J. Math. Biol. 50, 208–232 (2005)
Digital Object Identifier (DOI):
10.1007/s00285-004-0289-z

Mathematical Biology

Ola Hössjer

# Information and effective number of meioses in linkage analysis

Received: 16 December 2003 / Revised version: 7 May 2004 /
Published online: 11 November 2004 – © Springer-Verlag 2004

**Abstract.** In this paper we introduce two information criteria in linkage analysis. The setup is a sample of families with unusually high occurrence of a certain inheritable disease. Given phenotypes from all families, the two criteria measure the amount of information inherent in the sample for 1) testing existence of a disease locus harbouring a disease gene somewhere along a chromosome or 2) estimating the position of the disease locus.

Both criteria have natural interpretations in terms of effective number of meioses present in the sample. Thereby they generalize classical performance measures directly counting number of informative meioses.

Our approach is conditional on observed phenotypes and we assume perfect marker data. We analyze two extreme cases of complete and weak penetrance models in particular detail. Some consequences of our work for sampling of pedigrees are discussed. For instance, a large sibship family with extreme phenotypes is very informative for linkage for weak penetrance models, more informative than a number of small families of the same total size.

## 1. Introduction

The purpose of statistical linkage analysis is to map the location of gene(s) causing or increasing susceptibility to a certain disease or trait based on phenotype and DNA data from a number of pedigrees with occurrence of the disease. The technique is almost hundred years old and based on Morgan's discovery of crossovers during formation of germ cells. An overview of linkage analysis can be found in the books of Sham (1998) and Ott (1999).

DNA data is collected by typing as many family members as possible for genetic markers, which are DNA segments of known position along the genome, preferably highly polymorphic. Historically, the available genetic markers were few, and linkage analysis was carried out as a sequence of two-point comparisons between the disease locus and one marker locus at a time. The recombination parameter between the marker and disease locus is the unknown parameter of interest for two-point linkage analysis, with possible extra nuisance parameters (disease allele

O. Hössjer: Department of Mathematics, Stockholm University, Sweden

Research sponsored by the Swedish Research Council, under contract nr. 626-2002-6286.

Address for correspondence: Department of Mathematics, Stockholm University, 106 91 Stockholm, Sweden. e-mail: ola@math.su.se

*Key words or Phrases:* Effective number of meioses – Information criteria – Linkage analysis – Score functions

frequencies, penetrance parameters) from the genetic model. In parametric linkage analysis, all nuisance parameters are known, whereas the opposite is true in nonparametric linkage. Multipoint linkage analysis is concerned with simultaneous analysis of the trait locus together with a number (>1) of marker loci. In this case the order and distances between the markers become important, and there is no longer a single recombination parameter from any of the marker loci that could serve as linkage parameter. Instead, the map distance along the chromosome is used for this purpose.

It is important, for sampling and planning of linkage studies, that accurate measures of informativity are found for a single pedigree. In general, such a measure will depend on the pedigree structure, the genetic model (known or not) and possibly also on marker informativity and information on which pedigree members that are genotyped. It could be either conditional on observed phenotypes or unconditional. In the latter case, a sampling distribution for pedigrees must be known, and this introduces additional complication into model building (see e.g. Dudoit and Speed, 2000).

The statistical theory of two-point linkage analysis is well developed. The direct method is the classical procedure based on directly counting recombinant and nonrecombinant meioses. If $k$ out of $m$ meioses are recombinant, $k/m$ is an estimate of the recombination fraction and $m$ may serve as the information content of data. However, in general recombinations are not directly observed. In Chapters 4 and 5 of Ott (1999) two likelihood based information measures are outlined for parametric linkage analysis. The Fisher information of the recombination parameter $\theta$ is an *estimation based criterion* quantifying the ability to estimate $\theta$. The (maximum) expected linkage score (M)ELOD on the other hand is a *test related criterion* based on the ability to test the null hypothesis of unlinked trait and marker loci ($\theta = 0.5$). Both these criteria are proportional to $m$ in situations when the direct method applies.

The statistical theory of multipoint linkage analysis is still under development. The purpose of this paper is to define information bounds in multipoint linkage analysis under the following two assumptions: i) Perfect marker data is available, i.e. all (or sufficiently many for complete knowledge of DNA inheritance) pedigree members are genotyped at a dense set of genetic markers. ii) The information measures are conditional on observed phenotypes.

Under these premises, we define one test-related ($I^{\text{test}}$) and one estimation-related ($I^{\text{est}}$) information criterion. It turns out that $m^{\text{test}} = \log_2(I^{\text{test}} + 1)$ and $m^{\text{est}} = I^{\text{est}}$ can be interpreted as the effective number of meioses present in the pedigree for testing and estimation respectively. The effective number of meioses are added over families and provide a natural unified information framework for different kinds of genetic models.

We consider two extreme cases of penetrance models in particular detail: Complete and weak penetrance models. For complete penetrance models, the amount of testing or estimation information is roughly proportional to pedigree size. For weak penetrance models (the case of most interest in human genetics for complex diseases), the testing/estimation information is roughly proportional to squared pedigree size. However, close relationships (mostly sibs) and extreme phenotypes are much more informative than distant relationships and non-extreme phenotypes.

The derivation of $I^{\text{test}}$ is based on the common assumption in nonparametric linkage to standardize family scores to have zero mean and unit variance at loci unlinked to the trait locus. Such scores have distinct advantages over lod scores in having a clear connection to $p$-values (Kurbasic and Hössjer, 2003). We use the noncentrality parameter (Feingold et al. (1993)), a criterion closely related to power but mathematically more convenient. Recent results of Hössjer (2003c, 2003e), are employed, where the noncentrality parameter for weak genetic models is derived. The derivation of $I^{\text{est}}$ is based on results in Hössjer (2003a,b), where estimation accuracy under perfect marker information is analyzed. Both information bounds are defined for arbitrary pedigree structures and genetic models.

The paper is organized as follows: In Sections 2 and 3 we introduce the information bounds $I^{\text{test}}$ and $I^{\text{est}}$ for one and $N$ pedigrees. We also derive their optimality properties as information bounds. Complete and weak penetrance models are considered in Sections 4 and 5. Some extension are discussed in the last section and finally, proofs are collected in the appendix.

## 2. Information Bounds for One Pedigree

Consider a pedigree with $n$ individuals $1, \ldots, n$ numbered so that the first $f$ have no ancestors (the founders) and the remaining persons have both their parents in the pedigree (the nonfounders). For each nonfounder there are two meioses giving rise to the maternal and paternal germ cells. Thus the total number of meioses in the pedigree is $m = 2(n - f)$. Suppose we number them $1, \ldots, m$. The inheritance vector $v(t) = (v_1(t), \ldots, v_m(t))$ is a binary vector of length $m$ whose $j^{\text{th}}$ bit specifies the outcome (0 = grandpaternal, 1 = grandmaternal transmission) of the $j^{\text{th}}$ meiosis at locus $t$. It was originally introduced by Donnelly (1983) and subsequently used in linkage analysis as a general tool for handling extended pedigrees (Kruglyak et al., 1996).

For a chromosome of length $L$ Morgans, allele transmission for the pedigree is completely determined by the stochastic process $V = \{v(t);\ 0 \leq t \leq L\}$. The purpose of genetic markers is to retain as much information as possible about $V$. Under perfect marker information we assume $V$ to be known[1]. According to Mendel's law of segregation, $P(v_j(t) = 0) = P(v_j(t) = 1) = 0.5$ for each meiosis. If the outcomes of all meioses are independent, the a priori distribution of $v(t)$ becomes

$$P_0(w) := P(v(t) = w) = 2^{-m} \tag{1}$$

for all binary vectors $w$ of length $m$ and all loci $t$. Let $Y = (Y_1, \ldots, Y_n)$ be the collection of phenotypes of the pedigree and $\tau$ the unknown position of a gene which

[1] To be precise, the phase of the founders is usually not known. In this case, $V$ is not known even for perfect marker data (Kruglyak et al., 1996). However, the score functions used in linkage analysis are usually invariant w.r.t. this uncertainty. Therefore, it is no loss of generality to assume that $V$ is known.

causes or increases susceptibility to the disease. Knowledge of $Y$ gives information about $v = v(\tau)$, so that the aposteriori distribution

$$P_1(w) := P(v = w|Y) = 2^{-m} P(Y|v = w)/P(Y) \tag{2}$$

differs from the prior $P_0$. We refer to $P_1$ as the conditional inheritance distribution. It is of crucial importance in linkage analysis, see Dudoit and Speed (2000) and Hössjer (2003e) for a detailed discussion. The factor $P(Y|v)$ in (2) depends on genetic model parameters. Let $G_k = (a_{2k-1}a_{2k})$ be the genotype of the $k^{\text{th}}$ individual at the disease locus, with $a_{2k-1}$ and $a_{2k}$ the alleles received from the father and mother respectively. With $G = (G_1, \dots, G_n)$ the collection of genotypes and $a = (a_1, \dots, a_{2f})$ the founder alleles, we have

$$P(Y|v) = \sum_G P(Y|G)P(G|v) = \sum_a P(Y|a, v)P(a). \tag{3}$$

We used the fact that $G$ is uniquely determined by $a$ and $v$, since $v$ specifies how founder alleles are transmitted to all nonfounders. Further, we assumed independence of $a$ and $v$ (no segregation distortion).

The genetic model consists of disease allele frequencies and penetrance parameters. The former enter into $P(a)$ and the latter into $P(Y|a, v) = P(Y|G)$. Suppose there are $M$ possible alleles $0, \dots, M - 1$ at the disease locus with $p_i$ the frequency (probability) of the $i^{\text{th}}$ allele. Under random mating all founder alleles are independent and

$$P(a) = \prod_{k=1}^{2f} p_{a_k}. \tag{4}$$

Since $a$ is an *ordered* vector of founder alleles, no factors 2 are needed for heterozygous founders.

*Example 1 (Complete penetrance.).* When genotypes can be determined unambiguously from phenotypes we have complete penetrance and put $Y = G$. The penetrance factor $P(Y|G)$ is then one if $Y = G$ and zero if $Y = G'$ for any other genotype configuration $G'$. □

*Example 2 (Binary phenotypes.).* Binary phenotypes are usually encoded as '1 = affected' and '0 = unaffected'. For a single individual, the affection probability $P(Y_k = 1|G_k)$ is a function of $(a_{2k-1}a_{2k})$, with the order of $a_{2k-1}$ and $a_{2k}$ unimportant. Hence there are $M(M + 1)/2$ affection probabilities. When $M = 2$ we usually interpret the two alleles as 'normal' (allele 0) and 'disease causing' (allele 1). Then the penetrance parameters are

$$\psi = (\psi_0, \psi_1, \psi_2)$$

where $\psi_j$ is the affection probability for an individual with $j$ disease alleles. If $|G_k| = a_{2k-1} + a_{2k}$ is the number of disease alleles of $G_k$, the penetrance factor for individual $k$ becomes

$$P(Y_k|G_k) = \psi_{|G_k|}^{Y_k}(1 - \psi_{|G_k|})^{1-Y_k}.$$

The penetrance factor for the whole pedigree is

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

if phenotypes are conditionally independent given genotypes. This is the case if there are no other genes contributing to the disease and no shared environmental components among the pedigree members. □

*Example 3 (Gaussian phenotypes.).* When each individual $k$ has a continuous phenotype $P(Y|G)$ is a density in (3). It is common to model the phenotypes as normally distributed. Assume $M = 2$ alleles with 0 the normal and 1 the disease allele. Put $Y_k|G_k \in N(m_{|G_k|}, \sigma^2)$. Here $m_0$, $m_1$ and $m_2$ are the mean phenotype values for an individual with 0, 1 and 2 disease alleles and the residual variance $\sigma^2$ is caused by polygenic and/or environmental effects. In vector form we have

$$Y|G \in N(\mu, \sigma^2 C), \tag{5}$$

where $\mu = (m_{|G_1|}, \dots, m_{|G_n|})$ and $C$ is an $n \times n$ correlation matrix.

