}}$ are independent of $v$. Hence they can be absorbed into $C$. By symmetry, it suffices to sum over indeces $k < l$. The unknown parameters of $S_{\text{wpairs}}$ are $(m, \sigma^2, c, h_a^2, h_d^2)$. For additive models we put $c = h_d^2 = 0$ and in abscence of polygenic effects we reduce the parameter vector further by letting $h_a^2 = 0$. This special case of $S_{\text{wpairs}}$ is the weighted pairwise correlation statistic ($S_{\text{WPC}}$) of Commenges (1994). It contains $\sigma^{-2}$ as a multiplicative constant which can be dropped. The only remaining parameter to be estimated for $S_{\text{WPC}}$ is $m$.

If the phenotype $Y_k$ is unknown for some pedigree members derivation of $S$ is analogous, with $Y$ and $\Sigma$ containing entries from known phenotype members only. In particular, the sum in (23) is taken over pairs of pedigree members with known phenotypes. □

The traditional score function approach for Gaussian mixed models is to use the simplifying assumption $Y|v \in N(m1, \Sigma_\varepsilon)$, where $\Sigma_\varepsilon = \sigma^2(\Sigma + \varepsilon^2\Lambda)$ and $\Lambda = \Lambda(v)$ is a matrix with entries $(1-c)\text{IBD}_{kl}/2 + c1_{\{\text{IBD}_{kl}=2\}}$ representing interactions at the main locus. By treating the multivariate Gaussian approximation of $Y|v$ as the true likelihood $P_{\theta_\varepsilon}(Y|v)$ it can be shown that the same score function $S$ is obtained as in (23). One first computes the score $\bar{S}$ and then uses Proposition 1 for centering. This has been done for nuclear families with parents and siblings by Tang and Siegmund (2001), Putter et al. (2002) and Wang and Huang (2002).

**Example 3 (Gaussian mixed model with inbreeding.)** We now assume there is inbreeding in the pedigree. Let $\sigma_d = pq(m_2^* - 2m_1^* + m_0^*)$ and $\kappa = \sigma_d/\sigma_g$. Notice that $\kappa^2 = c$, and $\kappa$ is negative, zero or positive for a dominant, additive or recessive reference model. If individual $k$ receives

both of it's alleles from the same founder ($j_{2k-1} = j_{2k}$), the two alleles are homozygous by descent (HBD). We let $\text{HBD}_k = \text{HBD}_k(v)$ be one if this is the case and zero otherwise. It is shown in the appendix that

$$E(u_k) = \kappa \cdot \text{HBD}_k. \tag{24}$$

Combining (21) with Proposition 2 we find that $\rho = 1$ and

$$S(v) = \kappa \sum_k \omega_k \text{HBD}_k - C, \tag{25}$$

where $C$ is a centering constant. Notice that $S$ contains multiplicative constants $\kappa$ and $\sigma^{-1}$ which can be removed. The parameters that need to be estimated or put to prior values are $(m, h_a^2, h_d^2)$.

If the reference model is recessive, it is natural to put the constraint $\varepsilon \geq 0$ in order to maintain monotonicity of the three mean parameters. Then $\theta_0$ 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 have an additive model with mean parameters $(m_0, (m_0 + m_2)/2, m_2)$ and $m_2 > m_0$ when $\varepsilon = 0$. The argument leading to (25) carries over to this case. If any deviation from additivity is of interest, we put no sign constraint on $\varepsilon$ and use $Z(t) = W(t)^2$ as test statistic at locus $t$. □

## 4.2 Rare Disease Models

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

**Proposition 3** *Assume random mating (17) and let $e_j$ and 0 be binary vectors of length $j$ with $e_j$ having a one in the $j^{th}$ position and zeros elsewhere and 0 having zeros everywhere. Then $\rho = 1$ and*

$$S(v) = \sum_{j=1}^{2f} \frac{P(Y|e_j, v)}{P(Y|0, v)} - C \tag{26}$$

*whenever thr right hand side is a non-constant function of $v$. The constant $C$ is chosen so that $E_{\theta_0}(S) = 0$.*

Notice that $\theta_0$ is at the boundary of the parameter space because of the constraint $p \geq 0$ on the disease allele frequency.

**Example 4 (Gaussian mixed models for rare diseases.)** Assume that the pedigree is outbred. Define $b_j = b_j(v) = (b_{j1}, ..., b_{jn})$, where $b_{jk}$ is one iff individual $k$ receives the $j^{\text{th}}$ founder allele via one of it's parents (either $j_{2k-1}(v)$ or $j_{2k}(v)$ equals $j$). Put $K = \exp((m_1 - m_0)/\sigma)$, and let $r = (Y - m_0)/\sigma$ be a standardized residual vector in absence of disease alleles ($m = m_0$). Then, inserting $Y|e_j, v \sim N(m_0 1 + (m_1 - m_0)b_j, \sigma^2\Sigma)$ into (26) we arrive at

$$S_{\text{normdom}}(v) = \sum_{j=1}^{2f} K^{b_j \Sigma^{-1}(r - 0.5\log(K)b_j)'} - C, \tag{27}$$

where $C$ is a centering constant. We use the score function name $S = S_{\text{normdom}}$ introduced in Hössjer (2001) for the special case $h_a^2 = h_d^2 = 0$ of no polygenic effects. Notice that $m_2$ does enter into $S_{\text{normdom}}$. The reason is that for rare disease alleles it is very unlikely that there are 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 $S_{\text{normdom}}$ are $(K, m_0, \sigma^2, h_a^2, h_d^2)$. Of these $K$ is most important, since it measures the strength of the major genetic component. For rare disease alleles one has $E(Y_k) \approx m_0$ and $V(Y_k) \approx \sigma^2$. This motivates why $m_0$ and $\sigma$ are used for standardizing phenotypes. □

# 5 Multiparameter Score Functions

So far, $\{\theta_\varepsilon\}$ has been a one-dimensional trajectory of genetic model parameters. It is possible to generalize this to $d$ degrees of freedom (df), i.e. $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)$. This is a way to increase the nonparametric part of the score function, since we no longer need to make any parametric assumption regarding which one-dimensional trajectory to choose. If all components of $\varepsilon$ are zero there is no genetic effect at the locus of interest.

