The Annals of Statistics
2003, Vol. 31, No. 4, 1075–1109

ASYMPTOTIC ESTIMATION THEORY OF MULTIPOINT LINKAGE ANALYSIS UNDER PERFECT MARKER INFORMATION[1]

BY OLA HÖSSJER

Stockholm University

We consider estimation of a disease susceptibility locus τ at a chromosome. With perfect marker data available, the estimator $\hat{\tau}_N$ of τ based on N pedigrees has a rate of convergence N^{-1} under mild regularity conditions. The limiting distribution is the arg max of a certain compound Poisson process. Our approach is conditional on observed phenotypes, and therefore treats parametric and nonparametric linkage, as well as quantitative trait loci methods within a unified framework. A constant appearing in the asymptotics, the so-called asymptotic slope-to-noise ratio, is introduced as a performance measure for a given genetic model, score function and weighting scheme. This enables us to define asymptotically optimal score functions and weighting schemes. Interestingly, traditional $N^{-1/2}$ theory breaks down, in that, for instance, the *ML*-estimator is not asymptotically optimal. Further, the asymptotic estimation theory automatically takes uncertainty of τ into account, which is otherwise handled by means of multiple testing and Bonferroni-type corrections.

Other potential applications of our approach that we discuss are general sampling criteria for planning of linkage studies, appropriate grid size of marker maps, robustness w.r.t. choice of map function (dropping assumption of no interference) and quantification of information loss due to heterogeneity (with linked or unlinked trait loci).

We also discuss relations to pointwise performance criteria and pay special attention to weak genetic models, so-called local specificity models.

1. Introduction. Linkage analysis is concerned with localization of disease susceptibility genes. This is done by studying genetic linkage between observed quantities related to the disease and a number of marker genes, located at known positions along the chromosomes. Statistically, this entails carrying out a number of tests, which might give evidence that disease gene(s) are located at some chromosomal regions. Once such evidence is found, it is of interest to know the precision of the estimate of the disease locus.

For parametric methods based on likelihoods and a fixed number of markers, it is well known that the disease locus estimator follows standard asymptotics for *ML*-estimators [cf., e.g., Ott (1999), Chapter 5]. For instance, for a single marker, the recombination fraction estimator is asymptotically normal with

Received July 2001; revised May 2002.

[1]Supported in part by Swedish Natural Science Research Council Contract 6152-8013.

AMS 2000 subject classifications. Primary 62E20; secondary 62P10, 62M05, 92D10.

Key words and phrases. Arg max of stochastic processes, compound Poisson process, crossovers, linkage analysis, perfect marker information.

convergence rate $N^{-1/2}$, where N is the number of pedigrees. On the other hand, misspecification of the parametric model may result in inconsistent estimators, as noticed, for example, by Clerget-Darpoux, Bonaïti-Pellié and Hochez (1986).

In this article, we investigate the asymptotic behavior of disease locus estimators under perfect marker information, corresponding to a dense set of markers, when all (or sufficiently many) pedigree members are being typed. Under rather mild regularity conditions, the convergence rate is N^{-1}, with a nonstandard limiting distribution, which is the arg max of a compound Poisson process. This fast rate of convergence for confidence intervals has previously been noted in the genetics literature by several authors. Kong and Wright (1994) establish a special case of our result for backcross designs, and mention the possibility of generalizing this to other situations. Darvasi, Weinreb, Minke, Weller and Soller (1993) and Darvasi and Soller (1997) show by simulations that the lengths of confidence intervals are inversely proportional to the sample size for backcross and F_2 designs and Dupuis and Siegmund (1999) give theoretical justification of their results using an asymptotic expansion of the expected length of the confidence interval. Kruglyak and Lander (1995) give analytical expressions for the distribution function of confidence interval lengths for affected relative pairs in nonparametric linkage (NPL).

Our approach is conditional on observed phenotypes and treats parametric and nonparametric linkage, as well as quantitative trait loci (QTL) methods within a unified framework. Further, arbitrary pedigree structures are allowed for. The basic tool is arg max theory of stochastic processes and Markov properties of the inheritance vector process.

We argue that a certain *asymptotic slope-to-noise* ratio, analogous to the inverse of the asymptotic variance for $N^{1/2}$-consistent estimators, is an appropriate performance criterion for the whole sample in terms of estimation accuracy. This criterion enables us to derive asymptotically optimal weighting schemes and score functions, given a certain genetic model.

The paper is organized as follows: In Sections 2 and 3 we define concepts from linkage analysis, needed for the rest of the paper. Parametric, nonparametric and QTL score functions are introduced in Section 4. In Section 5, a general arg max result for stochastic processes is derived, which is then used in our linkage application in Section 6. Optimal weighting schemes and score functions are considered in Section 7, and relation to analogous pointwise criteria is described in Section 8. Particular attention is given to weak genetic models in Section 9. A local (efficacy related) version of the asymptotic slope-to-noise ratio is introduced, and locally optimal score functions and weighting schemes are derived. In Section 10 we discuss further consequences of our work, and finally, proofs and some technical regularity conditions are collected in the Appendix.

2. Some concepts from linkage analysis. Our objective is to locate that locus of a particular chromosome of length l that causes or contributes to a particular

inheritable disease. To this end we have family data consisting of N pedigrees $\mathcal{P}_1, \ldots, \mathcal{P}_N$, with $\mathcal{P}_i$ having n_i individuals. For some subset $\bar{\mathcal{P}}_i$ of $\mathcal{P}_i$, we have registered a vector $Y_i = (Y_{ik};\ k \in \bar{\mathcal{P}}_i)$ of *disease phenotypes*, where Y_{ik} measures some genetically influenced characteristic(s) of the kth individual of $\mathcal{P}_i$. In principle, the phenotypes can be both continuous (quantitative) or discrete (usually binary) random variables; in the former case, for example, blood pressure or body weight and in the latter case disease status. We may also associate a vector of covariates to each individual.

Suppose there is a disease-causing locus τ on some chromosome. The objective of linkage analysis is to test

$$\begin{aligned} &H_0 : \tau = \infty, \\ &H_1 : \tau \in [0, l], \end{aligned} \tag{2.1}$$

where $\tau = \infty$ means that the disease locus is located on another chromosome. For each locus $t \in [0, l]$ we define a (pointwise) test statistic $Z_N(t)$. It measures the degree of compatibility between the inheritance patterns observed at t (by studying the inheritance of a number of marker genes with known positions) and the disease phenotypes. Large positive values of $Z_N(t)$ give evidence that τ is located on the same chromosome as t in its close vicinity. An overall (nonlocal) test statistic for H_0 versus H_1 is

$$\sup_{0 \le t \le l} Z_N(t), \tag{2.2}$$

with

$$\hat{\tau}_N = \arg \max_{0 \le t \le l} Z_N(t) \tag{2.3}$$

the corresponding estimate of τ. Of course, $\hat{\tau}_N$ makes sense only under H_1. In this paper, we will focus on the properties of $\hat{\tau}_N$ as N grows.