If there are no shared environmental effects then

$$C = (1 - h_a^2 - h_d^2)I + h_a^2 R + h_d^2 \Delta, \tag{6}$$

where $I$ is an $n \times n$ identity matrix and $h_a^2$ and $h_d^2$ are polygenic heritabilities (i.e. the fraction of $\sigma^2$ due to additive and dominance effects). The $n \times n$ matrices $R = (r_{kl})$ and $\Delta = (\delta_{kl})$ are defined as follows: Let $\text{IBD}_{kl}(w)$ be the number of alleles that $k$ and $l$ share identical by descent, i.e. from the same founder alleles, if $w$ is the inheritance vector. Then $r_{kl} = E_0(\text{IBD}_{kl}(w)/2)$ and $\delta_{kl} = P_0(\text{IBD}_{kl}(w) = 2)$ is the fraction of alleles that $k$ and $l$ share on average (coefficient of relationship) and the probability that they share both alleles IBD respectively when $w \sim P_0$.

The penetrance parameters of the Gaussian model (5)–(6) are

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

See Fisher (1918), Kempthorne (1955) and Lynch and Walsh (1998) for more details. □

For one pedigree, the information bounds $I^{\text{test}}$ and $I^{\text{est}}$ are defined as functions of the prior and posterior distributions $P_0$ and $P_1$. The first one is a distance measure between $P_1$ and $P_0$ defined as

$$\begin{aligned} I^{\text{test}} &= \sum_w \frac{(P_1(w) - P_0(w))^2}{P_0(w)} = 2^m \sum_w P_1^2(w) - 1 \\ &= E_0 \left( \frac{P_1(w)}{P_0(w)} - 1 \right)^2, \end{aligned} \tag{7}$$

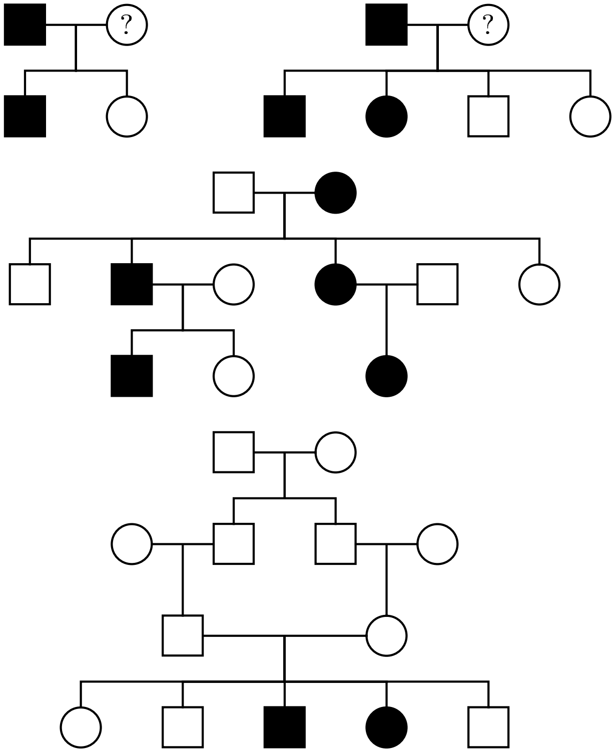

**Fig. 1.** Four pedigrees used in simulations, upper left (1), upper right (2), middle (3) and lower (4). Males and females are depicted as squares and circles. Affected individuals have black and unaffected ones have white symbols. Individuals with unknown phenotypes have question marks.

where $E_0$ denotes expectation under $P_0$. The sum ranges over $\mathbb{Z}_2^m$, the space of all binary vectors of length $m$, which is a group under component-wise modulo 2 addition. The second quantity measures the non-uniformity of $P_1$ as

$$I^{\text{est}} = \sum_{w \sim w'} \frac{(P_1(w') - P_1(w))^2}{P_1(w) + P_1(w')}, \tag{8}$$

where the sum ranges over all $m \cdot 2^{m-1}$ (unordered) pairs $w, w' \in \mathbb{Z}_2^m$ which are neighbours ($w \sim w'$), i.e. differ at exactly one bit. We use the convention $0/0 = 0$ in (8).

Figure 1 depicts binary phenotypes for four pedigrees of various size. The phenotypes of the first three show a clear autosomal dominant inheritance pattern whereas the last one corresponds to an autosomal recessive disease. See also page 39 of Haines and Paricak-Vance (1998). In Figures 2 and 3, $\log_2(I^{\text{test}} + 1)$ and $I^{\text{est}}$ are plotted as function of disease allele frequency $p_1$ for various penetrance models. It is seen that presence of phenocopies ($\psi_0 > 0$) reduces informativity of pedigrees a lot, especially for small $p_1$. In most cases informativity increases as $p_1$

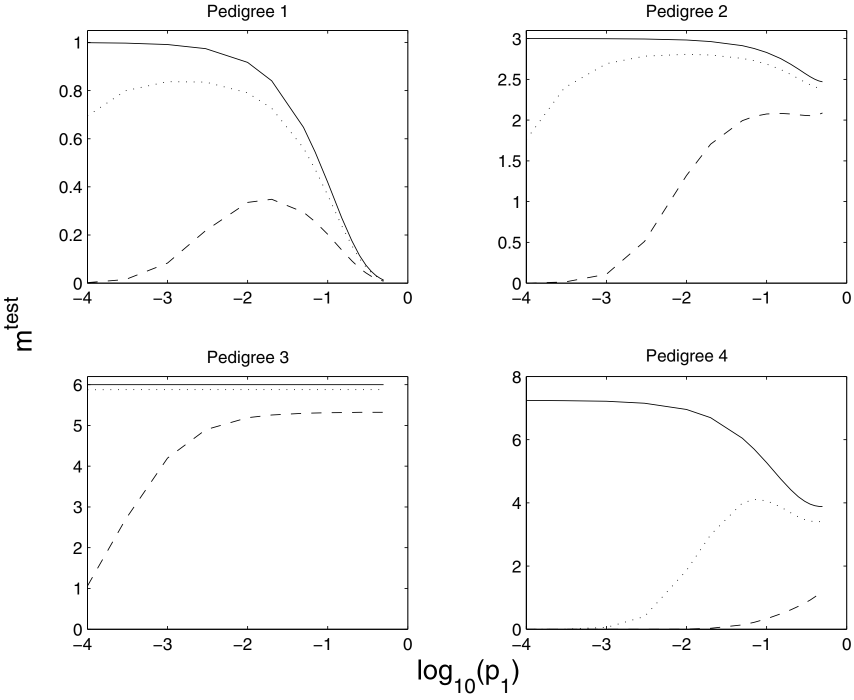

**Fig. 2.** Effective number $m^{\text{test}} = \log_2(I^{\text{test}} + 1)$ of meioses for testing as function of disease allele frequency $p_1$. The pedigrees, with binary phenotypes, are given in Figure 1. For the first three pedigrees the penetrance parameters $\psi$ correspond to a dominant trait; (0, 1, 1) (solid), (0.02, 1, 1) (dotted) and (0.1, 1, 1) (dashed). The fourth pedigree has recessive penetrance parameters; $(0, 0, 1)$ (solid), $(0.02, 0.02, 1)$ (dotted) and $(0.1, 0.1, 1)$ (dashed).

gets small. The reason is that phenotypes give more information about genotypes as $p_1$ decreases.

## 3. Information Bounds for $N$ pedigrees

Assume we have phenotype and marker data from $N$ pedigrees. Let $m_i$, $Y_i$, $v_i(t)$, $P_{0i}$ and $P_{1i}$ denote number of meioses, phenotype vector, inheritance vector at locus $t$, prior and posterior distributions (1) and (2) for the $i^{\text{th}}$ pedigree. Let $\boldsymbol{Y} = (Y_1, \dots, Y_N)$, $\boldsymbol{v}(t) = (v_1(t), \dots, v_N(t))$, $\boldsymbol{v} = \boldsymbol{v}(\tau)$ and $\boldsymbol{w} = (w_1, \dots, w_N)$. Assuming independent phenotype and marker data between pedigrees we get

$$\begin{aligned} P_0(\boldsymbol{w}) &:= P_0(\boldsymbol{v}(t) = \boldsymbol{w}) = \textstyle\prod_{i=1}^N P_{0i}(w_i) = 2^{-m}, \\ P_1(\boldsymbol{w}) &:= P_1(\boldsymbol{v} = \boldsymbol{w}|\boldsymbol{Y}) = \textstyle\prod_{i=1}^N P_{1i}(w_i), \end{aligned}$$

where $m = \sum_{i=1}^N m_i$ is the total number of meioses and $P_{0i}$ and $P_{1i}$ the distributions (1) and (2) for the $i^{\text{th}}$ pedigree. The information bounds for $N$ pedigrees are

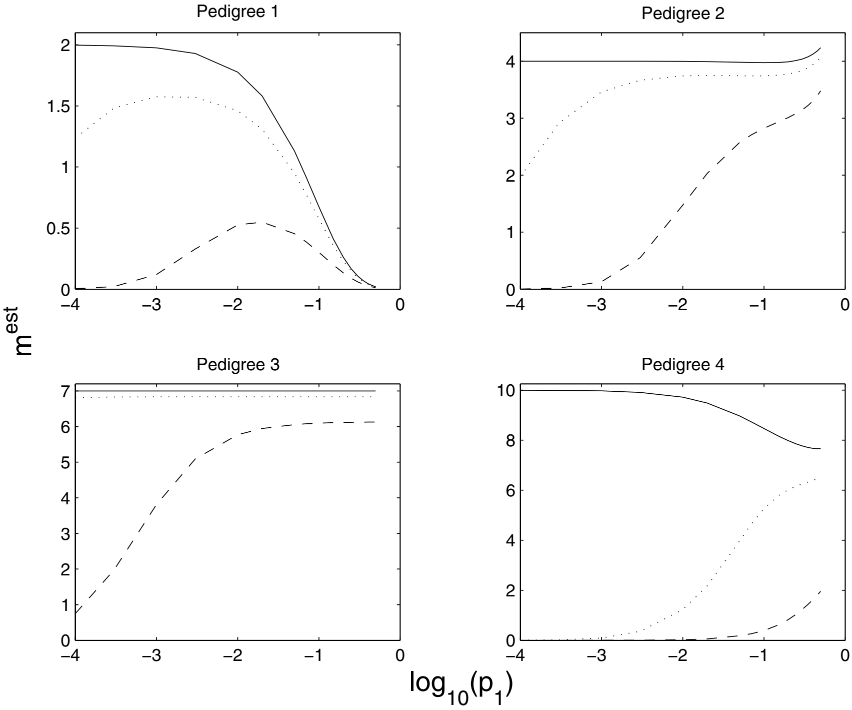

**Fig. 3.** Effective number $m^{\text{est}} = I^{\text{est}}$ of meioses for estimation as function of disease allele frequency $p_1$. For details on pedigrees and penetrance parameters, see Figures 1 and 2.