For simplicity, we restrict ourselves to local penetrance models. The definition of the score function (7) yields a vector $S(v) = (S^1(v), \ldots, S^d(v))$ when $\rho = 1$ and a $d \times d$ matrix $S(v) = (S^{ij}(v))_{i,j=1}^d$ when $\rho = 2$. Here $S^i(v) = \partial \log L(t, \theta_\varepsilon; \text{MD}^{\text{perf}})/\partial\varepsilon_i\big|_{\varepsilon=0}$ and $S^{ij}(v) = \partial^2 \log L(t, \theta_\varepsilon; \text{MD}^{\text{perf}})/\partial\varepsilon_i d\varepsilon_j\big|_{\varepsilon=0}$. If the score function $\bar{S}$ is extended in a similar way, Proposition 1 remains valid. Finally, if $\tilde{S}_k(a, v)$ is generalized to a $1 \times d$ vector when $k = 1$ and to a $d \times d$ matrix when $k = 2$, we get the following analogue of Proposition 2:

**Proposition 4** *For a weak penetrance model with $d$ df the conditional likelihood score function is given by*

$$S(v) = \begin{cases} E(\tilde{S}_1(a,v)) - C_1, & \text{if } \rho = 1, \\ E(\tilde{S}_1(a,v)\tilde{S}_1'(a,v)) + E(\tilde{S}_2(a,v)) - C_2, & \text{if } \rho = 2. \end{cases} \tag{28}$$

*where $C_1$ is a centering vector and $C_2$ a centering matrix, both chosen so that $E_{\theta_0}(S(v)) = 0$.*

**Example 5 (Gaussian mixed model with $d = 2$.)** Consider the Gaussian mixed model of Example 2 with $d = 2$ reference models. These will be chosen as two separate vectors $(m_0^*, m_1^*, m_2^*)$ of mean phenotype values. It will be convenient to introduce the inner product $< x, y >= q^2x_0y_0+2pqx_1y_1+p^2x_2y_2$ for vectors in $\mathbb{R}^3$. Three orthogonal unit vectors are $e_0 = (1,1,1)$, $e_1 = (-2p, q-p, 2q)/\sqrt{2pq}$ and $e_2 = (q^{-1}-1, -1, p^{-1}-1)$. For simlicitly, we write $\psi = (m_0, m_1, m_2)$ for the three mean parameters, implicitly treating the other penetrance parameters as fixed. A genetic model with mean parameters $\psi$ has expected phentype $< \psi, e_0 >$, additive genetic variance $< \psi, e_1 >^2$ and dominant genetic variance $< \psi, e_2 >^2$.

Only the three major gene mean paramters depend on $\varepsilon$, so we write $\psi_\varepsilon = (m_{\varepsilon 0}, m_{\varepsilon 1}, m_{\varepsilon 2})$, where

$$\psi_\varepsilon = me_0 + \sigma\varepsilon_1 e_1 + \sigma\varepsilon_2 e_2. \tag{29}$$

This corresponds to having two reference models $(m_0^*, m_1^*, m_2^*) = me_0 + e_1$ or $me_0 + e_2$, both with genetic variance 1. The first reference model has only additive, and the second one only dominant genetic variance. It is shown in the appendix that Proposition 4 holds with $\rho = 2$ for an outbred pedigree, with off diagonal elements zero for the $2 \times 2$ matrix $S(v)$. That is, $S(v) = \text{diag}(S^{11}(v), S^{22}(v))$, where $S^{11}$ and $S^{22}$ are the two score functions in (23) obtained by letting $c = 0$ and $c = 1$ respectively. Viewing $S(v)$ as a $1 \times 2$ vector with components $S^{11}(v)$ and $S^{22}(v)$, the Fisher information w.r.t. the parameter $\epsilon = (\epsilon_1, \epsilon_2) = (\varepsilon_1^2, \varepsilon_2^2)/2$ for perfect marker data and one pedigree is the $2 \times 2$ matrix $I^{\text{perf}} = 2^{-m}\sum_w S'(w)S(w)$. For non-perfect marker data and one pedigree, the score test is

$$\begin{aligned} W(t) &= d \log L(t, \theta_\varepsilon; \text{MD})/d\epsilon|_{\epsilon=0} I(t)^{-1/2} \\ &= E(W^{\text{perf}}(t)|\text{MD})(I^{\text{perf}})^{1/2} I(t)^{-1/2}, \end{aligned} \tag{30}$$

where

$$I(t) = E_{\theta_0}\left(E(S'(v(t))|\mathrm{MD})E(S(v(t))|\mathrm{MD})\right)$$

is the Fisher information matrix and $W^{\mathrm{perf}}(t) = S(v(t))(I^{\mathrm{perf}})^{-1/2}$ is defined as $W(t)$ with perfect marker data $\mathrm{MD}^{\mathrm{perf}}$ instead of MD.

For $N$ pedigrees, we simply add the familywise Fisher information matrices $I_i(t)$. The total score function has the form $\sum_{i=1}^{N} W_i(t)\gamma_i(t)$, where $W_i(t)$ is the likelihood score (30) for the $i^{\mathrm{th}}$ family and $\gamma_i(t) = I_i(t)^{1/2}(\sum_{j=1}^{N} I_j(t))^{-1/2}$ the corresponding $2\times 2$ weight matrix. By definition of the Fisher information matrix we have

$$\begin{array}{rcl} E_{\theta_0}(W(t)) & = & (0,0) \\ V_{\theta_0}(W(t)) & = & I_2. \end{array} \tag{31}$$

For large samples, $W(t)$ has an approximate bivariate normal distribution.

Because of the restrictions $\epsilon_1 \geq 0$ and $\epsilon_2 \geq 0$, $\theta_0$ is not an interior point of the parameter space. After standardization with $I(t)^{-1/2}$, partial derivatives w.r.t. $\epsilon_1$ and $\epsilon_2$ are along unit vectors $f_1 = (f_{11}, f_{12})$ and $f_2 = (f_{21}, f_{22})$ parallell to $(1,0)I(t)^{-1/2}$ and $(0,1)I(t)^{-1/2}$ respectively. Let $\Omega_1$ be region between $f_1$ and $f_2$, i.e. the image of the first quadrant after transformation $I(t)^{-1/2}$. A two-dimensional argument along the lines of (13) leads to a pointwise test statistic $Z(t) = \sup_{f\in\Omega_1} \max(0, (f, W(t)))^2$ for testing $H_0$ against the alternative $\tau = t$. Here the supremum ranges over all unit vectors $f$ 'between' $f_1$ and $f_2$ and $(\cdot,\cdot)$ is the scalar product in $\mathbb{R}^2$. We can rewrite this as

$$Z(t) = \begin{cases} ||W(t)||^2, & W(t) \in \Omega_1, \\ (W(t), f_2)^2, & W(t) \in \Omega_2, \\ 0, & W(t) \in \Omega_3, \\ (W(t), f_1)^2, & W(t) \in \Omega_4, \end{cases}$$

where $\|\cdot\|$ is the Euclidean norm and $\Omega_2$ the region between $f_2$ and $(-f_{12}, f_{11})$. Similarly, $\Omega_4$ is the region between $f_1$ and $(f_{22}, -f_{21})$ and $\Omega_3$ is the complement of $(\Omega_1 \cup \Omega_2 \cup \Omega_4)$.