In order to define $Z_N(t)$ we must first specify what is meant by an inheritance pattern of a pedigree. We assume that $\mathcal{P}_i$ has f_i *founders* (individuals without ancestors in the pedigree) and $n_i - f_i$ *nonfounders*. If each nonfounder k has both parents included in the pedigree, there are two meioses (production of ova and sperm cells) deciding which grandparental alleles k will receive at different loci on the chromosome. Thus the whole pedigree $\mathcal{P}_i$ contains $m_i = 2(n_i - f_i)$ meioses. The *inheritance vector* $v_i(t) = (v_{i1}(t), \ldots, v_{im_i}(t))$ of $\mathcal{P}_i$ at locus $0 \le t \le l$ is a binary vector of length m_i such that $v_{ij}(t)$ equals 0 or 1 depending on whether meiosis j transmits a grandpaternal or grandmaternal allele, $j = 1, \ldots, m_i$. Hence $v_i(t)$ specifies the mode of inheritance across $\mathcal{P}_i$ at locus t. The objective of linkage analysis is to test, for each locus t of interest, if $v_i(t)$ is independent of the observed phenotypes Y_i, $i = 1, \ldots, N$. A priori, without using information from the phenotypes, one has

$$P\big(v_i(t) = w\big) = 2^{-m_i} \tag{2.4}$$

for all $w \in \mathbb{Z}_2^{m_i}$, $0 \le t \le l$ and $i = 1, \dots, N$, where $\mathbb{Z}_2^{m_i}$ is the additive vector space over the field of two elements. This reflects the Mendelian mode of inheritance. Observation of Y_i gives the a posteriori distribution of $v_i(t)$,

$$P\big(v_i(t) = w | Y_i\big) = \frac{P(Y_i | v_i(t) = w)}{\sum_{w' \in \mathbb{Z}_2^{m_i}} P(Y_i | v_i(t) = w')}. \tag{2.5}$$

Let $S : \mathbb{Z}_2^{m_i} \to \mathbb{R}$ be a score function which to each inheritance vector $w \in \mathbb{Z}_2^{m_i}$ assigns a number $S(w) = S(w; \mathcal{P}_i, Y_i)$ which measures how compatible w is with the observed disease phenotypes Y_i. The score function $S(\cdot)$ depends on both the pedigree structure $\mathcal{P}_i$, the observed phenotypes Y_i and sometimes also on a set of known (i.e., beforehand estimated) parameters. Then

$$\bar{Z}_i(t) := S\big(v_i(t)\big) | Y_i \tag{2.6}$$

is the *family score* of the ith pedigree at locus t, defined conditionally on the observed phenotypes.

The test statistic $Z_N(\cdot)$ is defined according to

$$Z_N(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \gamma_i \bar{Z}_i(t), \qquad 0 \le t \le l, \tag{2.7}$$

where γ_i is the weight assigned to the ith pedigree and is chosen larger for more informative pedigrees in order to increase the accuracy of the estimate $\hat{\tau}_N$, or alternatively the power of the test based on (2.2). It is clear from (2.6) and (2.7) that the $\arg\max\ \hat{\tau}_N$ is not a single point but rather a finite union of bounded intervals. This turns up in the asymptotic behavior of $\hat{\tau}_N$ as $N \to \infty$ and will be discussed more extensively in Section 6.

Information about $v_i(t)$ is attained by typing members of $\mathcal{P}_i$ and observing the inheritance pattern of a number of marker genes located along the chromosome. Notice that (2.7) requires that $S(v_i(t))$ is known along the chromosome. This requires a dense set of genetic markers and that sufficiently many pedigree members ("sufficiently" depends on the pedigree structure) are genotyped. Even with perfect marker data, the phase of all founders is typically unknown; that is, it is not known which allele of a founder comes from the father and which from the mother. This means that we can never distinguish a fixed inheritance vector $w \in \mathbb{Z}_2^{m_i}$ from $w + u_{ik}$, $k = 1, \dots, f_i$, where the addition is componentwise modulus two and $u_{ik} \in \mathbb{Z}_2^{m_i}$ is one for all meioses originating from founder number k and zero otherwise. However, for most score functions of practical interest one has

$$S(w) = S(w + u_{ik}), \qquad k = 1, \dots, f_i, \tag{2.8}$$

for any $w \in \mathbb{Z}_2^{m_i}$ and $i = 1, \dots, N$. Thus, as long as (2.8) holds, the unknown phase of founders does not cause a problem.

3. The genetic model. In order to determine the distribution of the family score (2.6), we need to specify the conditional distribution $v_i(t)|Y_i$ given in (2.5). It depends on the *type* ϕ_i of $\mathcal{P}_i$, which includes the graphical structure of $\mathcal{P}_i$, the phenotype vector Y_i and genetic model parameters. Introduce $P_{\phi_i t}(w) = P(v_i(t) = w|Y_i)$, to emphasize that this conditional probability only depends on ϕ_i, t and w. Under H_0, $P_{\phi_i t}(\cdot)$ has a uniform distribution (2.4) over $\mathbb{Z}_2^{m_i}$, since $v_i(t)$ is then independent of Y_i. Assuming that H_1 holds, we first derive $P_{\phi_i t}$ at the disease locus as follows: Let $G_{ik} = (G_{ik}^p, G_{ik}^m)$ be the genotype of the kth individual in $\mathcal{P}_i$ at locus τ. For ease of exposition, we restrict ourselves to biallelic disease loci, although the setup is equally valid in the multiallelic case. Thus we assume that the paternally and maternally transmitted alleles G_{ik}^p and G_{ik}^m both belong to $\{a, A\}$, where A is the disease-causing allele and a the normal allele. Let $\mathcal{F}_i \subset \mathcal{P}_i$ be the set of founders for $\mathcal{P}_i$ and $G_{i\mathcal{F}_i} = (G_{ik})_{k\in\mathcal{F}_i}$ their genotypes. If w is the inheritance vector of $\mathcal{P}_i$, we let $G_i = G_i(G_{i\mathcal{F}_i}, w)$ be that collection of genotypes in the pedigree which can be uniquely inferred from $G_{\mathcal{F}_i}$ and w. Then (2.5) implies

$$P_{\phi_i\tau}(w) \propto P\big(Y_i|v_i(\tau) = w\big) = \sum_{G_{i\mathcal{F}_i}} P(Y_i|G_i)P\big(G_{i\mathcal{F}_i}\big), \tag{3.1}$$

where the sum ranges over all founder genotype configurations.

The joint probability $P(G_{i\mathcal{F}_i})$ of the founder genotypes requires a population genetic model. For instance, under random mating, the genotypes of the founders are independent, that is,

$$P(G_{i\mathcal{F}_i}) = \prod_{k\in\mathcal{F}_i} P(G_{ik}). \tag{3.2}$$

Random mating also implies *Hardy–Weinberg equilibrium*, which means that each factor in (3.2) can be determined from the allele frequencies $p = P(A)$ and $q = P(a) = 1 - p$ of the disease and normal allele(s) according to $P(AA) = p^2$, $P(Aa) = 2pq$ and $P(aa) = q^2$.