natural generalizations of (7) and (8), given by

$$I^{\text{test}} = \sum_{\boldsymbol{w}} \frac{(P_1(\boldsymbol{w}) - P_0(\boldsymbol{w}))^2}{P_0(\boldsymbol{w})} = 2^m \sum_{\boldsymbol{w}} P_1^2(\boldsymbol{w}) - 1. \tag{9}$$

and

$$I^{\text{est}} = \sum_{\boldsymbol{w} \sim \boldsymbol{w}'} \frac{(P_1(\boldsymbol{w}') - P_1(\boldsymbol{w}))^2}{P_1(\boldsymbol{w}) + P_1(\boldsymbol{w}')}, \tag{10}$$

respectively. Here $\boldsymbol{w} \sim \boldsymbol{w}'$ means that the two vectors $\boldsymbol{w}$ and $\boldsymbol{w}'$ of length $m$ differ at exactly one bit, i.e. $w_i \sim w_i'$ some $i \in \{1, \dots, N\}$ and $w_i = w_i'$ for all other $i$. The following result can now be deduced:

**Lemma 1.** *The information bounds $I^{test}$ and $I^{est}$ in (9) and (10) are given by*

$$I^{test} = \prod_{i=1}^{N} (1 + I_i^{test}) - 1$$

*and*

$$I^{est} = \sum_{i=1}^{N} I_i^{est}$$

*respectively, where $I_i^{test}$ and $I_i^{est}$ are information bounds for the $i^{th}$ pedigree, obtained by replacing $P_0$ and $P_1$ by $P_{0i}$ and $P_{1i}$ in (7) and (8).*

We will now motivate the relevance of $I^{\text{test}}$ and $I^{\text{est}}$ for linkage analysis. Let $S : \mathbb{Z}_2^m \to \mathbb{R}$ be a score function, with large values of $S(\boldsymbol{w}) = S(\boldsymbol{w}; \boldsymbol{Y})$ indicating high compatibility between $\boldsymbol{Y}$ and the inheritance vector $\boldsymbol{w}$. We assume that $S$ has been centered so that $E_0(S(\boldsymbol{v}(t))) = 2^{-m} \sum_{\boldsymbol{w}} S(\boldsymbol{w}) = 0$. Then, under perfect marker information, a linkage score for all families at locus $t$ is

$$Z(t) = S(\boldsymbol{v}(t))/\sigma_0, \tag{11}$$

where $\sigma_0^2 = V_0(S(\boldsymbol{v}(t))) = 2^{-m} \sum_{\boldsymbol{w}} S^2(\boldsymbol{w})$ and $V_0$ denotes variance under $P_0$. Notice that $Z(t)$ is defined conditionally on observed phenotypes. The random variation of $Z(t)$ comes solely from marker data $\boldsymbol{v}(t)$. A special case of (11) is

$$Z(t) = \sum_{i=1}^{N} \gamma_i Z_i(t), \tag{12}$$

i.e. the total linkage score $Z(t)$ is a linear combination of family scores $Z_i(t) = S_i(v_i(t))/\sigma_{0i}$, cf. Kruglyak et al. (1996). The weights $\gamma_i$ satisfy the constraint $\sum_{i=1}^{N} \gamma_i^2 = 1$ and $S_i(w_i) = S_i(w_i; Y_i)$ is a score function for the $i^{\text{th}}$ pedigree. We assume that $S_i$ is centered so that $E_0(S_i(v(t)) = 0$ and $\sigma_{0i}^2 = V_0(S(v_i(t)))$.

Suppose we wish to test

$$\begin{aligned} H_0 &: \tau \notin \Omega, \\ H_1 &: \tau \in \Omega, \end{aligned}$$

where $\Omega$ is a region, consisting of one or several chromosomes. Since allele transmissions at unlinked loci are independent it follows that under $H_0$, $\boldsymbol{v}(t)|\boldsymbol{Y} \sim P_0$ at all $t \in \Omega$. Because of the definition of $S$ we thus have $E(Z(t)) = 0$ and $V(Z(t)) = 1$ under $H_0$ and perfect marker information.

As test statistic we use

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

and $H_0$ is rejected as soon as $Z_{\max}$ exceeds a given threshold $T$. As performance criterion we might use the power $\beta = P(Z_{\max} \geq T|H_1)$. It depends on the chosen threshold $T$ and the size of $\Omega$. Another criterion (Feingold et al., 1993), which is independent of these quantities, is the noncentrality parameter

$$\text{NCP} = E(Z(\tau)).$$

It is the maximal value of $E(Z(t))$ attained at $\tau$. When $\beta$ is viewed as a function of the significance level $\alpha = P(Z_{\max} \geq T|H_0)$ (rather than the threshold $T$), it is essentially determined by NCP, at least when $N$ is so large that $Z$ is well approximated by a Gaussian process (Hössjer, 2003d). For data sets with a few large pedigrees, skewness of $Z(t)$ under $H_0$ may sometimes increase $T$ compared to values suggested by a Gaussian process approximation. This decreases $\beta$ and makes NCP somewhat less accurate as surrogate for power.

The following result gives upper bounds for NCP and $I^{\text{test}}$:

**Proposition 1.** *With a general score function (11), the maximum possible noncentrality parameter is*

$$\sup_S NCP = \sqrt{I^{test}}, \tag{13}$$

*and the maximum is attained for $S \propto P_1 - P_0$. Restricting ourselves to linear score functions (12), the maximal noncentrality parameter is*

$$\sup_{\{S_i\},\{\gamma_i\}} NCP = \sqrt{\sum_{i=1}^{N} I_i^{test}}, \tag{14}$$

*and the maximum is attained for $S_i \propto P_{1i} - P_{0i}$ and weights $\gamma_i \propto NCP_i = E(Z_i(\tau))$.*

It follows from Lemma 1 that the right hand side of (13) is strictly larger than (14) if $I_i^{\text{test}} > 0$ for at least two $i$. The difference can be notable for complete penetrance models (Section 4) but is small for weak penetrance models (Section 5).

Given $H_1$, it is of interest to locate the position of $\tau$ as well as possible. It turns out that $I^{\text{est}}$ is closely related to the size of confidence regions. Such a confidence region can be defined e.g. as

$$\tilde{\Omega} = \{t;\ Z_{\max} - Z(t) \leq \tilde{T}\}$$

for some threshold $\tilde{T}$ controlling the coverage probability $P(\tau \in \tilde{\Omega})$, see Siegmund (1986) and Kruglyak and Lander (1995). It turns out that asymptotically for large samples $N$, it is the local behaviour of $Z(\cdot)$ around $\tau$ that determines the length of the confidence region. It is shown in Hössjer (2003a) that constants $a, \sigma^2 > 0$ exists such that

$$\begin{aligned} E(Z(\tau) - Z(t)) &= a|t - \tau| + o(|t - \tau|), \\ V(Z(\tau) - Z(t)) &= \sigma^2|t - \tau| + o(|t - \tau|) \end{aligned} \tag{15}$$

as $t \to \tau$. We refer to $a$ as the mean slope and $\sigma^2$ as the diffusion coefficient at the disease locus. The slope-to-noise ratio is defined as

$$\text{SLNR} = a^2/\sigma^2. \tag{16}$$

It is shown in Hössjer (2003b) that the expected length of the confidence region depends (essentially) on the coverage probability and $\text{SLNR}^{-1}$. For instance, a 95% (50%) confidence region has asymptotically expected length $3.11 \cdot \text{SLNR}^{-1}$ ($0.63 \cdot \text{SLNR}^{-1}$) Morgans for weak genetic models as $N \to \infty$. For this reason, it is of interest to find an upper bound for SLNR.

**Proposition 2.** *Upper bounds for the slope-to-noise ratio are*

$$\sup_{\{S_i\},\{\gamma_i\}} SLNR = \sum_{i=1}^{N} \sup_{\{S_i\}} SLNR_i \leq \sup_S SLNR \leq I^{est}, \tag{17}$$

**Table 1.** Values of $\sup_{S_i}$ $\mathrm{SLNR}_i$ and $I_i^{\mathrm{est}}$ for various genetic models and phenotype combinations. The parent's phenotypes are unknown and the listed phenotypes are those of the sibs.

| Model | Pedigree | $p_1$ | $\psi$ | Phenotypes | $\sup_{S_i}$ $\mathrm{SLNR}_i$ | $I_i^{\mathrm{est}}$ |
|---|---|---|---|---|---|---|
| Binary | Sib pair | 0.01 | (0,0,1) | (1,1) | 3.8432 | 3.8432 |
| | | 0.5 | | | 0.4444 | 0.4444 |
| | | 0.01 | (0,1,1) | | 1.2024 | 1.2024 |
| | | 0.5 | | | 0.0229 | 0.0229 |
| | Sib quartet | 0.01 | (0,0,1) | (0,0,0,1) | 3.4013 | 3.4013 |
| | | 0.5 | | | 3.1621 | 3.2000 |
| | | 0.01 | (0,1,1) | (0,0,1,1) | 3.6551 | 3.6620 |
| | | 0.5 | | | 3.7362 | 3.7653 |
| Gaussian | Sib pair | 0.01 | (0,1,2,1,0,0) | (2,2) | 0.0090 | 0.0090 |
| | | 0.5 | | | 0.0144 | 0.0144 |
| | Sib quartet | 0.01 | | (-2,-2,2,2) | 0.0129 | 0.0129 |
| | | 0.5 | | | 1.5247 | 1.5296 |

*where the first supremum is for linkage scores (12) and the second for general linkage scores (11). In the former case, the optimal weighting scheme is $\gamma_i \propto a_i^2/\sigma_i^2$ and $SLNR_i = a_i^2/\sigma_i^2$ is the slope-to-noise ratio (16) for the $i^{th}$ pedigree. Here $a_i$ and $\sigma_i^2$ are the mean slopes and diffusion coefficients obtained by replacing $Z(\cdot)$ with $Z_i(\cdot)$ in (15).*

It turns out that $\sup_{S_i}$ $\mathrm{SLNR}_i$ and $I_i^{\mathrm{est}}$ are very close, see Table 1 for some examples[2]. This indicates that the left and right hand sides of (17) are very close. Therefore, it is no essential restriction for estimation purposes to consider the reduced class of linkage scores (12).