Under perfect marker information, $W(t)$ has approximately a standard bivariate normal distribution for large samples. This leads to $Z(t)$ being a mixture of three $\chi^2$ distributions,

$$Z(t) \in (0.5 - \alpha)\chi^2(0) + 0.5\chi^2(1) + \alpha\chi^2(2),$$

where $\alpha$ and $0.5 - \alpha$ are the angles between the boundaries of $\Omega_1$ and $\Omega_3$ respectively. This kind of limit distribution typically arises for likelihood

ratio tests when the null hypothesis parameter is at the boundary of the parameter space, see Self and Liang (1987).

There is a simple interpretation of the two df linkage score $Z(t)$. Let $Z(t;c)$ be the one df linkage score (11) when $c$ is the assumed proportion of dominant genetic variance at the disease locus. Then each $Z(t;c) = (f, W(t))$ for some unit vector $f = f(c)$ in $\Omega_1$. Hence

$$Z(t) = \max\left(0, \sup_{0\le c\le 1} Z(t;c)\right)^2$$

Since we are only concerned with large (i.e. positive) values of $Z(t)$, this means that $Z(t)$ is essentially equivalent to $\sup_{0\le c\le 1} Z(t;c)$. This is true both for a fixed $t$ or when we maximize $Z(t)$ w.r.t. $t$. □

# 6 A Simulation Study

In this section, we investigate the performance of $S_{\text{wpairs}}$ and $S_{\text{normdom}}$ for various choices of score function parameters. For simplicity, we don't include dominance effects in the score functions and hence put $c = h_d^2 = 0$ in the definition of $S_{\text{wpairs}}$ and $h_d^2 = 0$ in the definition of $S_{\text{normdom}}$. We also assume that the phenotype mean $E(Y_k) = m$ and total variance $V(Y_t) = \sigma_t^2 = \sigma_g^2 + \sigma^2$ have been estimated from populations data. We use the residual vector $r = (Y - m)/\sigma_t$ both in (23) and (27). The three essential genetic model characteristics are then $h_a^2$ and

$$\begin{aligned} \text{Disp} &= (m_2 - m_0)/\sigma \\ \text{Dom} &= (2m_1 - m_0 - m_2))/(m_2 - m_0). \end{aligned}$$

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 non-decreasing we have Disp $\ge 0$ and $-1 \le$ Dom $\le 1$, with Dom taking values -1,0,1 for recessive, additive and dominant models respectively.

We only consider outbred pedigrees, and hence the linkage score function is $Z(t) = W(t)$, i.e. the second row of (14) is used. As performance criterion we use the noncentrality parameter NCP $= E(Z(\tau)|Y)$, the expected value w.r.t. marker data and conditional on phenothpes of the linkage score function at the disease locus. This criterion is related to the power to detect linkage (Feingold et. al. (1993)), but does not require specification of a threshold, significance level or chromosomal regions to be tested. For a genomewise scan, a 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, 2003).

For perfect marker data and one pedigree one has

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

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

$$\text{NCP} = \sqrt{N} \cdot \frac{\sum_{i=1}^N \gamma_i \text{NCP}_i / N}{\sqrt{\sum_{i=1}^N \gamma_i^2 / N}} \tag{32}$$

with $\text{NCP}_i$ the noncentrality parameter and $\gamma_i$ the weight of the $i^{\text{th}}$ pedigree. If we assume pedigrees (including their phenotypes) are drawn from a population, the second factor of (32) converges to the asymptotic noncentrality parameter $\text{ANCP} = \int \gamma(Y) \text{NCP}(Y) dP(Y) / \sqrt{\int \gamma^2(Y) dP(Y)}$ as $N$ grows, where $dP(Y)$ is the sampling distribution of pedigrees (including their phenotype vectors $Y$), see Hössjer (2001, 2003a). Hence

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

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 $\alpha$ ($0 < \alpha \leq 1$) of randomly sampled $Y$ ($P_\theta(Y) = \sum_G P_\psi(Y|G) P_p(G)$) with largest weights[2] $\gamma(Y)$.

In Figures 1-6 we have plotted $50 \times$ANCP for perfect marker data and various score functions, genetic models (Disp, Dom, $h_a^2$), pedigrees and sampling fractions $\alpha$. This corresponds to a noncentrality parameter of a sample with $N = 2500$. For comparison, we have also included the optimal (in terms of NCP) score function $S_{\text{optimal}}$, which is the centered version of $P(v|Y)$ (Hössjer, 2003a). We assume, for simplicity of interpretation, that all families in the populations have the same pedigree structure. We have included four pedigree structures in the simulations - sib pair (SP), sib trio (Strio), sib

[2]For a correctly specified model, this means that a fraction $\alpha$ of the most informative pedigrees are considered. This is because the weights are proportional to the square root of the Fisher informations.

quartet (Squart) and first cousin (Cous) families. In all cases, all individuals except the two parents of the first generation are genotyped for markers.

Notice that ANCP for $S_{\text{optimal}}$ is a measure of informativity for the particular combination of pedigree structure, genetic model parameters and sampling fraction. It is evident from Figure 1-6 that the informativity in general increases with increased polygenic heritability $h_a^2$. We may explain this by the fact 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 $S_{\text{normdom}}$. It turns out that a value of $h_a^2$ around 0.5 gives overall best performance when the true $h_a^2$ varies between 0 and 1. The same conclusions can be drawn for other pedigree structures, genetic models and sampling fractions, both for $S_{\text{normdom}}$ and $S_{\text{wpairs}}$. For this reason, we have compared $S_{\text{normdom}}$ and $S_{\text{wpairs}}$ with assumed $h_a^2 = 0.5$ with $S_{\text{HE}}$ and $S_{\text{optimal}}$. Here

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

is the score function analogue of the classical Haseman-Elston regression method for quantitative traits, see Haseman and Elston (1972) and Hössjer (2001). Of the three non-ideal score functions, $S_{\text{normdom}}$ and $S_{\text{wpairs}}$ have the best performance, with $S_{\text{normdom}}$ slightly better. The Haseman-Elston score function is competitive for large $h_a^2$ and large disease allele frequencies $p$.

# 7 Conclusions

In this paper, we have presented a general semiparametric framework based on conditional likelihoods for choosing score functions in linkage analsysis. We believe that a strategic parametric choice of the fixed parameter often leads to robust procedures with good performance. When the chosen model is not too misspecified, the decreased number of degrees of freedom compared to a fully nonparametric approach can be worthwile. However, more work is needed to compare the two approachess.

A disadvantage of the familywise score (9) and the total score (11) is that computation of $I(t)$ requires extensive simulation. A Monte Carlo estimate of $I(t)$ based on multiple imputation can be defined (Clayton, 2001). It is unbiased and increasingly accurate as $N$ grows. Another alternative is to

normalize by $\sqrt{I^{\text{perf}}}$ rather than $\sqrt{I(t)}$, since $I^{\text{perf}}$ is easy to compute and does not depend on $t$. Now $I(t) \leq I^{\text{perf}}$ follows from Jensen's inequality (this is in fact a general inequality relating Fisher informations for incomplete and complete data, cf. e.g. Dempster et. al., 1977). As a result, $V_{\theta_0}(W(t)) \leq 1$, both for the familywise and total score $W(t)$. Thresholds based on the perfect data approximation $V_{\theta_0}(W(t)) = 1$ then lead to conservative tests, see Kruglyak et. al. (1996) and Kong and Cox (1997).