The conditional distribution of phenotypes given genotypes can be described with various degrees of generality. In the simplest case, one assumes conditional independence of individual phenotypes given genotypes, meaning that the first factor of (3.1) can be written as

$$P(Y_i|G_i) = \prod_{k\in\tilde{\mathcal{P}}_i} P(Y_{ik}|G_{ik}), \tag{3.3}$$

where the product ranges over pedigree members with known phenotype. The factors $P(Y_{ik}|G_{ik})$ are referred to as *penetrances* and depend on the model being used. Formula (3.3) can be generalized to incorporate environmental effects and contributions from other loci (unlinked to τ) as well.

EXAMPLE 1 (Binary phenotypes). We let $Y_{ik} = 1$ for an affected individual and 0 for an unaffected one. Then the penetrances

$$P(Y_{ik}|G_{ik}) = \begin{cases} g_0^{Y_{ik}}(1-g_0)^{1-Y_{ik}}, & G_i = (aa), \\ g_1^{Y_{ik}}(1-g_1)^{1-Y_{ik}}, & G_i = (Aa), \\ g_2^{Y_{ik}}(1-g_2)^{1-Y_{ik}}, & G_i = (AA), \end{cases} \tag{3.4}$$

depend on three numbers g_0, g_1 and g_2, which denote the probabilities of observing $Y_{ik} = 1$ for an individual with zero, one or two disease alleles. The penetrance parameters of the genetic model are $\psi = (g_0, g_1, g_2)$.

EXAMPLE 2 (Gaussian phenotypes). We assume that $Y_{ik}|G_{ik} \in N(m_{|G_i|} + \sum_{j=1}^r \beta_j x_{ikj}, \sigma^2)$, where $|G_{ik}|$ is the number of disease alleles of G_{ik}, σ^2 is the residual (environmentally caused) variance, $x_{ik} = (x_{ik1}, \ldots, x_{ikr})$ is the set of covariates of individual k in $\mathcal{P}_i$ and $(\beta_1, \ldots, \beta_r)$ are regression coefficients. Then the penetrance factor is not a probability but a density,

$$P(Y_{ik}|G_{ik}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(Y_{ik} - m_{|G_{ik}|} - \sum_{j=1}^r \beta_j x_{ikj}\right)^2\right), \tag{3.5}$$

and the vector of penetrance parameters is $\psi = (m_0, m_1, m_2, \beta_1, \ldots, \beta_r, \sigma^2)$.

Let $P_{\phi_i t}$ be a 2^{m_i}-dimensional row vector containing all probabilities $P_{\phi_i t}(w)$, $w \in \mathbb{Z}_2^{m_i}$. By combining (3.1)–(3.3), we arrive at an expression for $P_{\phi_i \tau}$. In order to evaluate $P_{\phi_i t}$ at other loci along the chromosome, we will assume that $v_i(\cdot)$ evolves as two independent time homogeneous Markov processes on $\mathbb{Z}_2^{m_i}$ with the same intensity matrix in either direction from τ, and with $P_{\phi_i \tau}$ as initial distribution. Thus

$$P_{\phi_i t} = P_{\phi_i \tau} Q_{\phi_i |t-\tau|}, \tag{3.6}$$

where $Q_{\phi_i h} = \{Q_{\phi_i h}(w', w)\}_{w', w}$ is a $2^{m_i} \times 2^{m_i}$ transition matrix corresponding to lag h. The entries of $Q_{\phi_i h}$ are described by assuming that the components of $v_i(\cdot)$ evolve as independent Markov processes on $\{0, 1\}$, with jumps occuring according to a Poisson process with constant intensity λ (Haldane's model of no interference). Each jump of $v_{ij}(\cdot)$ corresponds to a *crossover* of the jth meiosis of $\mathcal{P}_i$ between grandpaternal and grandmaternal chromosomes. When the genetic map distance is measured in centiMorgans (meaning that on the average $0.01h$ crossovers occur between two loci at distance h), one has $\lambda = 0.01$. The recombination fraction between two loci at distance h cM from each other is the probability that a certain meiosis transmits alleles from different grandparents at the two loci. Under Haldane's model it is given by

$$\theta_h = P\big(v_{ij}(\tau \pm h) \neq v_{ij}(\tau)\big) = \tfrac{1}{2}\big(1 - \exp(-2\lambda h)\big). \tag{3.7}$$

Since the components of $v_i(\cdot)$ evolve independently, we obtain

$$P\big(v_i(t) = w | v_i(\tau) = w'\big) = \theta_{|t-\tau|}^{|w'-w|}(1-\theta_{|t-\tau|})^{m_i-|w'-w|} \tag{3.8}$$

as an expression for $Q_{\phi_i|t-\tau|}(w, w')$, where $|w'-w| = \sum_{j=1}^{m_i}|w'_j - w_j|$ is the Hamming distance between w' and w.

4. Examples of score functions.

4.1. *Parametric linkage.* Following Kruglyak, Daly, Reeve-Daly and Lander (1996), let us introduce

$$V_i(t) = \left\{ v_i(t) + \sum_{k=1}^{f_i} \alpha_k u_{ik};\ \alpha_1, \ldots, \alpha_{f_i} \in \{0,1\} \right\}, \tag{4.1}$$

the equivalence class of all 2^{f_i} inheritance vectors that can be formed by starting from $v_i(t)$ and then changing phase of any founders in $\mathcal{P}_i$. Under perfect marker information, the available data is Y_i and $V_i(\cdot) = \{V_i(t);\ 0 \le t \le l\}$ and the likelihood function is $L_i(t) = P(Y_i, V_i(\cdot)|\tau = t)$. Suppose we wish to test H_0 against the pointwise alternative $H_1(t): \tau = t$. It is shown in Hössjer (2001a) that the likelihood ratio of $\mathcal{P}_i$ can be written as

$$\frac{L_i(t)}{L_i(\infty)} = 2^{m_i} P_{\phi_i \tau}\big(v_i(t)\big). \tag{4.2}$$

The key ingredient in the proof is to verify that $V_i(\cdot)$, which is a function of the Markov process $v_i(\cdot)$, is itself a Markov process [cf. also Proposition 1 in Dudoit and Speed (1999)]. The lod score, defined as the base 10 logarithm of the likelihood ratio of the whole data set, thus equals

$$\log_{10} \frac{\prod_{i=1}^N L_i(t)}{\prod_{i=1}^N L_i(\infty)} = \sum_{i=1}^N \log_{10}\big(2^{m_i} P_{\phi_i \tau}\big(v_i(t)\big)\big). \tag{4.3}$$

Apart from a factor $N^{1/2}$, the lod score is identical to $Z_N(t)$ in (2.7), with uniform weights $\lambda_i \equiv 1$ and a score function

$$S_{\text{lod}}(w) = \log_{10}\big(2^m P_{\phi\tau}(w)\big) \tag{4.4}$$

for a pedigree $\mathcal{P}$ of type ϕ with m meioses. This connection between score functions and likelihood analysis was noted by Kruglyak, Daly, Reeve-Daly and Lander (1996) in the context of incomplete marker information.