It is clear from (15) that $\mathrm{SLNR} = a^2/\sigma^2$ depends on fluctuations of $Z(\cdot)$, caused by crossovers in close vicinity of $\tau$. In the appendix, explicit formulas for $a$ and $\sigma^2$ are given in (A.1). Equivalent formulations are $a = m \cdot E_1(S(\boldsymbol{w}) - S(\boldsymbol{w}'))/\sigma_0$ and $\sigma^2 = m \cdot E_1((S(\boldsymbol{w}') - S(\boldsymbol{w}))^2)/\sigma_0^2$, where $\boldsymbol{w}$ has distribution $P_1$ and given $\boldsymbol{w}$, $\boldsymbol{w}'$ is uniformly distributed on all $m$ neighbours $\boldsymbol{w}' \sim \boldsymbol{w}$ that result when a crossover occurs in $\boldsymbol{w}$. Hence, for a standardized score function $S$ with $\sigma_0^2 = 1$, $a$ and $\sigma^2$ are the intensity of crossovers $m$ times the average increment and the average squared increment respectively of $S$ caused by one crossover $\boldsymbol{w} \to \boldsymbol{w}'$ under $H_1$. In the upper bound $I^{\mathrm{est}}$ of $a^2/\sigma^2$ in formula (10), each of the $m \cdot 2^{m-1}$ possible pairs $\boldsymbol{w} \sim \boldsymbol{w}'$ in the sum corresponds to two crossovers, either $\boldsymbol{w} \to \boldsymbol{w}'$ or $\boldsymbol{w}' \to \boldsymbol{w}$.

## 4. Complete Penetrance Models

In this section we derive formulas for $I^{\mathrm{test}}$ and $I^{\mathrm{est}}$ under the complete penetrance model of Example 1. To begin with, we consider one pedigree and omit family index $i$.

[2] In fact, $\sup_{S_i}$ $\mathrm{SLNR}_i$ was used as definition of information in Hössjer (2003a). However, we prefer to use $I_i^{\mathrm{est}}$ here because this quantity has a simpler and more easily interpretable expression.

Let $\mathcal{A}$ and $\mathcal{B}$ be the sets of individuals $k$ that are homozygotes ($a_{2k-1} = a_{2k}$) and heterozygotes with unknown phase ($a_{2k-1} \neq a_{2k}$ but the parental origin of $a_{2k-1}$ and $a_{2k}$ is unknown) respectively. For each $k$ we also let $\text{Par}_k$ and $\text{Off}_k$ denote the set of parents and offspring of $k$ that belong to the pedigree. Notice that either $\text{Par}_k$ or $\text{Off}_k$ may be empty. For instance, $\text{Par}_k = \emptyset$ when $k$ is founder. Then define the set

$$\mathbb{C} = \left\{ \sum_{k \in \mathcal{A}} \sum_{j \in \text{Off}_k} \alpha_{kj} 1_j + \sum_{k \in \mathcal{B}} \alpha_k 1_{\text{Par}_k \cup \text{Off}_k}, \;\; \alpha_k, \alpha_{kj} \in \{0, 1\} \right\}, \tag{18}$$

where addition is component-wise modulo 2, $1_A$ is a vector with ones in positions $A \subset \{1, \dots, m\}$ and zeros elsewhere and $1_j$ is short for $1_{\{j\}}$. In other words, $\mathbb{C}$ is the linear span, in $\mathbb{Z}_2^m$, of all vectors $1_j$ and $1_{\text{Par}_k \cup \text{Off}_k}$ appearing in (18).

**Theorem 1.** *Consider one pedigree. Assume random mating (4) and that the genotype vector $G$ is consistent with Mendelian inheritance, i.e. there is at least one pair $(a, v)$ such that $G$ is obtained when the founder alleles $a$ are segregated according to $v$ ($(a, v) \rightarrow G$).Then*

$$P_1(w) = |\mathbb{C}|^{-1} 1_{\{w \in w_0 + \mathbb{C}\}} \tag{19}$$

*for some coset $w_0 + \mathbb{C} = \{w_0 + w;\; w \in \mathbb{C}\}$ depending on $G$.*

Given the expression (19) for $P_1$, we can easily plug it into (7) and (8) and derive the following formulas for $I^{\text{test}}$ and $I^{\text{est}}$:

**Corollary 1.** *Under the assumptions of Theorem 1 we have*

$$I^{test} = 2^{m^{test}} - 1, \tag{20}$$

*where $m^{test} = m - \dim(\mathbb{C})$ and $\dim(\mathbb{C})$ is the dimension of $\mathbb{C}$. Further,*

$$I^{est} = m^{est}, \tag{21}$$

*where $m^{est}$ is the number of inheritance vectors $w \sim 0$ such that $w \notin \mathbb{C}$.*

We interpret $m^{\text{test}}$ and $m^{\text{est}}$ as the effective number of meioses in the pedigree for testing and estimation respectively.

*Example 4 (Fully polymorphic disease locus.)* The disease gene is fully polymorphic for a pedigree if all founder alleles $a_1, \dots, a_{2f}$ are distinct. This event occurs with probability one in the limit of a large number of disease alleles, each having small probability ($M \rightarrow \infty$, $\max_{0 \leq i \leq M-1} p_i \rightarrow 0$). For a pedigree without loops this means that all individuals are heterozygous. It is shown in the appendix that

$$m^{\text{test}} = m - f, \quad m^{\text{est}} = m - f' \tag{22}$$

if the phases of all founder genotypes are unknown. Here $f'$ is the number of founders with precisely one offspring. The formula for $m^{\text{test}}$ agrees with the founder phase symmetry reduced number of meioses used in multipoint linkage analysis.

See Kruglyak et al. (1996) for details. If the phase of all founders is known (this requires allelic information from previous generations) one has

$$m^{\text{test}} = m^{\text{est}} = m. \tag{23}$$

□

*Example 5 (Backcross.)* Consider a nuclear family with $n_{\text{off}}$ offspring. The father is heterozygous (01) and the mother homozygous (00). We assume that $n_{\text{het}}$ offspring are heterozygous (01) and the remaining $n_{\text{off}} - n_{\text{het}}$ offspring homozygous (00). It is shown in the appendix that

$$m^{\text{test}} = m^{\text{est}} = n_{\text{off}} \tag{24}$$

and

$$m^{\text{test}} = n_{\text{off}} - 1, \quad m^{\text{est}} = n_{\text{off}} 1_{\{n_{\text{off}}>1\}} \tag{25}$$

if the father has known and unknown phase respectively. □

*Example 6 (Intercross.)* In the previous example, assume instead that both the father and mother are heterozygous (01) and $n_{\text{het}}$ offspring are heterozygous (01). The remaining $n_{\text{off}} - n_{\text{het}}$ homozygous offspring can have two genotypes, either (00) or (11). It is shown in the appendix that

$$m^{\text{test}} = 2n_{\text{off}} - n_{\text{het}}, \quad m^{\text{est}} = 2n_{\text{off}} \tag{26}$$

and

$$m^{\text{test}} = 2n_{\text{off}} - n_{\text{het}} - 2 + 1_{\{n_{\text{het}}=n_{\text{off}}\}}, \quad m^{\text{est}} = 2n_{\text{off}} 1_{\{n_{\text{off}}>1\}} \tag{27}$$

if the parents have known and unknown phase respectively. □

We end this section by considering $N$ pedigrees. Lemma 1 and Corollary 1 suggest that the effective number of meioses for testing and estimation should be defined as

$$\begin{aligned} m^{\text{test}} &= \log_2(I^{\text{test}} + 1) = m + \log_2\left(\sum_{\boldsymbol{w}} P_1^2(\boldsymbol{w})\right) \\ m^{\text{est}} &= I^{\text{est}} \end{aligned} \tag{28}$$

for general genetic models, not necessarily those with complete penetrance. By combining Corollary 1 with Lemma 1, we immediately get:

**Corollary 2.** *The effective number of meioses (28) for testing and estimation are added over families, i.e.* $m^{test} = \sum_{i=1}^{N} m_i^{test}$ *and* $m^{est} = \sum_{i=1}^{N} m_i^{est}$. *Here* $m_i^{test} = \log_2(I_i^{test} + 1)$ *and* $m_i^{est} = I_i^{est}$ *are the effective number of meioses for testing and estimation in the* $i^{th}$ *pedigree. These are always non-negative integers for complete penetrance models.*

Notice that $m^{\text{test}}$ and $m^{\text{est}}$ tend to integer values as $p_1 \to 0$ for the dominant models of pedigrees 1–3 when there are no phenocopies ($\psi = (0, 1, 1)$). The reason is that $p_1 \to 0$ corresponds to a complete penetrance model with $Y_k = 0 \Rightarrow G_k = (00)$ and and $Y_k = 1 \Rightarrow G_k = (01)$. This is not the case for the recessive model $\psi = (0, 0, 1)$ without phenocopies in pedigree 4. In this case, when $p_1 \to 0$, $Y_k = 0$ implies either $G_k = (00)$ or $(01)$ for the three unaffected children of the last generation. Further, it is not certain which of the two grandparents in the first generation that carries the disease allele.

## 5. Weak Penetrance Models

As in the previous section we first consider one fix pedigree and drop index $i$. The genetic model consists of disease allele frequencies $p = (p_0, \ldots, p_{M-1})$ and penetrance parameters $\psi$. The stronger the genetic model, the more information $Y$ carries about $v = v(\tau)$. As a result, the distribution $P_1$ is less uniform.

In this section we let the penetrance parameters $\psi = \psi_\varepsilon$ depend on the scalar $\varepsilon$. When $\varepsilon = 0$ there is no genetic effect at $\tau$, meaning that $P(Y|a, v) = P(Y)$ is independent of $a$ and $v$ in (3) and $P_1 = P_0$. The larger $\varepsilon$ is, the stronger is the genetic component and the more $P_1$ departs from $P_0$. By essentially Taylor expanding (3) with respect to $\varepsilon$ around $\varepsilon = 0$, it has been shown (McPeek, 1999, Hössjer, 2003c, 2003e) that

$$P_1(w) = 2^{-m}(1 + \varepsilon^\rho S_{\text{opt}}(w)) + o(\varepsilon^\rho) \tag{29}$$

as $\varepsilon \to 0$ for some positive integer $\rho$ and score function $S_{\text{opt}}$. It turns out that $S_{\text{opt}}$ is a locally optimal score function for small $\varepsilon$ in terms of maximizing both NCP and SLNR (Hössjer, 2003a). For pedigrees without loops (outbred pedigrees) one has $\rho = 2$ and

$$S_{\text{opt}}(w) = \sum_{k<l} \omega_{kl} S_{kl}(w), \tag{30}$$

for many genetic models. The sum in (30) ranges over all pairs of individuals with known phenotype. Here

$$S_{kl}(w) = (1 - c) \cdot (\text{IBD}_{kl}(w)/2 - r_{kl}) + c \cdot (1_{\{\text{IBD}_{kl}(w)=2\}} - \delta_{kl}) \tag{31}$$

and $c$ is the fraction of genetic variance at the main locus $\tau$ due to dominance effects. The weights $\omega_{kl}$ depend on phenotypes and genetic model parameters. McPeek (1999) derived (30) for binary phenotypes and Hössjer (2003c, 2003e) for general genetic models.