The likelihood score functions based on a Gaussian mixed model are very sensitive to outliers. When these are present, it is advisable to replace the vector $r$ of standardized residuals in $S_{\text{wpairs}}$ and $S_{\text{normdom}}$ by a more robust transformation of $Y$. See Wang and Huang (2002) for one proposal along these lines.

The lod score of Morton (1955) is defined as

$$\log_{10} \frac{P_{t,\theta}(\text{MD}, Y)}{P_{\infty,\theta}(\text{MD}, Y)} = \log_{10} \frac{P_{t,\theta}(\text{MD}, Y)}{P(\text{MD}) P_\theta(Y)} = \text{constant} + \log_{10} L(t, \theta; \text{MD}),$$

where $t = \infty$ corresponds to $H_0$. Maximization with respect to $\theta$ at each locus $t$ leads to the mod score of Risch (1984) and Clerget-Darpoux et. al. (1986). Therefore, mod scores are equivalent to using the profile conditional likelihood (13). This approach is computationally demanding for larger pedigrees, although a faster version can be defined by replacing the original conditional likelihood $L(t, \theta)$ by an empirical one based on the score function $S$ (Kong and Cox, 1997).

# Appendix. Proofs

**Proof of Proposition 1.** By taking the logarithm of (15) and differentiating $k$ times, it follows that

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

where $C = d^k \log P_{\theta_\varepsilon}(Y)/d\varepsilon^k \Big|_{\varepsilon=0}$. The proposition is proved if we can show that $E_{\theta_0}(S) = 2^{-m} \sum_w S(w) = 0$. But this follows by diffentiating the relationship $\log\left(\sum_w P_{\theta_\varepsilon}(v = w|Y)\right) = 0$ $k$ times w.r.t. $\varepsilon$ at $\varepsilon = 0$. □

**Proof of Proposition 2.** Let $\tilde{l}(\psi) = \tilde{l}(\psi; a, v) = \log P_\psi(Y|a, v)$ be the log likelihood for the complete data. Then

$$P_\psi(Y|v) = E \exp\left(\tilde{l}(\psi)\right), \tag{A.1}$$

where expectation is w.r.t. $a$. Let $\tilde{l}^{(k)}(\psi_\varepsilon)$ be the $k^{\text{th}}$ derivative of $\tilde{l}(\psi_\varepsilon)$ w.r.t. $\varepsilon$. Notice that $\tilde{l}^{(k)}(\psi_0) = \tilde{S}_k(a,v)$. Differentiating the logarithm of (A.1) twice we obtain

$$\begin{aligned} d\log P_{\psi_\varepsilon}(Y|v)/d\varepsilon|_{\varepsilon=0} &= E\tilde{S}_1(a,v), \\ d^2\log P_{\psi_\varepsilon}(Y|v)/d\varepsilon^2|_{\varepsilon=0} &= E\tilde{S}_1^2(a,v) - E^2\tilde{S}_1(a,v) + E\tilde{S}_2(a,v). \end{aligned}$$

Now (19) follows from the definition of $\tilde{S}(v)$, $S(v)$ and Proposition 1. □

**Expansion of $u$ in Examples 2 and 3.** Regard the founder allele vector $a$ as random and the inheritance vector $v$ at the major disease locus as fixed. Let $\xi_j = (a_j - p)/\sqrt{pq}$ for $j = 1, \ldots, 2f$. Then $\{\xi_j\}$ are i.i.d. random variables with mean zero and unit variance. For an individual $k$ with both of it's alleles originating from different founders, it is proved in Hössjer (2001a) that

$$u_k = (\sigma_a(\xi_{j_{2k-1}} + \xi_{j_{2k}})/\sqrt{2} + \sigma_d\xi_{j_{2k-1}}\xi_{j_{2k}})/\sigma_g. \tag{A.2}$$

For an outbred pedigree, $j_{2k-1}$ and $j_{2k}$ must be different with probability one for all $k$. Hence (22) follows from (18) and (A.2) by averaging over $a$. If $k$ receives both of it's alleles from the same founder ($\text{HBD}_k = 1$), the analogous expansion is

$$u_k = (\sigma_d + \sigma_l\xi_{j_{2k}})/\sigma_g,$$

where $\sigma_l = \sqrt{pq}(m_2^* - m_0^*)$. Averaging w.r.t. $a$ we arrive at (24). □

**Proof of Proposition 3.** Since $\psi$ is fixed, we omit it as index in the notation. Thus $P_{\theta_\varepsilon}(Y|v) = \sum_a P(Y|v)P_\varepsilon(a)$, with $P_\varepsilon(a)$ as in (17). Notice that $P_\varepsilon(0) = 1 - 2f\varepsilon + o(\varepsilon)$ and $P_\varepsilon(e_j) = \varepsilon + o(\varepsilon)$ as $\varepsilon \to 0$. Taking the logarithm of $P_{\theta_\varepsilon}(Y|v)$ and differentiating with respect to $\varepsilon$ at $\varepsilon = 0$ we obtain

$$\begin{aligned} \bar{S}(v) &= \textstyle\sum_a P_0'(a)P(Y|a,v)/P(Y|0,v) \\ &= \textstyle\sum_{j=1}^{2f} P(Y|e_j,v)/P(Y|0,v) - 2f. \end{aligned}$$

The result now follows from Proposition 1. □

**Deriving the score function of Example 5.** Assume w.l.o.g. that $m = 0$. Then $e_1 = (e_{10}, e_{11}, e_{12})$ and $e_2 = (e_{20}, e_{21}, e_{22})$ are the mean parameters of purely additive and dominant reference models with genetic variance 1 at the main locus. Define vectors $u_1 = (u_{11}, \ldots, u_{1n})$ and $u_2 = (u_{21}, \ldots, u_{2n})$ by $u_{ik} = e_{i,a_{j_{2k-1}(v)}+a_{j_{2k}(v)}}$. Then, an expansion of the log likelihood as in Example 2 gives

$$\begin{aligned} \tilde{S}_1(a,v) &= (u_1\Sigma^{-1}r', u_2\Sigma^{-1}r'), \\ \tilde{S}_2(a,v) &= -\begin{pmatrix} u_1\Sigma^{-1}u_1' & u_1\Sigma^{-1}u_2' \\ u_2\Sigma^{-1}u_1' & u_2\Sigma^{-1}u_2' \end{pmatrix}. \end{aligned} \tag{A.3}$$

Now use Proposition 4 and average w.r.t. $a$ to get the likelihood score function $S(v) = (S^{ij}(v))_{i,j=1}^2$. Comparing (A.3) with (21), it is clear that $S^{11}(v)$ and $S^{22}(v)$ must equal (23) with $c = 0$ and 1 respectively. This is because $e_1$ and $e_2$ correspond to purely additive and dominant reference models. The off-diagonal elements $S^{12}(v) = S^{21}(v)$ are zero since $E(u_{1k}u_{2l}) = 0$ for any $1 \leq k, l \leq n$. This in turn is a consequence of the expansions

$$\begin{aligned} u_{1k} &= (\xi_{j_{2k-1}} + \xi_{j_{2k}})/\sqrt{2} \\ u_{2l} &= \xi_{j_{2l-1}}\xi_{j_{2l}} \end{aligned}$$

and the zero mean and independence of $\{\xi_j\}$. □

# Acknowledgement

The author wishes to thank Yudi Pawitan for pointing out valuable references.