4.2. *Nonparametric linkage.* The parametric score function (4.4) requires knowledge of the genetic model. For many complex diseases this is not known. For models with binary phenotypes, as in Example 1, *nonparametric linkage* uses score functions based on allele sharing among affected individuals, which neither requires knowledge of the disease allele frequency nor penetrance parameters. If the affected individuals in the pedigree share more alleles from the same founders identical by descent at locus t than what could be expected under H_0, this gives evidence for linkage. An example of such a score function was introduced by Whittemore and Halpern (1994). Given two pedigree members k_1 and k_2 of $\mathcal{P}_i$, let $\mathrm{IBD}_{ik_1k_2} = \mathrm{IBD}_{ik_1k_2}(w)$ denote the number of founder alleles shared identical by descent by k_1 and k_2 in $\mathcal{P}_i$ when the inheritance vector is w. With $\mathcal{P}_i^{\mathrm{aff}} \subset \mathcal{P}_i$ the set of affected pedigree members with known phenotype, we then define

$$S_{\mathrm{pairs}}(w) = S_{\mathrm{pairs}}(w; \mathcal{P}_i, Y_i) = \sum_{k_1<k_2\in\mathcal{P}_i^{\mathrm{aff}}} \mathrm{IBD}_{ik_1k_2} \tag{4.5}$$

as the total number of alleles shared pairwise IBD among affecteds.

4.3. *Quantitative trait loci.* Consider Example 2, where the phenotypes Y_{ik} are conditionally Gaussian, given the genotypes. It is common to use multivariate normal distributions and variance components theory in order to map quantitative trait loci (QTL) [cf., e.g., Almasy and Blangero (1998)]. An alternative procedure is proposed in Commenges (1994) and Hössjer (2001b), using a score function

$$S_{\mathrm{add}}(v) = \sum_{k_1<k_2\in\bar{\mathcal{P}}_i} r_{ik_1}r_{ik_2}\mathrm{IBD}_{ik_1k_2}, \tag{4.6}$$

with $\bar{\mathcal{P}}_i$ the set of pedigree members with known phenotype. Notice that S_{add} is a weighted version of S_{pairs}, with $r_{ik} = Y_{ik} - m - \sum_j \beta_j x_{ikj}$ the residual of individual k and

$$m = q^2m_0 + 2pqm_1 + p^2m_2, \tag{4.7}$$

the average genetic effect. The score function S_{add} requires no multivariate normality assumptions, and yet it is asymptotically equivalent to the variance components technique described above for local alternatives and additive models $m_1 = (m_0 + m_2)/2$ [cf. Hössjer (2001b) for details]. Notice that S_{add} requires estimation of m and $\{\beta_j\}_{j=1}^r$, but not of any variance components. Such estimation can be achieved using methods from segregation analysis. For a model without covariates only $m = E(Y_{ik})$ needs to be estimated, and this can easily be done from population data.

5. An asymptotic argmax result. We regard $Z_N(\cdot)$ as a random element of $D[0, l]$, the space of right continuous functions $[0, l] \to \mathbb{R}$ with left-hand limits, which we equip with the Skorohod topology and the associated Borel

sigma algebra. We will first develop a more general asymptotic arg max result (Theorem 1), which is then specialized to our linkage application (Theorem 2) in the next section. For simplicity we assume

$$\tau \in (0, l), \tag{5.1}$$

corresponding, in the linkage application, to the disease locus being located at an inner point of the chromosome. We tacitly assume in this section that $\hat{\tau}_N$ is uniquely defined with probability 1. Although this is not the case in the linkage application, we discuss in Section 6 how to circumvent this difficulty.

The rate at which $\hat{\tau}_N$ tends to τ is critically dependent on the local behavior of the mean value and (co)variance functions of $Z_N(\cdot)$ around τ. Essentially, we assume, as $t \to \tau$, that $N^{-1/2}(E(Z_N(\tau)) - E(Z_N(t))) = a|t - \tau|^\alpha + o(|t - \tau|^\alpha)$ and $\mathrm{Var}(Z_N(t) - Z_N(\tau)) = \sigma^2|t - \tau|^{2\beta} + o(|t - \tau|^{2\beta})$, for some constants $1/2 \le \beta < \alpha$, $\beta \le 1$ and $a, \sigma^2 > 0$. As we will see below, this implies a rate of convergence N^{-d} of $\hat{\tau}_N - \tau$ towards zero as $N \to \infty$, where

$$d = d(\alpha, \beta) = \frac{1}{2(\alpha - \beta)}.$$

The rationale for this can be seen by transforming $Z_N(\cdot) - Z_N(\tau)$ both on the horizontal and vertical scales according to

$$\tilde{Z}_N(s) = \kappa_2 N^{\beta d}\big(Z_N(\tau + \kappa_1 s N^{-d}) - Z_N(\tau)\big) \tag{5.2}$$

with $s \in \tilde{\mathbb{S}}_N := N^d([0, l] - \tau)/\kappa_1$, $\kappa_1 = (\sigma/a)^{1/(\alpha-\beta)}$ and $\kappa_2 = a^{\beta/(\alpha-\beta)} \times \sigma^{-\alpha/(\alpha-\beta)}$. Then notice that

$$\left(\frac{a}{\sigma}\right)^{2d} N^d(\hat{\tau}_N - \tau) = \arg\max_{s \in \tilde{\mathbb{S}}_N} \tilde{Z}_N(s). \tag{5.3}$$

We extrapolate $\tilde{Z}_N(\cdot)$ outside $\tilde{\mathbb{S}}_N$, so that it becomes a random element of $D(-\infty, \infty)$ (but otherwise arbitrarily), and endow it with the Skorohod topology. The transformation $Z_N \to \tilde{Z}_N$ is made in such a way that $E(\tilde{Z}_N(s)) \to -|s|^\alpha$ and $\mathrm{Var}(\tilde{Z}_N(s)) \to |s|^{2\beta}$ as $N \to \infty$. Thus, under appropriate regularity conditions,

$$\tilde{Z}_N(\cdot) \xrightarrow{\mathcal{L}} \tilde{Z}(s) := W(s) - |s|^\alpha, \tag{5.4}$$

where $W(\cdot) \in D(-\infty, \infty)$ satisfies $E(W(s)) = 0$ and $\mathrm{Var}(W(s)) = |s|^{2\beta}$. The standard Brownian motion $B(\cdot)$ is the most well-known example of such a limiting distribution $W(\cdot)$, and more generally fractional Brownian motion, which corresponds to $1/2 < \beta \le 1$. Non-Gaussian limiting processes $W(\cdot)$ are also possible; see the discussion at the end of this section.