*Example 7 (Gaussian phenotypes.).* Continuing Example 3, let $m_0^*$, $m_1^*$ and $m_2^*$ be the first three penetrance parameters of a fixed reference model. Then $m = E(Y_k) = p_0^2 m_0^* + 2p_0 p_1 m_1^* + p_1^2 m_2^*$ is the phenotype mean and $\sigma_g^2 = \text{Var}(m^*_{|G_k|}) = E(m^*_{|G_k|} - m)^2$ the genetic variance at $\tau$. Consider a family $\{\psi_\varepsilon\}_{\varepsilon \geq 0}$ of genetic models

$$\begin{aligned}\psi_\varepsilon = (m, m, m, \sigma^2, h_a^2, h_d^2) + \varepsilon(\sigma/\sigma_g)(m_0^* \\ -m, m_1^* - m, m_2^* - m, 0, 0, 0).\end{aligned} \tag{32}$$

Notice that the phenotype mean is $m$ for all $\varepsilon$. When $\varepsilon$ grows, only the first three parameters are changed, starting from a model with no genetic effects at $\varepsilon = 0$ and going in direction towards the reference model. The main locus heritability $h^2 = \text{Var}(m_{|G_k|})/\text{Var}(Y_k) = \varepsilon^2\sigma^2/(\varepsilon^2\sigma^2 + \sigma^2) = \varepsilon^2/(1+\varepsilon^2)$ gives a natural interpretation of $\varepsilon$. Let $r = (Y - m)/\sigma = (r_1, \ldots, r_n)$ be the vector of standardized residuals. Then (30) holds with

$$\omega_{kl} = (rC^{-1})_k(rC^{-1})_l - C_{kl}^{-1} \overset{h_a^2=h_d^2=0}{=} r_k r_l, \tag{33}$$

where $C_{kl}^{-1}$ is the $(k,l)^{\text{th}}$ entry of the inverse correlation matrix in (6). See Commenges (1994) in the special case of no polygenic effects and Tang and Siegmund (2001) and Hössjer (2003c) for the general case. The reference model's genetic variance at $\tau$ can be split into additive and dominance components, $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$. Here $\sigma_a^2 = 2p_0p_1(p_1(m_2^* - m_1^*) + p_0(m_1^* - m_0^*))^2$, $\sigma_d^2 = (p_0p_1)^2(m_2^* - 2m_1^* + m_0^*)^2$ and $c = \sigma_d^2/\sigma_g^2$ in (31).

Figures 4–5 display $m^{\text{test}}$ and $m^{\text{est}}$ for a trajectory (32) of penetrance parameters as function of main locus heritability. It is seen that the number of sibs, the phenotypes, and the amount of polygenic variance all influence $m^{\text{test}}$ and $m^{\text{est}}$ greatly. This will be further discussed later on in this section. □

*Example 8 (Binary phenotypes.)* In Example 2, let $\psi_0^*$, $\psi_1^*$ and $\psi_2^*$ be the penetrance parameters of a fixed reference model with prevalence $K_p = P(Y_k = 1) = (1-p_0)^2\psi_0^* + 2p_0p_1\psi_1^* + p_1^2\psi_2^*$. Then introduce the family of genetic models $\{\psi_\varepsilon\}_{0\le\varepsilon\le\varepsilon_{\max}}$ by

$$\psi_\varepsilon = (K_p, K_p, K_p) + \varepsilon(K_p/\sigma_g)(\psi_0^* - K_p, \psi_1^* - K_p, \psi_2^* - K_p),$$

where $\sigma_g^2 = \text{Var}(\psi_{|G_k|}^*) = E(\psi_{|G_k|}^* - K_p)^2$ is the genetic variance at $\tau$. The upper bound $\varepsilon_{\max}$ guarantees that all components of $\psi_\varepsilon$ are probabilities. The prevalence is $K_p$ for all $\varepsilon$ and the genetic component grows with $\varepsilon$ so that the relative risk ratio $P(Y_k = 1|Y_l = 1)/P(Y_k = 1)$ of a monozygotic twin pair $(k,l)$ is $1+\varepsilon^2$, cf. Risch (1990a). In this case (30) holds with

$$\omega_{kl} = (Y_k - K_p)(Y_l - K_p)/(1-K_p)^2, \tag{34}$$

see McPeek (1999). As for Gaussian phenotypes, the genetic variance $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$ at $\tau$ can be split into additive and dominance components. Here $\sigma_a^2 = 2p_0p_1(p_1(\psi_2^* - \psi_1^*) + p_0(\psi_1^* - \psi_0^*))^2$, $\sigma_d^2 = (p_0p_1)^2(\psi_2^* - 2\psi_1^* + \psi_0^*)^2$ and $c = \sigma_d^2/\sigma_g^2$ in (31). □

By inserting (30) into (29), we get the following result:

**Proposition 3.** *Consider one pedigree. Assume a weak penetrance expansion (29) with $\rho = 2$ and $S_{opt}$ as in (30). Then*

$$I^{test} = \omega\Sigma^{test}\omega' \cdot \varepsilon^4 + o(\varepsilon^4) \tag{35}$$

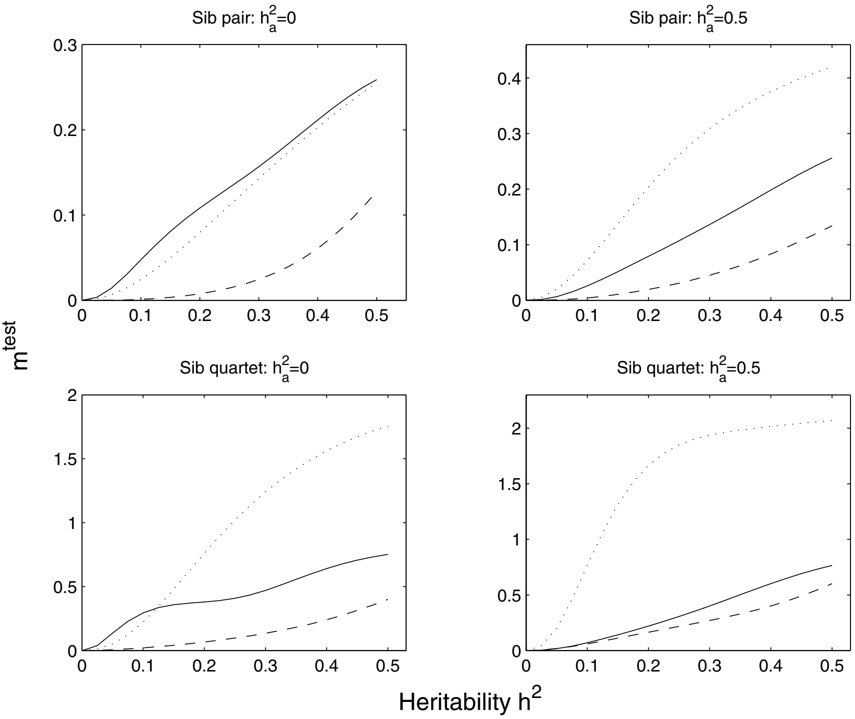

**Fig. 4.** Effective number $m^{\text{test}} = \log_2(I^{\text{test}} + 1)$ of meioses for testing as function of main locus heritability $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma^2)$. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). At the main locus, it's disease allele frequency $p_1 = 0.1$ and it's penetrance parameters are $m_0^* = 0$, $m_1^* = 1$, $m_2^* = 2$ and $\sigma = 1$. Hence $m = E(Y_k) = 0.2$. The parents have unknown phenotypes and the children in the sib pair family have phenotypes (2.2,2.2) (solid), (-1.8 2.2) (dotted) and (0.2,2.2) (dashed). The children in the sib quartet family have phenotypes (2.2,2.2,2.2,2.2) (solid), (-1.8,-1.8,2.2,2.2) (dotted) and (-1.8,0.2,0.2,2.2) (dashed).

*and*

$$I^{est} = \omega\Sigma^{est}\omega' \cdot \varepsilon^4 + o(\varepsilon^4) \tag{36}$$

*as* $\varepsilon \to 0$. *Here* $\omega = (\omega_{kl}; k < l)$ *is a row vector with indexes ranging over all pairs of individuals with known phenotype. Further,* $\Sigma^{test} = (\Sigma^{test}_{kl,k'l'})$ *and* $\Sigma^{est} = (\Sigma^{est}_{kl,k'l'})$ *are matrices with entries*

$$\Sigma^{test}_{kl,k'l'} = 2^{-m} \sum_w S_{kl}(w)S_{k'l'}(w)$$

*and*

$$\Sigma^{est}_{kl,k'l'} = 2^{-m-1} \sum_{w\sim w'} (S_{kl}(w') - S_{kl}(w))(S_{k'l'}(w') - S_{k'l'}(w))$$

*respectively.*

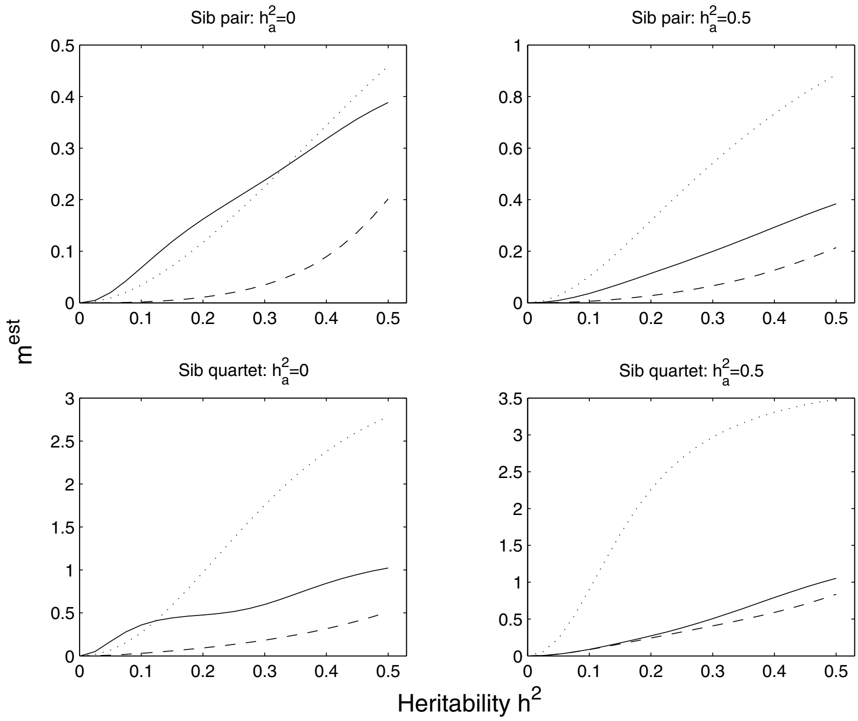

**Fig. 5.** Effective number $m^{\text{est}} = I^{\text{est}}$ of meioses for estimation as function of main locus heritability $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma^2)$. See Figure 4 for more details.