# References

Bottai, M. (2003). Confidence regions when the Fisher information is zero. *Biometrics* **90**, 73-84.

Clayton, D. (2001). Tests for genetic linkage and association with incomplete data. Invited talk given at the Easter North American Region of the 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 Epidmiol.* **11**, 189-200.

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

Dupuis, J. and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics*, **151**, 373-386.

Elston, R.C. (1989). Man bites dog? The validity of maximizing lod scores to determine mode of inheritance. *Am. J. Med. Genet.* **34**, 487-488.

Ewens, W.J. and Shute, N.C.E. (1986). A resolution of the ascertainment problem. I. Theory. *Theory Popul. Biol.* **30**, 388-412.

Feingold, E., O'Brown, P. and Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. of Hum. Genet.* **53**, 234-251.

Feingold, E. (2002). Regression-based quantitative-trait-locus mapping in the 21st century. *Am. J. of Hum. Genet.* **71**, 217-222.

Haseman, J.K. and Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**, 3-19.

Hössjer, O. (2001). Determining inheritance distributions via stochastic penetrances. Report 2001:17, Centre for Mathematical Sciences, Mathematical Statistics, Lund University. Under revision for *J. Amer. Statist. Assoc.*.

Hössjer, O. (2003a). Asymptotic estimation theory of multipoin linkage analysis under perfect marker information. *Ann. Statist* **31**, 1075-1109.

Hössjer, 0. (2003b). Assessing accuracy in linkage analysis by means of confidence regions. *Genetic Epidemiology*, **25**, 59-72.

Kong, A. and Cox, N.J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *Am. J. of Hum. Genet.* **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. *Am. J. Hum. Genet.*, **58**, 1347-1363.

Kruglyak, L. and Lander, E. (1995). High resolution genetic mapping of complex traits. *Am. J. of Hum. Genet.* **56**, 1212-1223.

Lander, E. and Green, P. (1987). Construction of multilocus genetic maps in humans. *Proc. Nat. Acad. Sci. USA* **84**, 2363-2367.

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

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates Inc.

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

Morton, N.E. (1955). Sequential tests for the detection of linkage. *Am. J. of Hum. Genet.* **61**, 277-318.

Ott, J. (1999). *Analysis of Human Genetic Linkage*, third ed., John Hopkins Univ. Press.

Putter, H., Sandkuijl, L.A. and van Houwelingen, J.C. (2002). Score test for detecting linkage to quantitative traits. *Genet. Epidemiol.* **22**, 345 355.

Risch, N. (1984). Segregation analysis incorporating genetic markers. I. Single-locus models with an application to type I diabetes. *Am. J. Hum.*

*Genet.*, **36**, 363-386.

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

Rotnitzky, A., Cox, D.R., Bottai, M. and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli* **6**, 243-284.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82**, 605-610.

Sham, P. (1998). *Statistics in Human Genetics.* Arnold Applications in Statistics, London.

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

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

Wang, K. and Huang, J. (2002). A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. *Am. J. Hum. Genet.* **70**, 412-424.

Winter, R.M. (1980). The estimation of phenotype distributions from pedigree data. *Am. J. Med. Genet.* **7**, 537-542.

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

Ängquist, L. and Hössjer, O. (2003). Improving the calculation of statistical significance in genome-wide scans. Report 2003:3, Mathematical Statistics, Stockholm University.