We are now ready to formulate an asymptotic result for $\hat{\tau}_N$:

THEOREM 1 (Asymptotic arg max result, general case). *Suppose* (5.1) *holds and consider a sequence* $\{Z_N(\cdot)\}_N$ *of stochastic processes satisfying* (G1)–(G5) *in the Appendix. Then, asymptotically as* $N \to \infty$, *the* arg max *of* $Z_N(\cdot)$ *satisfies*

$$\left(\frac{a}{\sigma}\right)^{2d} N^d(\hat{\tau}_N - \tau) \xrightarrow{\mathcal{L}} \arg\max_{s\in\mathbb{R}} \big(W(s) - |s|^\alpha\big), \tag{5.5}$$

with W *as defined in* (G3).

The classical parametric convergence rate $N^{-1/2}$ corresponds to $d(2, 1) = 1/2$. It occurs for instance in regular *ML* or M-estimation with i.i.d. data. Other rates of convergence are treated by Kim and Pollard (1990) and Arcones (1994, 1998). Theorem 1 differs from the results of Kim and Pollard and Arcones in that we only consider a one-dimensional (real-valued) index set and use different types of regularity conditions, tailored for the linkage application in Section 6. Further, our regularity conditions automatically give consistency *and* the asymptotic distribution.

Notice that $\alpha = 1$ and $\beta = 1/2$ yield N^{-1}-convergence. This has been noted in the change point literature, with $W(\cdot)$ a certain partial sum process [cf. Siegmund (1986) and Dümbgen (1991)] and for isotone functional estimation at a point of discontinuity [cf. Anevski and Hössjer (2002b)]. In the latter case $W(\cdot)$ is either a certain discretized Brownian motion (regression) or a centered Poisson process (density estimation). In the present paper (Section 6), as well as in Kong and Wright (1994), $W(\cdot)$ is a centered compound Poisson process.

6. Asymptotics in linkage analysis. Let us now specialize Theorem 1 to our linkage application. The *type* of a pedigree $\mathcal{P}$ can be represented as

$$\phi = (\mathcal{P}, \bar{\mathcal{P}}, Y, \text{genetic model}), \tag{6.1}$$

where $\bar{\mathcal{P}} \subset \mathcal{P}$ consists of those pedigree members with known phenotypes and $Y = (Y_k,\ k \in \bar{\mathcal{P}})$. The genetic model can be represented with various degrees of complexity. In the simplest case we have

$$\text{genetic model} = \big(p, \psi, P_\psi(\cdot\,|(aa)), P_\psi(\cdot\,|(Aa)), P_\psi(\cdot\,|(AA))\big), \tag{6.2}$$

where p is the disease allele frequency, ψ is the vector of penetrance parameters and $\{P_\psi(\cdot|\text{genotype})\}$ describes the conditional distribution of the phenotype given all possible genotypes. More complex genetic models can be defined by allowing, for example, for multiallele disease loci and multilocus models.

Assume there are finitely many pedigree graphs, say K_1, possible, and let $\mathcal{X}$ be the *sample space*, that is, the set of permissible values of the phenotypes. For instance, $\mathcal{X} = \{0, 1\}$ and $\mathbb{R}$ in Examples 1 and 2, respectively. Then we define the *type space*,

$$\begin{aligned}\Phi = \big\{&\phi;\ \mathcal{P} \in \{1, \ldots, K_1\},\ \bar{\mathcal{P}} \in \{\mathcal{P}' \subset \mathcal{P};\ |\mathcal{P}'| \geq 2\},\\ &Y \in \mathcal{X}^{\bar{\mathcal{P}}}, \text{genetic model} \in \{1, \ldots, K_2\}\big\},\end{aligned} \tag{6.3}$$

where we have coded the K_1 and K_2 possible pedigree graphs and genetic models as positive integers.

We can allow for heterogeneity by letting $K_2 = 2$, with one of the two genetic models corresponding to a disease locus unlinked to τ. Another possibility is to let the sample space depend on the genetic model, thereby slightly generalizing (6.3), so that several $\mathcal{X}$ are allowed for. For instance, when the phenotypes of some families are discretized, we need to represent this with several genetic models in the sense of (6.2).

EXAMPLE 3 (Sib pairs). To make (6.1) more concrete, we consider a population with pedigrees $\mathcal{P} = \{1, 2, 3, 4\}$, consisting of parents 1, 2, and offspring 3, 4. Suppose only sib pairs have known phenotypes in the population and that only one genetic model is of concern. Then $\bar{\mathcal{P}} = \{3, 4\}$, $Y = (Y_3, Y_4)$, and the type vector can be simplified to $\phi = Y$. For a binary genetic model, as in Example 1, the type space is $\Phi = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, although, by symmetry, $(0, 1)$ and $(1, 0)$ are equivalent. Since affected sib pairs are by far most powerful for genetic linkage for most binary genetic models of interest, it is common to reduce the type space to $\Phi = \{(1, 1)\}$, meaning that all pedigrees in the population are of the same type. Thus the method of ascertainment of the pedigree has reduced the type space.

For the Gaussian genetic model of Example 2, the type space is $\Phi = \mathbb{R}^2$. Risch and Zhang (1995, 1996) found that extremely discordant or concordant sib pairs were most powerful for linkage, suggesting that one should reduce the type space to a subset of $\mathbb{R}^2$.

If we also include the possibility that one or two parents can have a known phenotype, the type space consists of binary sequences of length between 2 and 4 (discrete case) or it equals $\mathbb{R}^2 \cup \mathbb{R}^3 \cup \mathbb{R}^4$ (continuous case). By symmetry, we equate the two possible cases when one parent has known phenotype.

More generally, since only finitely many pedigree structures and genetic models are allowed, we assume that the type space

$$\Phi = \bigcup_{k=1}^{K} \Phi_k$$

is a finite union of disconnected sets of (possibly) varying dimension. Let d_k be the dimension Φ_k. Then Φ_k is a point if $d_k = 0$ or $\Phi_k = J_{k1} \times \cdots \times J_{kd_k}$ if $d_k \geq 1$, where each J_{ki} is a (finite, half-infinite or infinite) one-dimensional interval. We assume that $\mathcal{P}$, $\bar{\mathcal{P}}$ and the genetic model are fixed throughout each Φ_k. If $m(\phi)$ is the number of meioses of pedigree type ϕ, it thus follows that $m(\cdot)$ is constant over each Φ_k. For a continuous trait, the phenotype vector Y varies over Φ_k, so that $\Phi_k = \chi^{|\bar{\mathcal{P}}|}$, where χ is an interval. Hence $d_k = |\bar{\mathcal{P}}|$ is a number between 2 and $|\mathcal{P}|$. In the binary case, the phenotype vector $Y \in \{0, 1\}^{\bar{\mathcal{P}}}$ is kept fixed at each Φ_k so that $d_k \equiv 0$ and Φ becomes a finite set of points.