An expression for $\omega\Sigma^{\text{test}}\omega'$ appears in Tang and Siegmund (2001) for Gaussian phenotypes in the special case $c = 0$. In general, the phenotypes enter into (35) and (36) only through $\omega$. The pedigree structure, on the other hand, enter into the matrices $\Sigma^{\text{test}}$ and $\Sigma^{\text{est}}$. Table 2 shows values of the diagonal elements $\Sigma_{kl,kl}^{\text{test}}$ and $\Sigma_{kl,kl}^{\text{est}}$ for various relative pairs $(k, l)$. It is seen that these decrease rapidly with distance of relationship.

In most cases, at least some nondiagonal entries of $\Sigma^{\text{test}}$ and $\Sigma^{\text{est}}$ are nonzero. An important exception, however, occurs for a nuclear family with 2 parents and $n_{\text{off}} = n - 2$ children. Then $\Sigma_{kl,kl}^{\text{test}} = \Sigma_{kl,kl}^{\text{est}} = 0$ if a least one of $k$ and $l$ is a parent. Therefore, it suffices to consider sib pairs $(k, l)$. It can be shown that $\Sigma^{\text{test}}$ and $\Sigma^{\text{est}}$ are proportional to identity matrices of order $n_{\text{off}}(n_{\text{off}} - 1)/2$, with diagonal entries taken from Table 2. Plugging this into (35)–(36) we get

$$\begin{aligned} \omega\Sigma^{\text{test}}\omega' &= (0.125 + 0.0625 \cdot c^2) \sum_{k<l} \omega_{kl}^2, \\ \omega\Sigma^{\text{est}}\omega' &= (0.25 + 0.25 \cdot c^2) \sum_{k<l} \omega_{kl}^2, \end{aligned} \tag{37}$$

where the sums range over all $n_{\text{off}}(n_{\text{off}} - 1)/2$ sib pairs $(k, l)$. Hence the amount of testing and estimation information depends only on the dominance ratio $c$ and the

**Table 2.** Values of $\Sigma^{\text{test}}_{kl,kl}$ and $\Sigma^{\text{est}}_{kl,kl}$ for various pairs of relatives $(k, l)$ as function of the fraction $c$ of dominance variance at the main locus.

| $(k, l)$ | $r_{kl}$ | $\delta_{kl}$ | $\Sigma^{\text{test}}_{kl,kl}$ | $\Sigma^{\text{est}}_{kl,kl}$ |
|---|---|---|---|---|
| Parent-offspring | 0.5 | 0 | 0 | 0 |
| Sibs | 0.5 | 0.25 | $0.125 + 0.0625 \cdot c^2$ | $0.25 + 0.25 \cdot c^2$ |
| Grandp-grandch | 0.25 | 0 | $0.0625 \cdot (1 - c)^2$ | $0.0625 \cdot (1 - c)^2$ |
| Uncle-nephew | 0.25 | 0 | $0.0625 \cdot (1 - c)^2$ | $0.15625 \cdot (1 - c)^2$ |
| First cousins | 0.125 | 0 | $0.0469 \cdot (1 - c)^2$ | $0.125 \cdot (1 - c)^2$ |
| Second cousins | 0.03125 | 0 | $0.0146 \cdot (1 - c)^2$ | $0.046875 \cdot (1 - c)^2$ |
| Double first cousins | 0.25 | 0.0625 | $0.0562 + 0.0441 \cdot (c - 0.8)^2$ | $0.1875 + 0.25 \cdot (c - 0.5)^2$ |

sum $\sum_{k<l} \omega_{kl}^2$ for sibling families and weak penetrance models. An informative sib pair is one with large $\omega_{kl}^2$. For instance, this confirms the well known fact that affected sib pairs are most informative for binary traits when $p_1$ and $K_p$ are small, cf. (34) and Risch (1990b). For Gaussian phenotypes, it follows from (33) that concordant (both $Y_k$ and $Y_l$ large) or discordant ($Y_k$ large, $Y_l$ small or vice versa) sib pairs are the most informative ones. The larger the residual correlation is because of additive/dominant polygenic variance, the more informative are discordant sib pairs in relation to concordant ones. See Figures 4–5 and Risch and Zhang (1995).

We close this section by considering $N$ pedigrees. The following result can easily be deduced from Lemma 1 and Proposition 3:

**Corollary 3.** *The conclusions of Proposition 3 remain valid for N pedigrees, provided the quadratic forms in (35) and (36) are replaced by* $\sum_{i=1}^{N} \omega_i \Sigma_i^{test} \omega_i'$ *and* $\sum_{i=1}^{N} \omega_i \Sigma_i^{est} \omega_i'$ *respectively. Here* $\omega_i$, $\Sigma_i^{test}$ *and* $\Sigma_i^{est}$ *are the values of* $\omega$, $\Sigma^{test}$ *and* $\Sigma^{est}$ *for the* $i^{th}$ *pedigree.*

## 6. Conclusions

In this paper, we have introduced two information measures, $I^{\text{test}}$ and $I^{\text{est}}$. We have motivated their usefulness for testing presence or estimating location of a disease locus in linkage analysis. Both criteria have natural interpretations in terms of effective number of meioses for testing and estimation, thereby naturally generalizing the direct method based on fully informative meiotic events. Particular attention has been paid to complete and weak penetrance models.

An alternative performance criterion for testing is expected lod score. The classical lod score corresponds to a score function

$$S(\boldsymbol{w}) = \log_{10}(P_1(\boldsymbol{w})/P_0(\boldsymbol{w})).$$

In our framework of perfect marker data and conditioning on phenotypes, the conditional expected lod score at locus $t$ is defined as CELOD$(t) = E_1(S(\boldsymbol{v}(t)))$. It's maximum is attained at $\tau = t$ (see Hössjer, 2003a), which is the maximum conditional lod score

$$\text{MCELOD} = \text{CELOD}(\tau) = \sum_{\boldsymbol{w}} \log_{10}(P_1(\boldsymbol{w})/P_0(\boldsymbol{w}))P_1(\boldsymbol{w}). \tag{38}$$

This criterion is essentially the Kullback-Leibler distance between $P_0$ and $P_1$ and is closely related to $\log_{10}(I^{\text{test}} + 1) = m^{\text{test}}/\log_2(10)$. Therefore, an alternative definition of effective number of meioses for testing is

$$\tilde{m}^{\text{test}} = \log_2(10) \cdot \text{MCELOD}. \tag{39}$$

This criterion is also added over families, and it is easy to see that $m^{\text{test}}$ and $\tilde{m}^{\text{test}}$ are equivalent for complete penetrance models and in the limit $\varepsilon \to 0$ for weak penetrance models they differ by a factor $\log_2(e)$. The reason why we have used NPL scores rather than lod scores is the formers closer relationship to $p$-values (Kurbasic and Hössjer, 2003).

It is interesting to note the relationship between $I^{\text{est}}$ and $I^{\text{test}}$ with Fisher information. Firstly, $I^{\text{est}}$ resembles a Fisher information with an $m$-dimensional 'score vector' $\{(P_1(w + 1_j) - P_1(w))/(P_1(w + 1_j) + P_1(w))\}_{j=1}^m$ at $w$. Secondly, the proportionality constant $I = \omega\Sigma^{\text{test}}\omega'$ in $I^{\text{test}} = I\varepsilon^4 + o(\varepsilon^4)$ can be interpreted as a Fisher information for estimating the 'genetic model strength parameter' $\epsilon = \varepsilon^2$ at $\epsilon = 0$, cf. Hössjer (2003c).

One consequence of this work is sampling of pedigrees. For complete penetrance models, $m^{\text{test}}$ and $m^{\text{est}}$ are roughly proportional to the pedigree size (measured in terms of number of nonfounders). Since $m^{\text{est}}$ and $m^{\text{test}}$ are added over families, it is equally informative to sample one large pedigree as sampling many small pedigrees of the same total size. The situation is different for weak penetrance models, then both $I^{\text{est}}$ and (to a good approximation) $I^{\text{test}}$ are added over families. Both these quantities grow quadratically in pedigree size (measured in terms of number of individuals with known phenotype). This suggests that one large pedigree is more informative than many small pedigrees of the same total size. However, the situation is not as simple as this since distant relationships are much less informative than close ones. Still, we can conclude from (37) that one nuclear family with $n_{\text{off}}$ sibs is much more informative than, for instance, three nuclear families with $n_{\text{off}}/3$ sibs and similar phenotypes.

It is striking that the maximal noncentrality parameter $\text{NCP} = \sqrt{2^{m^{\text{test}}} - 1}$ for a correctly specified model grows so rapidly with $m^{\text{test}}$, cf. (13) and (28). Figure 6 shows that less than five fully informative meioses are sufficient for obtaining highly significant linkage for a genomewide scan with affected sib pairs (Lander and Kruglyak, 1995, Ängquist and Hössjer, 2003). This explains the early success of positional cloning for diseases with a clear Mendelian inheritance pattern, i.e. strong penetrance (Haines and Paricak-Vance, 1998). Of course, in practice the effective number of meioses required will be larger because of unknown genetic model and imperfect marker data.

We have defined $m^{\text{test}}$ and $m^{\text{est}}$ conditionally on phenotypes $Y$ for one pedigree. By averaging this quantities over a population we get, for each pedigree, the average number of meioses for testing and estimation in the population. For instance, the

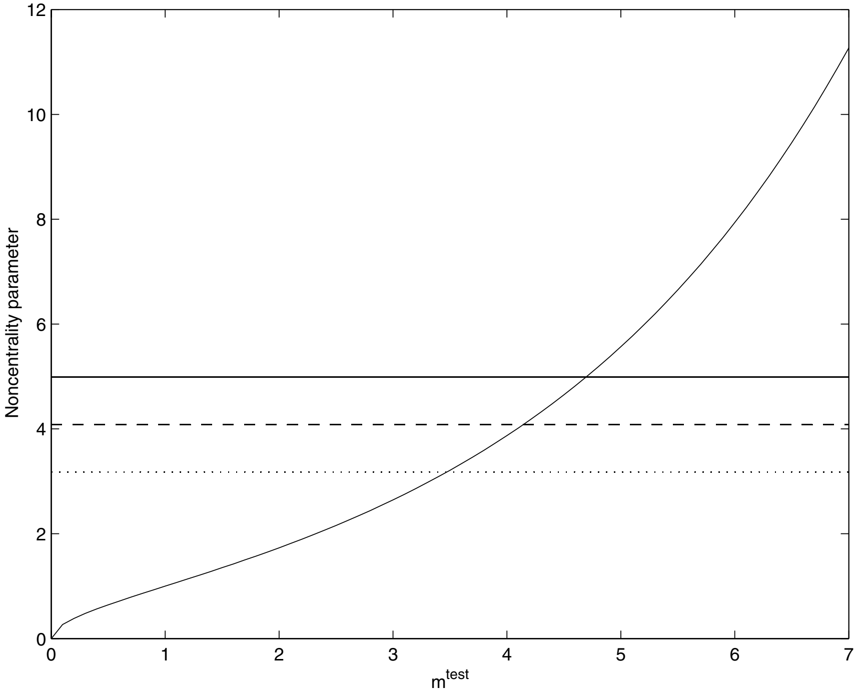

**Fig. 6.** Maximal noncentrality parameter ($\sqrt{2^{m^{\text{test}}} - 1}$) as function of effective number of meioses for testing ($m^{\text{test}}$). Horizontal lines are thresholds for genomewide significance for affected sib pairs (crossover rate 2) based on 23 pairs of chromosomes of total genetic length 33 M. They correspond to suggestive linkage (dotted), significant linkage (dashed) and highly significant linkage (solid).

maximum expected lod score (MELOD) is obtained by averaging MCELOD in (38) with respect to $Y$. In veiw of (39), $\log_2(10) \cdot$ MELOD can be interpreted as the average effective number of meioses for testing. See Hössjer (2003a, 2003e) for more work along this line. The conditional approach does not require specification of any sampling distribution of $Y$ (Winter, 1980). This is an advantage since the sampling distribution depends on the ascertainment scheme and is typically very involved.