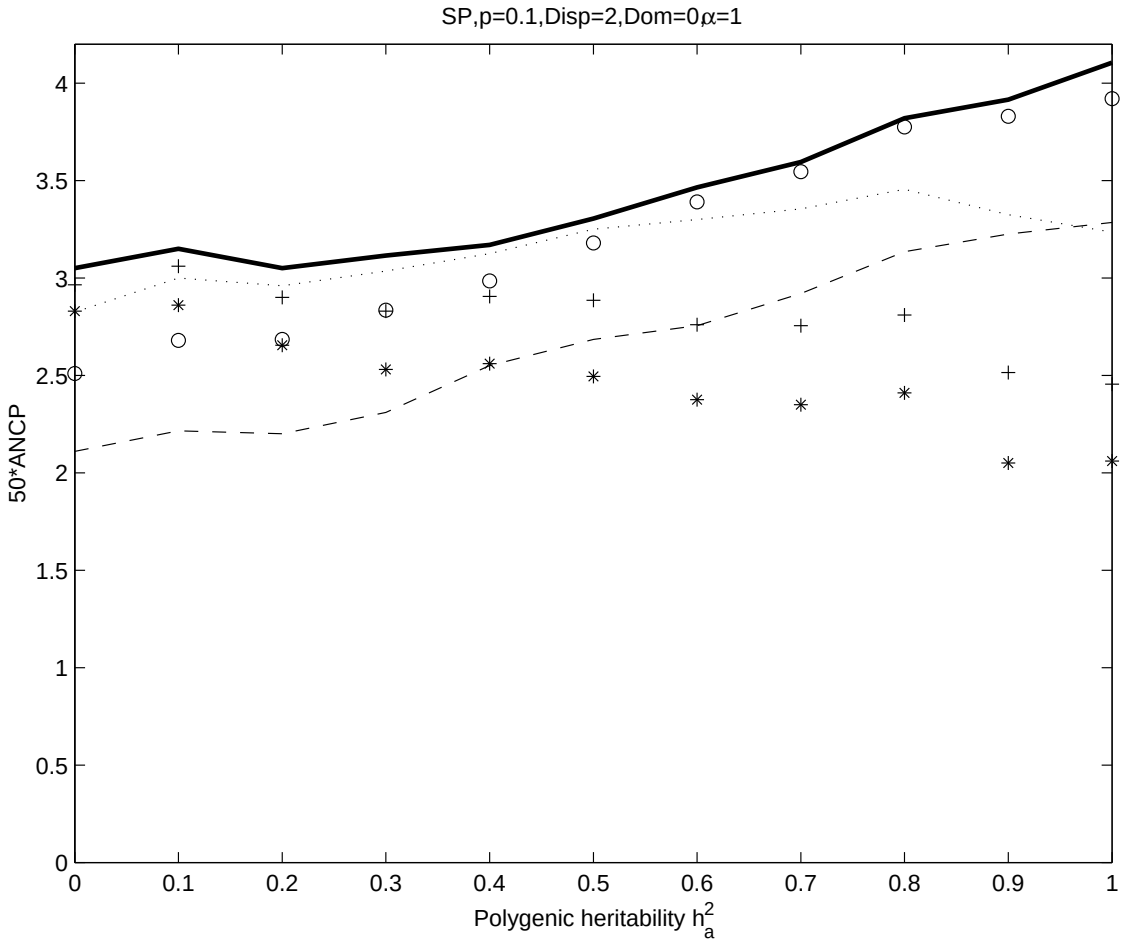


Figure 1: $50 \cdot \mathrm{ANCP}$ as function of true $h_a^2$ for optimal (thick solid line), Haseman-Elston (dashed line) and $S_{\mathrm{normdom}}$ score functions with different choices of assumed $h_a^2$: $h_a^2 = 0$ $(*)$, $h_a^2 = 0.2$ $(+)$, $h_a^2 = 0.5$ (dotted line) and $h_a^2 = 0.8$ (o).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

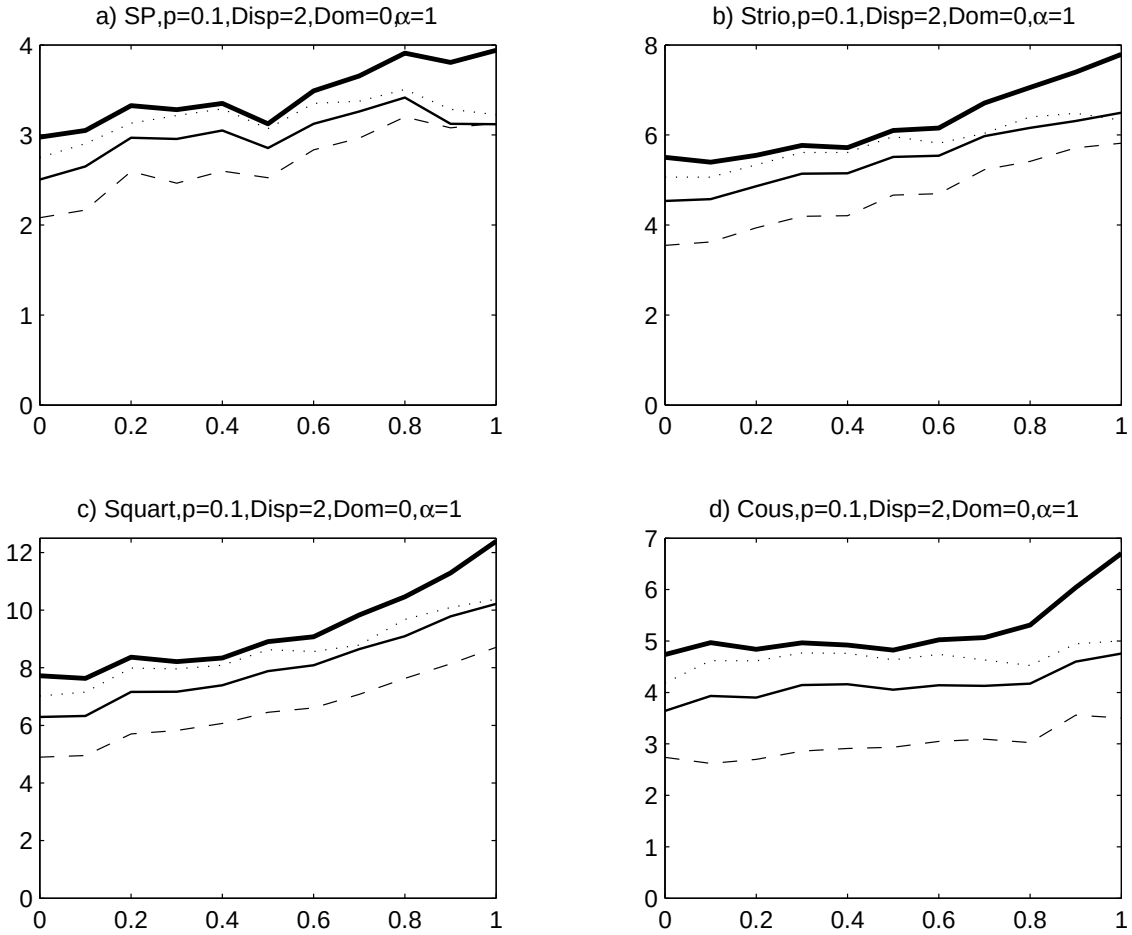


Figure 2: $50 \cdot$ ANCP as function of true $h_a^2$ for different score functions: Optimal (thick solid line), $S_{\text{normdom}}$ with assumed $h_a^2 = 0.5$ (dotted line), $S_{\text{wpairs}}$ with assumed $h_a^2 = 0.5$ (thin solid line) and Haseman-Elston (dashed line). The four subplots correspond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