Now Φ is topologized by introducing the metric $d(\phi_1, \phi_2)$, which is set to $|\phi_2 - \phi_1|/(1 + |\phi_2 - \phi_1|)$ if ϕ_1 and ϕ_2 belong to the same component Φ_k, with $|\cdot|$ the Euclidean norm of $\mathbb{R}^{d_k}$. Otherwise, we put $d(\phi_1, \phi_2) = 1$.

Let ϕ_i be the type of $\mathcal{P}_i$ and

$$\nu_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\phi_i}$$

the empirical measure defined by the sample $\{\mathcal{P}_i\}_{i=1}^N$, where δ_ϕ is a point mass at ϕ.

(L1) $\nu_N \to_{\mathcal{L}} \nu$ as $N \to \infty$ for some Borel measure ν on Φ; that is, $\nu_N(\cdot \cap \Phi_k) \to_{\mathcal{L}} \nu(\cdot \cap \Phi_k)$ in terms of weak convergence of measures on $\mathbb{R}^{d_k}$, for each $1 \le k \le K$.

Notice that (L1) includes the case when $\{\phi_i\}$ are drawn randomly from ν, and the weak convergence in particular implies that $\nu_N(\Phi_k) \to \nu(\Phi_k)$ for each k.

We will use the pedigree type ϕ_i of $\mathcal{P}_i$ to specify weights in (2.7) according to

$$\gamma_i = \gamma(\phi_i), \tag{6.4}$$

for some weight function $\gamma : \Phi \to (-\infty, \infty)$ that should be large in absolute value for types corresponding to "informative pedigrees." The fact that negative weights are allowed will be commented on in Section 10.2. Our next assumption, corresponding to (G1), is:

(L2) For each $t \neq \tau$, $\int \gamma(\phi)\mu_t(\phi)\, d\nu(\phi) < \int \gamma(\phi)\mu_\tau(\phi)\, d\nu(\phi)$,

where $\mu_t(\phi_i) = E(\bar{Z}_i(t))$ is the mean value of the ith family score at $t \in [0, l]$.

We will now investigate how the mean and variance functions of each family score $\bar{Z}_i(\cdot)$ scale locally around τ. To facilitate this, we introduce some additional notation: The conditional distribution $P_{\phi t}$ in (3.6) and the score function $S = \{S(w);\ w \in \mathbb{Z}_2^m\}$ can both be interpreted as row vectors of dimension M, where $m = m(\phi)$ and $M = M(\phi) = 2^m$, or as elements of $\mathbb{R}^M$. To highlight that the vector S actually depends on ϕ, we sometimes write $S = S_\phi$. In Lemma 1 of the Appendix, it is proved that

$$\mu_t(\phi) = \mu_\tau(\phi) - a(\phi)|t - \tau| + o(|t - \tau|) \tag{6.5}$$

and

$$\operatorname{Var}\big(\bar{Z}_i(t) - \bar{Z}_i(\tau)\big) = \sigma^2(\phi)|\tau - t| + o(|t - \tau|) \tag{6.6}$$

as $t \to \tau$, where $a(\phi)$ and $\sigma^2(\phi)$ are constants. The "mean slope" $a(\phi)$ is defined by

$$a(\phi) = \lambda\left(m \sum_w S(w) P_{\phi\tau}(w) - \sum_{j=1}^{m} \sum_w S(w) P_{\phi\tau}(w + e_j)\right) = \lambda S b_\phi^T, \tag{6.7}$$

where $e_j \in \mathbb{Z}_2^m$ is a unit vector with one in the jth position and zeros elsewhere, ${b_\phi}^T$ is the transpose of

$$b_\phi = -P_{\phi\tau} A_\phi \tag{6.8}$$

and $A_\phi = \lambda^{-1} dQ_{\phi h}/dh|_{h=0}$ is λ^{-1} times the infinitesimal generator of the Markov process (3.8). Further, the local variance $\sigma^2(\phi)$ is given by

$$\begin{aligned} \sigma^2(\phi) &= \lambda \sum_{w \in \mathbb{Z}_2^m} P_{\phi\tau}(w) \sum_{j=1}^{m} \big(S(w+e_j) - S(w)\big)^2 \\ &= \lambda S B_\phi S^T, \end{aligned} \tag{6.9}$$

where B_ϕ is a symmetric $M \times M$-matrix defined by

$$B_\phi = \operatorname{diag}(P_{\phi\tau} A_\phi) - \operatorname{diag}(P_{\phi\tau}) A_\phi - A_\phi \operatorname{diag}(P_{\phi\tau}). \tag{6.10}$$

The elements of A_ϕ are found by combining (3.7) and (3.8) and differentiating w.r.t. h,

$$A_\phi(w, w') = \begin{cases} -m, & w' = w, \\ 1, & |w' - w| = 1, \\ 0, & |w' - w| > 1. \end{cases} \tag{6.11}$$

Hence the elements of b_ϕ are $b_\phi(w) = m P_{\phi\tau}(w) - \sum_{j=1}^m P_{\phi\tau}(w + e_j)$. Similarly, the elements of B_ϕ have the form

$$B_\phi(w', w) = \begin{cases} m P_{\phi\tau}(w) + \displaystyle\sum_{j=1}^{m} P_{\phi\tau}(w + e_j), & w' = w, \\ -\big(P_{\phi\tau}(w') + P_{\phi\tau}(w)\big), & |w' - w| = 1, \\ 0, & |w' - w| > 1. \end{cases}$$

Equations (6.5) and (6.6) are crucial for establishing the local scaling of the linkage score process $Z_N(\cdot)$ in (2.7). In fact, we will prove below that Theorem 1 can be applied with $\alpha = 1$ and $\beta = 0.5$, giving a surprisingly fast rate of convergence N^{-1}, since $d(1, 0.5) = 1$. Further, the constants a and σ^2 appearing in Theorem 1 are weighted averaged versions of the quantities $a(\phi)$ and $\sigma^2(\phi)$, defined by

$$a = \int \gamma(\phi) a(\phi)\, d\nu(\phi) \tag{6.12}$$

and

$$\sigma^2 = \int \gamma^2(\phi) \sigma^2(\phi)\, d\nu(\phi), \tag{6.13}$$

respectively.

We will need some additional regularity conditions. For this we introduce $\sigma^2_{H_0}(\phi) = \mathrm{Var}_0(S) = E_0[(S - E_0(S))^2]$ as the variance under H_0 for $S = S_\phi(v)$. Subscript 0 here means that expectation is taken when $v \in \mathbb{Z}_2^{m(\phi)}$ has a uniform inheritance distribution (2.4).