The results in this paper are also valid for oligogenic diseases, provided that the other major genes are *unlinked* to $\tau$. It is only the probability $P(Y|v)$ in (3) that needs to be generalized. A challenge is to derive appropriate information bounds for several *linked* disease loci. Another generalization is to handle incomplete marker data.

## Appendix. Proofs

*Proof of Lemma 1.* To prove the first part, notice that

$$I^{\text{test}} = 2^m \sum_{\boldsymbol{w}} P_1(\boldsymbol{w})^2 - 1 = \prod_{i=1}^{N} 2^{m_i} \sum_{w_i} P_{1i}(w_i)^2 - 1 = \prod_{i=1}^{N} (I_i^{\text{test}} + 1) - 1.$$

The second part follows by noticing that

$$\frac{(P_1(\boldsymbol{w}') - P_1(\boldsymbol{w}))^2}{P_1(\boldsymbol{w}) + P_1(\boldsymbol{w}')} = \frac{(P_{1i}(w_i') - P_{1i}(w_i))^2}{P_{1i}(w_i) + P_{1i}(w_i')}$$

if $\boldsymbol{w} \sim \boldsymbol{w}'$ and differ at a bit which belongs to the $i^{\text{th}}$ pedigree. □

*Proof of Proposition 1.* Formula (14) follows from Hössjer (2003a), but to make the exposition self-contained we present a short proof here. First NCP $= \sum_{i=1} \gamma_i \text{NCP}_i$, so the constraint $\sum_{i=1}^N \gamma_i^2 = 1$ and Cauchy-Schwarz's inequality imply that $\gamma_i \propto \text{NCP}_i$ maximizes NCP *given a certain set of family-wise score functions* $\{S_i\}$. The maximal value is $\sqrt{\sum_{i=1}^N \text{NCP}_i^2}$. It remains to find score functions which maximize $\text{NCP}_i$ separately for each $i$. We introduce the inner product $(S_1, S_2) = 2^{-m_i} \sum_w S_1(w) S_2(w)$ for functions $\mathbb{Z}_2^{m_i} \to \mathbb{R}$. Then

$$\text{NCP}_i = E(Z_i(\tau)) = \frac{\sum_w S_i(w) P_{1i}(w)}{\sqrt{2^{-m_i} \sum_w S_i^2(w)}} = \frac{2^{m_i}(S_i, P_{1i} - P_{0i})}{\sqrt{(S_i, S_i)}}.$$

In the last equality we used the zero sum restriction $\sum_w S_i(w) = 0$ on $S_i$. When maximizing over all zero sum functions, it follows from Cauchy-Schwarz's inequality that $S_i \propto P_{1i} - P_{0i}$ gives the maximal value $\sqrt{I_i^{\text{test}}}$ of $\text{NCP}_i$.

The proof of (13), finally, is analogous to the proof of $\max_{S_i} \text{NCP}_i = \sqrt{I_i^{\text{test}}}$. □

*Proof of Proposition 2.* The slope-to-noise ratio for the total linkage score (12) is

$$\text{SLNR} = \frac{(\sum_{i=1}^N \gamma_i a_i)^2}{\sum_{i=1}^N \gamma_i^2 \sigma_i^2}.$$

*Given score functions* $\{S_i\}$, the optimal weighting scheme $\gamma_i \propto a_i/\sigma_i^2$ gives the maximal value SLNR $= \sum_{i=1}^N \text{SLNR}_i$. By maximizing this expression with respect to all $\{S_i\}$ we obtain the first identity of (17). See Hössjer (2003a) for more details. The first inequality of (17) follows immediately since on the left hand side we maximize SLNR over a smaller class of score functions (namely those that are linear combinations of family-wise scores). Hence it remains to prove the second inequality of (17).

Let $\mathcal{A}$ be the space of real-valued functions $(\boldsymbol{w}, \boldsymbol{w}') \to R(\boldsymbol{w}, \boldsymbol{w}')$ defined for all $m2^{m-1}$ (unordered) pairs $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{Z}_2^m$ which differ at precisely one bit. Introduce the inner product $\langle R_1, R_2 \rangle = \sum_{\boldsymbol{w} \sim \boldsymbol{w}'} R_1(\boldsymbol{w}, \boldsymbol{w}') R_2(\boldsymbol{w}, \boldsymbol{w}')(P_1(\boldsymbol{w}) + P_1(\boldsymbol{w}'))$ on $\mathcal{A}$. It can be shown that

$$\begin{aligned} a &= \sigma_0^{-1} \sum_{\boldsymbol{w} \sim \boldsymbol{w}'} (P_1(\boldsymbol{w}') - P_1(\boldsymbol{w}))(S(\boldsymbol{w}') - S(\boldsymbol{w})) = \sigma_0^{-1} \langle R_{\text{opt}}, R \rangle \\ \sigma^2 &= \sigma_0^{-2} \sum_{\boldsymbol{w} \sim \boldsymbol{w}'} (P_1(\boldsymbol{w}) + P_1(\boldsymbol{w}'))(S(\boldsymbol{w}') - S(\boldsymbol{w}))^2 = \sigma_0^{-2} \langle R, R \rangle, \end{aligned} \tag{A.1}$$

where $R(\boldsymbol{w}, \boldsymbol{w}') = S(\boldsymbol{w}') - S(\boldsymbol{w})$, $R_{\text{opt}}(\boldsymbol{w}, \boldsymbol{w}') = (P_1(\boldsymbol{w}') - P_1(\boldsymbol{w}))/(P_1(\boldsymbol{w}) + P_1(\boldsymbol{w}'))$ and $0/0 = 0$ in the definition of $R_{\text{opt}}$. Formula (A.1) is proved in Hössjer (2003a) for one pedigree and the proof for $N$ pedigrees is identical. Hence

$$\sup_S \text{SLNR} \le \sup_{R \in \mathcal{A}} \frac{\langle R_{\text{opt}}, R\rangle^2}{\langle R, R\rangle} = \langle R_{\text{opt}}, R_{\text{opt}}\rangle = I^{\text{est}}, \tag{A.2}$$

and the second supremum is attained for $R = R_{\text{opt}}$. This completes the proof.

Notice that the bound (A.2) need not be tight, since this requires existence of a score function $S$ such that $S(\boldsymbol{w}') - S(\boldsymbol{w}) = R_{\text{opt}}(\boldsymbol{w}, \boldsymbol{w}')$ for all $\boldsymbol{w} \sim \boldsymbol{w}'$. □

*Proof of Theorem 1.* Let $\mathcal{F} = \{1, \ldots, f\}$ be the set of founders and $\mathcal{F}_0 \subset \mathcal{F}$ the set of heterozygous founders. Further, let $\mathbb{A}$ be the set of founder allele vectors $a = (a_1, \ldots, a_{2f})$ which are compatible with $G$, i.e. such that $G_k = (a_{2k-1}a_{2k})$ or $(a_{2k}a_{2k-1})$ for each founder $k$. Clearly $\mathbb{A}$ has $2^{|\mathcal{F}_0|}$ elements, since switching $a_{2k-1} \leftrightarrow a_{2k}$ leaves $a$ within $\mathbb{A}$ for each $k \in \mathcal{F}_0$.

Since $Y = G$, (2) and (3) imply

$$P_1(w) \propto P(G|w) = \sum_{a \in \mathbb{A}} P(G|a, w)P(a) = \sum_{a \in \mathbb{A};\, (a,w) \to G} P(a) \propto n(w),$$

where $n(w) := |\{a \in \mathbb{A};\ (a, w) \to G\}|$. We used the fact that $P(G|a, w)$ is one if $(a, w) \to G$ and zero otherwise and that $P(a)$ has the same value for all $a \in \mathbb{A}$ under random mating (4). By assumption, we know that $n(w) > 0$ for at least one $w$. Hence it suffices to prove that i) $n(w) = n(w + c)$ for each $c \in \mathbb{C}$ and ii) if $w \notin v + \mathbb{C}$, then at least one of $n(w)$ and $n(v)$ is zero.

In order to prove i), notice that

$$\begin{aligned}
&(a, w) \to G \Leftrightarrow (a, w + 1_j) \to G, && \text{if } j \in \text{Off}_k \text{ and } k \in \mathcal{A},\\
&(a, w) \to G \Leftrightarrow (\pi_k a, w + 1_{\text{Off}_k}) \to G, && \text{if } k \in \mathcal{B} \cap \mathcal{F},\\
&(a, w) \to G \Leftrightarrow (a, w + 1_{\text{Par}_k \,\cup\, \text{Off}_k}) \to G, && \text{if } k \in \mathcal{B} \cap \mathcal{N},
\end{aligned} \tag{A.3}$$

where $\pi_k$, $k \in \mathcal{F}_0$, is the permutation of $(a_1, \ldots, a_{2f})$ that switches elements $2k - 1$ and $2k$ and $\mathcal{N}$ is the set of nonfounders. By repeated application of (A.3) for various $k$ and $j$ it follows that $n(\cdot)$ is constant on each coset $w + \mathbb{C}$, and this proves i).