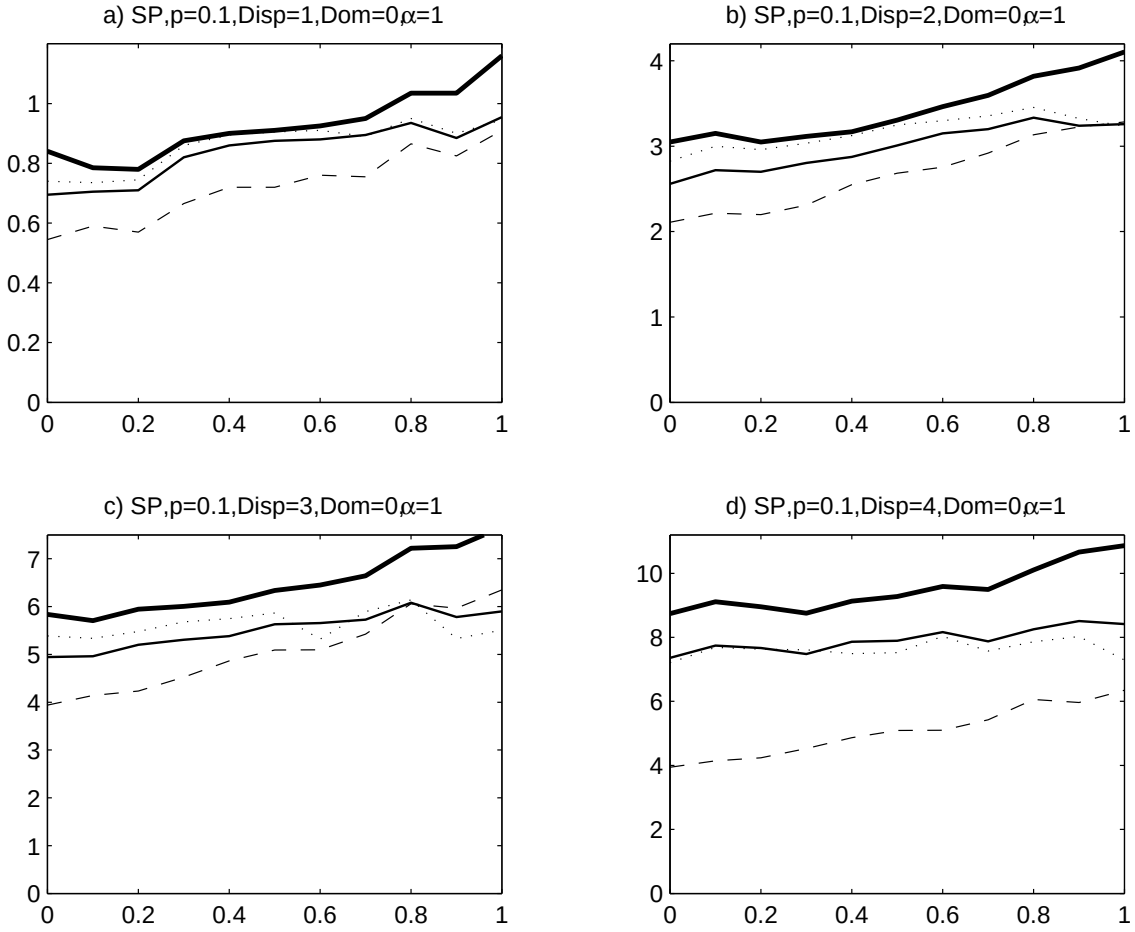


Figure 3: $50 \cdot \text{ANCP}$ as function of true $h_a^2$ for different score functions: Optimal (thick solid line), $S_{\text{normdom}}$ with assumed $h_a^2 = 0.5$ (dotted line), $S_{\text{wpairs}}$ with assumed $h_a^2 = 0.5$ (thin solid line) and Haseman-Elston (dashed line). The four subplots correspond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

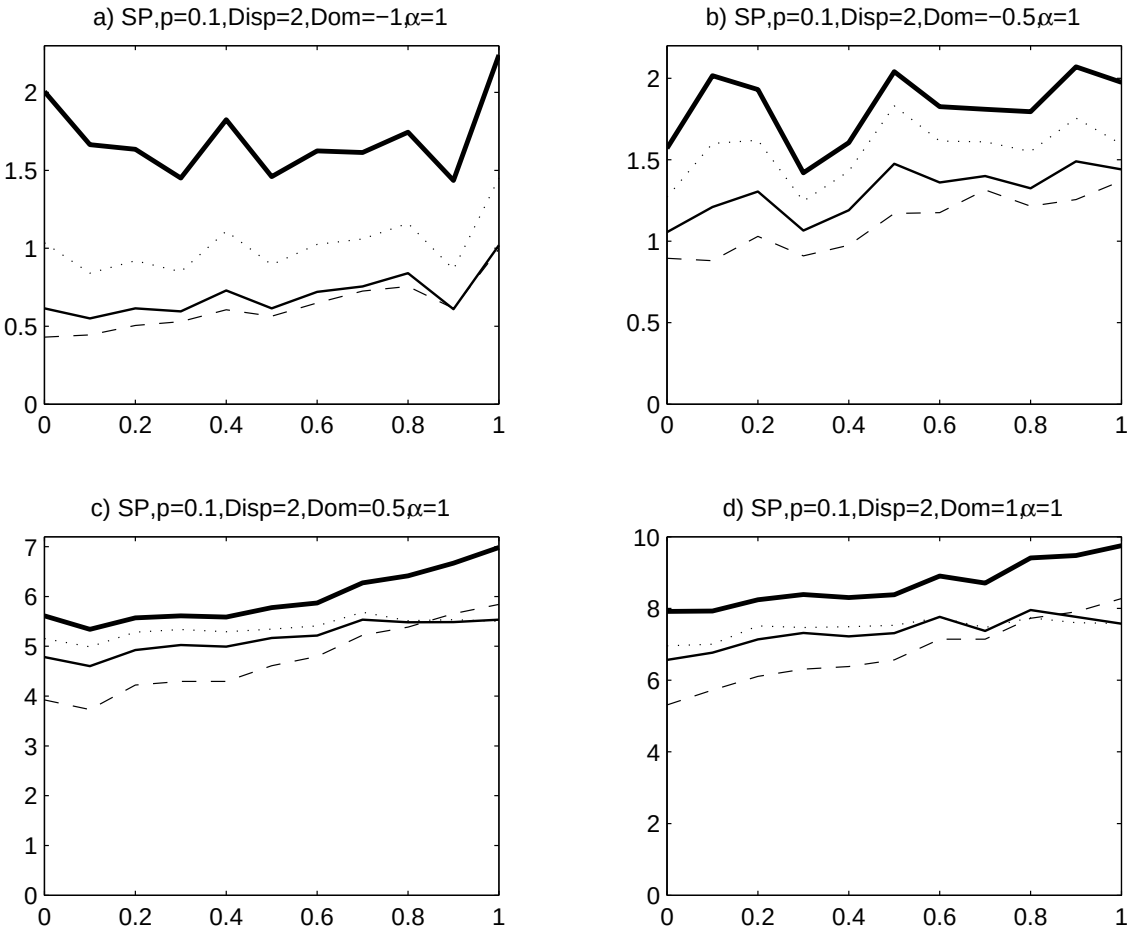


Figure 4: $50 \cdot$ ANCP as function of true $h_a^2$ for different score functions: Optimal (thick solid line), $S_{\text{normdom}}$ with assumed $h_a^2 = 0.5$ (dotted line), $S_{\text{wpairs}}$ with assumed $h_a^2 = 0.5$ (thin solid line) and Haseman-Elston (dashed line). The four subplots correspond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