(L3) The constants a and σ^2 are both positive.
(L4) Let m be the value of $m(\cdot)$ on Φ_k. Then $\phi \to S_\phi(w)$ is continuous on Φ_k for each $w \in \mathbb{Z}_2^m$.
(L5) Let $Y \in \mathcal{X}$ be the phenotype of a certain individual. Then $P_\psi(Y|(aa))$, $P_\psi(Y|(Aa))$ and $P_\psi(Y|(AA))$ are all continuous functions of Y on $\mathcal{X}$, for each (of finitely many possible) penetrance vectors ψ.
(L6) The weight function $\gamma(\cdot)$ is continuous on Φ.
(L7) The expression $\sup_N \int_{\gamma^2(\phi)\sigma^2_{H_0}(\phi)>A} \gamma^2(\phi)\sigma^2_{H_0}(\phi)\, d\nu_N(\phi)$ tends to zero as $A \to \infty$.

Notice that (L4)–(L6) are automatically satisfied for finite sample spaces $\mathcal{X}$, and (L5) can easily be generalized to incorporate two-locus and multiallelic genetic models. Further, (L7) is a technical uniform integrability condition that permits ν_N to be replaced by ν in certain integrals. It is automatically satisfied if, for example, $\gamma^2\sigma^2_{H_0}$ is bounded, which is the case for finite Φ.

The "consistency conditions" (L2) and (L3) can sometimes be checked by means of the following result:

PROPOSITION 1 [A sufficient condition for (L2) and (L3)]. *Consider a fixed pedigree type ϕ. Suppose that the score function $w \to S_\phi(w)$ is a strictly increasing function of $P_{\phi\tau}(w)$ and that $P_{\phi\tau}(w) \neq 2^{-m}$ for at least some w. Then*

$$\mu_t(\phi) < \mu_\tau(\phi) \qquad \text{for all } t \neq \tau \tag{6.14}$$

and

$$a(\phi) > 0. \tag{6.15}$$

The inequalities are reversed if S_ϕ is a strictly decreasing function of $P_{\phi\tau}$. Finally, $\sigma^2(\phi) > 0$ follows if $S_\phi(w) \neq S_\phi(w')$ for some pair (w, w') with $|w' - w| = 1$ and $P_{\phi\tau}(w) > 0$.

The lod score function (4.4) obviously satisfies the monotonicity requirement of Proposition 1, as do $S_\phi = P_{\phi\tau}$ and $S_\phi = -P^{-1}_{\phi\tau}$. The latter two score functions will be shown to have pointwise optimality properties in Section 8.

Before formulating the asymptotic arg max result, we need to specify the limiting distribution $\tilde{Z}$ in (5.4). Let $\bar{m}_N = \int_\Phi m(\phi)\, d\nu_N(\phi)$ denote the average number of meioses of a family type picked at random from $\{\phi_i\}_{i=1}^N$. Then $d\bar{\nu}_N(\phi) = m(\phi)\, d\nu_N(\phi)/\bar{m}_N$ is a measure on Φ corresponding to a randomly picked meiosis from $\{\phi_i\}_{i=1}^N$. The union of all crossovers in $\{\phi_i\}_{i=1}^N$ evolves as

a Poisson process with intensity $N\bar{\lambda}_N$, where $\bar{\lambda}_N = \bar{m}_N\lambda$. The corresponding asymptotic quantities $\bar{m}$, $\bar{\nu}$ and $\bar{\lambda}$ are defined by replacing ν_N by its asymptotic limit ν. Next, we define a *type-crossover space* according to

$$\Xi = \bigcup_{k=1}^{K} \Phi_k \times \mathbb{Z}_2^{m(\Phi_k)} \times \mathbb{Z}_2^{m(\Phi_k)},$$

where $m(\Phi_k)$ is the constant value of $m(\cdot)$ on Φ_k. An element $\xi = (\phi, w, w') \in \Xi$ corresponds to a crossover of a pedigree of type ϕ, so that the inheritance vector is changed from w to w' (and consequently $|w - w'| = 1$). We equip Ξ with the metric $\tilde{d}(\xi_1, \xi_2) = \tilde{d}((\phi_1, w_1, w_1'), (\phi_2, w_2, w_2'))$, which is set to $d(\phi_1, \phi_2)$ if ϕ_1 and ϕ_2 belong to the same Φ_k and $(w_1, w_1') = (w_2, w_2')$. In all other cases, we put $\tilde{d}(\xi_1, \xi_2) = 1$. Define a measure $\tilde{\nu}_N$ on the Borel sigma algebra on Ξ generated by $\tilde{d}(\cdot, \cdot)$ according to

$$d\tilde{\nu}_N(\xi) = d\bar{\nu}_N(\phi) P_{\phi\tau}(w) \mathbb{1}_{\{|w'-w|=1\}}/m(\phi),$$

which is the probability distribution of crossovers for the given sample at τ. The corresponding asymptotic measure $\tilde{\nu}$ is defined by replacing $\bar{\nu}_N$ by $\bar{\nu}$ in the definition of $\tilde{\nu}_N$. Next, we define a random variable $X : \Xi \to \mathbb{R}$ according to

$$X(\xi) = \gamma(\phi)\big(S_\phi(w') - S_\phi(w)\big), \tag{6.16}$$

which measures the change [after weighting with $\gamma(\phi)$] of the score function corresponding to a crossover ξ. Thus, if $\xi \sim \tilde{\nu}$, it is easy to see from (6.7) and (6.9) that $X = X(\xi)$ satisfies $E(X|\phi) = -\gamma(\phi)a(\phi)/(\lambda m(\phi))$ and $E(X^2|\phi) = \gamma^2(\phi)\sigma^2(\phi)/(\lambda m(\phi))$. After averaging out ϕ, we find that

$$\begin{aligned} E(X) &= \int E(X|\phi)\, d\bar{\nu}(\phi) = -a/\bar{\lambda}, \\ E(X^2) &= \int E(X^2|\phi)\, d\bar{\nu}(\phi) = \sigma^2/\bar{\lambda}. \end{aligned} \tag{6.17}$$

Define next doubly infinite sequences $\boldsymbol{T} = \{T_j\}_{j=-\infty}^{-1} \cup \{T_j\}_{j=1}^{\infty}$, with $0 < T_1 < T_2 < \cdots$ and $0 > T_{-1} > T_{-2} > \cdots$, and $\boldsymbol{\xi} = \{\xi_j\}_{j=-\infty}^{-1} \cup \{\xi_j\}_{j=1}^{\infty}$, of time points and crossovers, respectively. Let $\Omega = \{\boldsymbol{\omega}\}$ be the space of all such pairs of sequences $\boldsymbol{\omega} = (\boldsymbol{T}, \boldsymbol{\xi})$. We define the metric

$$\begin{aligned} \rho(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2) &= \rho\big((\{T_{1j}\}, \{\xi_{1j}\}), (\{T_{2j}\}, \{\xi_{2j}\})\big) \\ &= \sum_{j\neq 0} 2^{-|j|} \frac{|T_{2j} - T_{1j}|}{1 + |T_{2j} - T_{1j}|} + \sum_{j\neq 0} 2^{-|j|} \frac{\tilde{d}(\xi_{2j}, \xi_{1j})}{1 + \tilde{d}(\xi_{2j}, \xi_{1j})} \end{aligned}$$

on Ω. A measure $\check{\nu}$ on the Borel sigma algebra of Ω can be specified as follows: Assume that $\{T_j\}_{-\infty}^{-1}$ and $\{T_j\}_1^{\infty}$ evolve as two independent Poisson processes with intensity $\bar{\lambda}$, and that $\{\xi_j\}_{-\infty}^{-1} \cup \{\xi_j\}_1^{\infty}$ is an i.i.d. sequence of crossovers with $\xi_j \sim \tilde{\nu}$.