To prove ii), assume, on the contrary, the existence of two inheritance vectors $v = (v_1, \ldots, v_m)$ and $w = (w_1, \ldots, w_m)$ such that $w \notin v + \mathbb{C}$ and $n(v), n(w) > 0$. Then there exist founder allele vectors $a$ and $a'$ such that $(a, v) \to G$ and $(a', w) \to G$. By repeated application of the middle equation of (A.3), it follows that $a' = a$ can be assumed without loss of generality. We also assume that meioses have been numbered so that the two meioses giving rise to each nonfounder $k$ have lower number than all meioses corresponding to the offspring of $k$. Let $j = j(v, w)$ be the smallest number in $\{1, \ldots, m\}$ for which $v_j \neq w_j$. Assume also that $v$ and $w$ have been chosen to give the maximal possible value of $j(v, w)$. Denote the parent and offspring corresponding to $j$ by $k$ and $l$. If $k$ is a homozygote we must have $j(v + 1_j, w) > j(v, w)$ and $(a, v + 1_j) \to G$. Since $v + 1_j \in v + \mathbb{C}$, we could then replace $v$ by $v + 1_j$ and this would contradict, by the way we numbered meioses, our choice of $v$ and $w$ to maximize $j(v, w)$. Hence $k$ must be a heterozygote. Since

$v_j \neq w_j$, $k$ passes on different alleles (say $a_1$ and $a_2$) to $l$ according to $v$ and $w$. Let $k'$ be the other parent of $l$. Since $(a, v) \rightarrow G$ and $(a, w) \rightarrow G$, $k'$ must also be a heterozygote $(a_1a_2)$ which passes on different alleles to $l$ according to $v$ and $w$. But then $v' = v + 1_{\mathrm{Par}_l \cup \mathrm{Off}_l} \in v + \mathbb{C}$ satisfies $(a, v') \rightarrow G$ and $j(v', w) > j(v, w)$. Again, this contradicts our choice of $v$ and $w$ to maximize $j(v, w)$. Hence ii) is proved. □

*Proof of Corollary 1.* By definition (2) of $P_1$ we have

$$I^{\mathrm{test}} = 2^m \sum_w P_1(w)^2 - 1 = 2^m |\mathbb{C}|^{-1} - 1 = 2^{m^{\mathrm{test}}} - 1,$$

where the first equality follows as in the proof of Lemma 1 and $|\mathbb{C}| = 2^{\dim(\mathbb{C})} = 2^{m - m^{\mathrm{test}}}$ is the number of elements of $\mathbb{C}$. To prove (21), notice first that $m^{\mathrm{est}}$ can be written as $m^{\mathrm{est}} = |\{w;\ w \sim w_0 \text{ and } w \notin w_0 + \mathbb{C}\}|$. Then observe that each term of (2) is nonzero iff one of $w$ and $w'$ lies in $w_0 + \mathbb{C}$. The nonzero value is $|\mathbb{C}|^{-1}$, and the number of such pairs $w \sim w'$ is $m^{\mathrm{est}}|\mathbb{C}|$. Hence $I^{\mathrm{est}} = m^{\mathrm{est}}|\mathbb{C}| \cdot |\mathbb{C}|^{-1} = m^{\mathrm{est}}$.
□

*Details from Example 4.* When founder phases are unknown we have $\mathcal{A} = \emptyset$ and $\mathcal{B} = \mathcal{F}$, with $\mathcal{F}$ defined in the proof of Theorem 1. Now $\mathrm{Par}_k = \emptyset$ for each founder $k$. Further, the sets $\mathrm{Off}_1, \ldots, \mathrm{Off}_f$ are disjoint. Hence $\dim(\mathbb{C}) = f$ and $m^{\mathrm{test}} = m - f$. If $w \sim 0$ then $w = 1_j$ for some $j = 1, \ldots, m$. In order to determine $m^{\mathrm{est}}$ we must count the number of $j$ such that $1_j \notin \mathbb{C}$. It is clear that $1_j \in \mathbb{C}$ iff $j \in \mathrm{Off}_k$ for some founder $k$ with $|\mathrm{Off}_k| = 1$. The number of such $j$ is $f'$ and hence $m^{\mathrm{est}} = m - f'$. When founder phases are known we have $\mathcal{A} = \mathcal{B} = \emptyset$, implying $\mathbb{C} = \{0\}$ and hence (23). □

*Details from Example 5.* We number individuals as 1 = father, 2 = mother, $3, \ldots, n_{\mathrm{het}} + 2$ for the heterozygous offspring and $n_{\mathrm{het}} + 3, \ldots, n$ for the homozygous offspring. There are $m = 2n_{\mathrm{off}}$ meioses. Those corresponding to the father's and mother's offspring are numbered as 1, ..., $n_{\mathrm{off}}$ and $n_{\mathrm{off}} + 1, \ldots, 2n_{\mathrm{off}}$ respectively. Notice that

$$\begin{aligned} \mathrm{Off}_1 &= \{1, \ldots, n_{\mathrm{off}}\}, \\ \mathrm{Off}_2 &= \{n_{\mathrm{off}} + 1, \ldots, 2n_{\mathrm{off}}\}, \\ \mathrm{Par}_k &= \{k - 2, k - 2 + n_{\mathrm{off}}\},\ k = 3, \ldots, n. \end{aligned} \tag{A.4}$$

Assume first that the father has known phase. Then $\mathcal{A} = \{2, 3 + n_{\mathrm{het}}, \ldots, n\}$ and $\mathcal{B} = \emptyset$ since all heterozygous offspring have known phase. Hence $\mathbb{C}$ is spanned by all $1_j$, $j \in \mathrm{Off}_2$. These vectors are clearly linearly independent and hence $\dim(\mathbb{C}) = |\mathrm{Off}_2| = n_{\mathrm{off}}$ and $m^{\mathrm{test}} = m - n_{\mathrm{off}} = n_{\mathrm{off}}$. Further, $1_j \notin \mathbb{C}$ iff $j \in \mathrm{Off}_1$, so that $m^{\mathrm{est}} = |\mathrm{Off}_1| = n_{\mathrm{off}}$.

When the father has unknown phase $\mathcal{A}$ remains the same but $\mathcal{B} = \{1\}$. Hence $\mathbb{C}$ is spanned by $1_{\mathrm{Off}_1}$ and all $1_j$, $j \in \mathrm{Off}_2$, so that $m^{\mathrm{test}} = 2n_{\mathrm{off}} - (n_{\mathrm{off}} + 1) = n_{\mathrm{off}} - 1$. When $n_{\mathrm{off}} = 1$ there is no vector $1_j$ outside $\mathbb{C}$ whereas if $n_{\mathrm{off}} > 1$ we have $1_j \notin \mathbb{C}$ iff $j \in \mathrm{Off}_1$. This gives the formula for $m^{\mathrm{est}}$ in (25). □

*Details from Example 6.* Assume first that the parents have known phase. Then $\mathcal{A} = \{3 + n_{\mathrm{het}}, \dots, n\}$ and $\mathcal{B} = \{3, \dots, 2 + n_{\mathrm{het}}\}$, so that $\mathbb{C}$ is spanned by the vectors $1_{\mathrm{Par}_k}$, $k = 3, \dots, 2 + n_{\mathrm{het}}$. These vectors are linearly independent since the corresponding sets $\mathrm{Par}_k$ are disjoint. Hence $\dim(\mathbb{C}) = n_{\mathrm{het}}$ and $m^{\mathrm{test}} = 2n_{\mathrm{off}} - n_{\mathrm{het}}$. It is clear that no vector $1_j$ lies in $\mathbb{C}$, so that $m^{\mathrm{est}} = 2n_{\mathrm{off}}$.

When the parents have unknown phase we have $\mathcal{B} = \{1, \dots, 2 + n_{\mathrm{het}}\}$, and $\mathbb{C}$ is spanned by $1_{\mathrm{Off}_1}$, $1_{\mathrm{Off}_2}$ and $1_{\mathrm{Par}_k}$, $k = 3, \dots, 2 + n_{\mathrm{het}}$. After some calculations it can be shown that these vectors are linearly independent when $n_{\mathrm{het}} < n_{\mathrm{off}}$, giving $\dim(\mathbb{C}) = n_{\mathrm{het}} + 2$ and $m^{\mathrm{test}} = 2n_{\mathrm{off}} - n_{\mathrm{het}} - 2$. When $n_{\mathrm{het}} = n_{\mathrm{off}}$, the constraint $1_{\mathrm{Off}_1} + 1_{\mathrm{Off}_2} + \sum_{k=3}^{n_{\mathrm{het}}+2} 1_{\mathrm{Par}_k} = 0$ reduces $\dim(\mathbb{C})$ to $n_{\mathrm{het}} + 1$ and hence $m^{\mathrm{test}} = 2n_{\mathrm{off}} - n_{\mathrm{het}} - 1$. When $n_{\mathrm{off}} = 1$, both of the vectors $1_1 = (1, 0) = 1_{\mathrm{Off}_1}$ and $1_2 = (0, 1) = 1_{\mathrm{Off}_2}$ lie in $\mathbb{C}$, so that $m^{\mathrm{est}} = 0$, When $n_{\mathrm{off}} > 1$, it can be shown after some calculations that $1_j \notin \mathbb{C}$ for all $j$ and this gives $m^{\mathrm{est}} = m = 2n_{\mathrm{off}}$. □

*Proof of Proposition 3.* By plugging (29) into (7) and (8) we get

$$I^{\mathrm{test}} = 2^{-m} \sum_{w} S_{\mathrm{opt}}^2(w) \cdot \varepsilon^{2\rho} + o(\varepsilon^{2\rho}) \tag{A.5}$$

and

$$I^{\mathrm{est}} = 2^{-m-1} \sum_{w \sim w'} (S_{\mathrm{opt}}(w') - S_{\mathrm{opt}}(w))^2 \cdot \varepsilon^{2\rho} + o(\varepsilon^{2\rho}) \tag{A.6}$$

as $\varepsilon \to 0$. The proposition then follows by inserting (30) into (A.5) and (A.6) and interchanging sums with respect to $w$ (or $w$, $w'$ with $w' \sim w$) and $k, l$. □

*Acknowledgements.* The author wishes to thank two referees for helpful comments on an earlier version of the paper.

## References

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

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

Donnelly, P.: The probability that some related individuals share some section of the genome identical by descent. Theoret. Population Biol. **23**, 34–64 (1983)

Dudoit, S., Speed, T.P.: A score test for linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs. Biostatist. **1**, 1–26 (2000)

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

Fisher, R.: The correlation between relatives on the supposition of Mendelian inheritance. Proc. Roy. Soc. Edinburgh. **52**, 399–433 (1918)

Haines, J.L., Paricak-Vance, M.A.: Approaches to Gene Mapping in Complex Human Diseases. New York: Wiley-Liss, (1998)

Hössjer, O.: Determining inheritance distributions via stochastic penetrances. J. Amer. Statist. Assoc. **98**, 1035–1051 (2003e)

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

Hössjer, O.: Assessing accuracy in linkage analysis by means of confidence regions. Genet. Epidemiol. **25**, 59–72 (2003b)

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

Hössjer, O.: Spectral decomposition of score functions in linkage analysis, 2003d. Report 2003:21, Mathematical Statistics, Stockholm University.

Kempthorne, O.: Genetic Statistics. New York: Wiley, (1955)

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

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., Lander, E.S.: Parametric and nonparametric linkage analysis: A unified multipoint approach. Am. J. Hum. Genet. **58**, 1347–1363 (1996)

Kurbasic, A., Hössjer, O.: Computing $p$-values in parametric linkage analysis, 2003. Report 2003:18, Mathematical Statistics, Stockholm University. To appear in Human Heredity

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

McPeek, M.S.: Optimal allele-sharing statistics for genetic mapping using affected relatives. Genet. Epid. **16**, 225–249 (1999)

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

Risch, N.: Linkage strategies for genetically complex traits I. Multilocus models. Am. J. Hum. Genet. **46**, 222–228 (1990a)

Risch, N.: Linkage strategies for genetically complex traits II. The power of affected relative pairs. Am. J. Hum. Genet. **46**, 229–241 (1990b)

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

Siegmund, D.: Boundary crossing probabilities and statistical applications. Ann. Statist. **14**, 361–404 (1986)

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

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

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