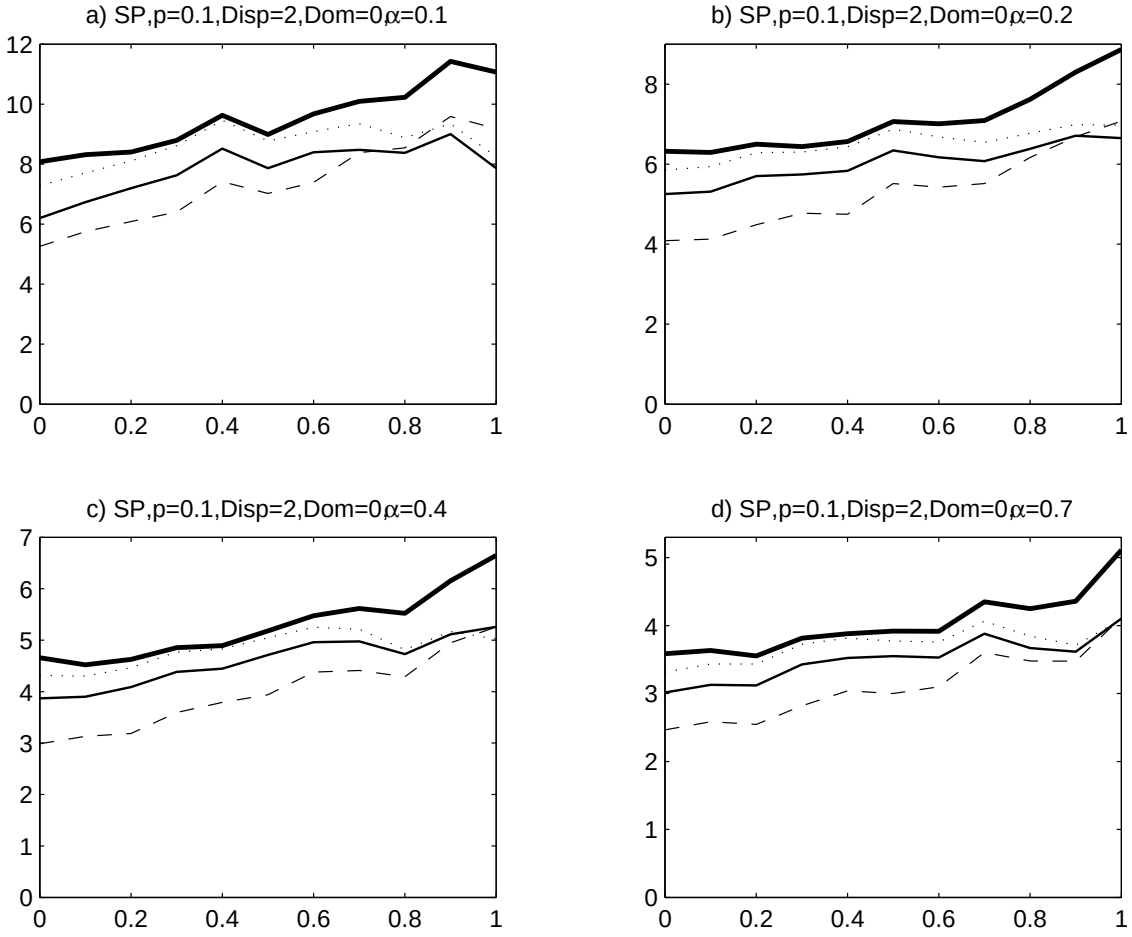


Figure 5: $50 \cdot$ ANCP as function of true $h_a^2$ for different score functions: Optimal (thick solid line), $S_{\text{normdom}}$ with assumed $h_a^2 = 0.5$ (dotted line), $S_{\text{wpairs}}$ with assumed $h_a^2 = 0.5$ (thin solid line) and Haseman-Elston (dashed line). The four subplots correspond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

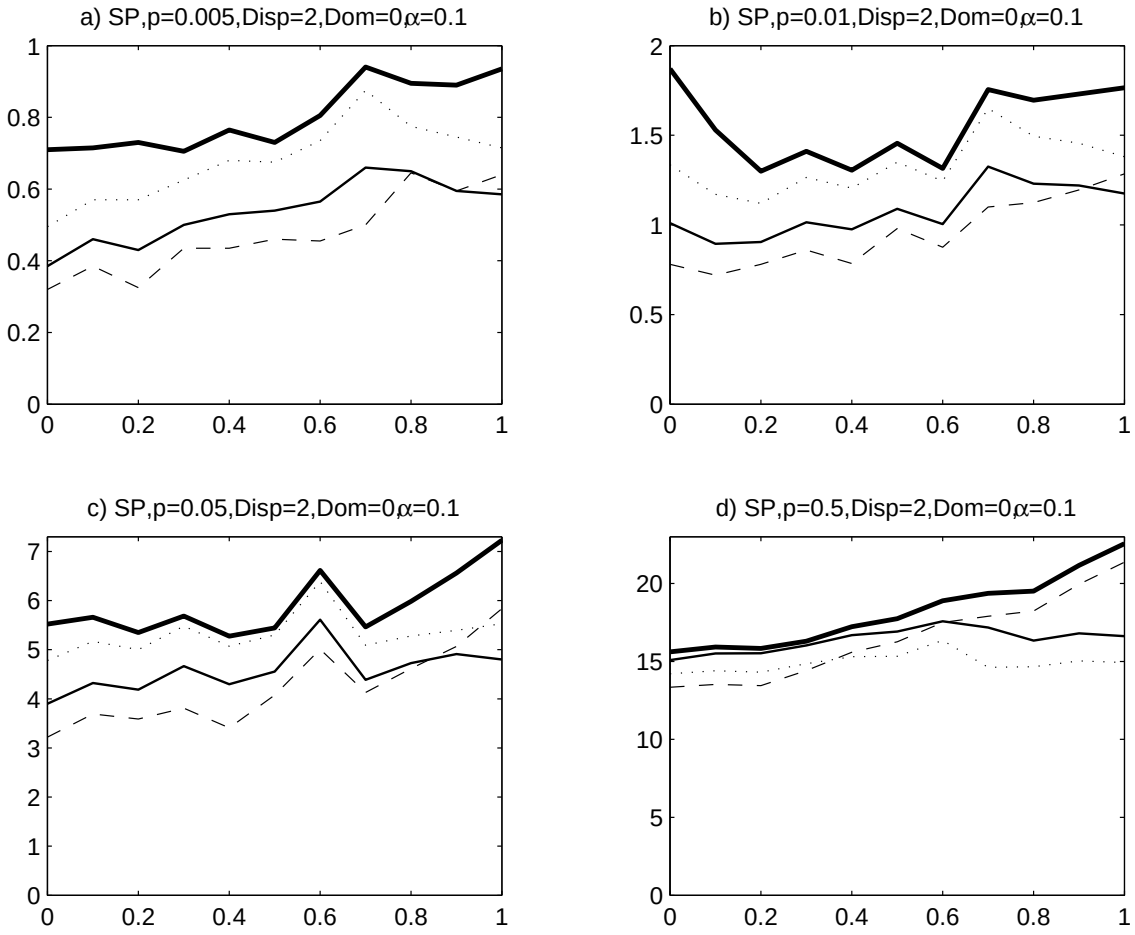


Figure 6: $50 \cdot$ ANCP as function of true $h_a^2$ for different score functions: Optimal (thick solid line), $S_{\text{normdom}}$ with assumed $h_a^2 = 0.5$ (dotted line), $S_{\text{wpairs}}$ with assumed $h_a^2 = 0.5$ (thin solid line) and Haseman-Elston (dashed line). The four subplots correspond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