Define a mapping $z:\Omega \to D(-\infty,\infty)$ according to

$$z(s)=z\big(s;(\boldsymbol{T},\boldsymbol{\xi})\big)=\begin{cases}\kappa_2 \displaystyle\sum_{T_j\in(0,\kappa_1 s]} X_j, & s\geq 0,\\ \kappa_2 \displaystyle\sum_{T_j\in(\kappa_1 s,0]} X_j, & s<0,\end{cases} \tag{6.18}$$

with $X_j=X(\xi_j)$, $\kappa_1=\sigma^2/a^2$ and $\kappa_2=a/\sigma^2$. The distribution of the limiting process $\tilde{Z}$, defined generally in (5.4), can now be expressed by the measure $\breve{\nu}\circ z^{-1}$ in the linkage application. This means, in view of (6.17), that

$$\begin{aligned} E\big(\tilde{Z}(s)\big)&=\kappa_1\kappa_2\bar{\lambda}E(X)|s|=-|s|,\\ \operatorname{Var}\big(\tilde{Z}(s)\big)&=\kappa_1\kappa_2^2\bar{\lambda}E(X^2)|s|=|s|.\end{aligned} \tag{6.19}$$

Thus, since $\tilde{Z}$ has independent increments on disjoint intervals, the centered process $W(s)=\tilde{Z}(s)+|s|$ has the same covariance function as a standard Brownian motion B.

Notice that the arg max of neither Z_N nor $\tilde{Z}$ is unique; they are both unions of finitely many intervals. In order to make the arg max unique, we define, for any $u\in[0,1]$, a new $\arg\max_u$-functional as follows: Let $z\in D(J)$, where J is any (finite or infinite) subinterval of $\mathbb{R}$. Then

$$\arg\max_u z=F_z^{-1}(u),$$

where F_z is the right-continuous inverse of the distribution function $F_z(x)=|(-\infty,x]\cap\arg\max z|/|\arg\max z|$, where $\arg\max z$ is the set of *all* maxima for z and $|\cdot|$ is the Lebesgue measure. To be precise, this definition requires that $\arg\max z$ be nonempty and $|\arg\max z|$ be finite, which holds with probability 1 for all processes of interest to us.

With $\hat{\tau}_N(u)=\arg\max_u Z_N$, we are now ready to formulate the following theorem.

THEOREM 2 (Asymptotic arg max result, linkage case). *Consider the genetic model described in Sections* 2 *and* 3 *with disease locus estimator* $\hat{\tau}_N$ *and test statistic process* $Z_N(\cdot)$ *as in* (2.3) *and* (2.7). *Then, if* (2.8), (L1) *and* (L4)–(L6) *hold, it follows that*

$$\tilde{Z}_N\xrightarrow{\mathcal{L}}\tilde{Z} \tag{6.20}$$

as $N\to\infty$, *where* $\tilde{Z}$ *is the two-sided compound Poisson process* $z(\cdot;(\boldsymbol{T},\boldsymbol{\xi}))$ *defined in* (6.18) *with* $(\boldsymbol{T},\boldsymbol{\xi})\sim\breve{\nu}$. *If also* (L2), (L3) *and* (L7) *hold, as well as* (L8) *in the Appendix, it follows that*

$$\frac{a^2}{\sigma^2}N\big(\hat{\tau}_N(u)-\tau\big)\xrightarrow{\mathcal{L}}\arg\max_u\tilde{Z} \tag{6.21}$$

for each $u\in[0,1]$, *with* a^2 *and* σ^2 *as defined in* (6.12) *and* (6.13), *respectively*.

Notice that the weak convergence (6.20) holds under both H_0 and H_1. In particular, it does not require the consistency conditions (L2) and (L3). Further, (6.21) implies that $\hat{\tau}_N$ converges to τ at rate N^{-1}. The limiting distribution is nonstandard, expressed as the arg max of the compound Poisson process $\tilde{Z}$. The constant

$$\text{ASLNR}(\nu) := \frac{a^2}{\sigma^2} = \frac{(\int \gamma(\phi)a(\phi)\,d\nu(\phi))^2}{\int \gamma^2(\phi)\sigma^2(\phi)\,d\nu(\phi)} \tag{6.22}$$

appearing in (6.21) will be referred to as the *asymptotic slope-to-noise ratio* of $\hat{\tau}_N$. We will use it as a performance criterion for the estimation procedure, since the right-hand side of (6.21) involves the arg max of a stochastic process whose two first moments are fixed. Still, the *distribution* of $\tilde{Z}$ depends on both the weighting scheme and the score function. Further discussion and motivation of ASLNR(ν) as a performance criterion can be found in Hössjer (2001a).

Further simplification of the limit distribution is possible in the limit ASLNR(ν) $\to 0$, corresponding to weak genetic models. Then $W(\cdot) \to_{\mathcal{L}} B(\cdot)$ on $D(-\infty, \infty)$. The arg max functional of $B(s) - |s|$ is unique with probability 1, and an explicit expression for its distribution is obtained by Siegmund (1986).

7. Optimal weights and score functions.

7.1. *Optimal weights.* Let us keep the score function fixed and vary the weights. This means that the functions $a(\cdot)$ and $\sigma^2(\cdot)$ in (6.7) and (6.9) are fixed, whereas $\gamma(\cdot)$ varies. By the Cauchy–Schwarz inequality, the weight function that maximizes ASLNR(ν) in (6.22) is given by

$$\gamma(\phi) \propto \frac{a(\phi)}{\sigma^2(\phi)}, \tag{7.1}$$

corresponding to an asymptotic signal-to-noise ratio

$$\text{ASLNR}(\nu) = \int \frac{a^2(\phi)}{\sigma^2(\phi)}\,d\nu(\phi) = \int \text{ASLNR}(\delta_\phi)\,d\nu(\phi), \tag{7.2}$$

which is a weighted average of the asymptotic slope-to-noise ratios of all pedigree types.

In practice we only know ν_N, not ν. However, the optimal weighting function (7.1) is independent of the weighting measure ν applied to the type space. In particular, (7.1) optimizes ASLNR(ν_N), and so (7.2) implies

$$\text{SLNR}_N \le \sum_{i=1}^{N} \text{ASLNR}(\delta_\phi), \tag{7.3}$$

where

$$\begin{aligned} \mathrm{SLNR}_N &= N\,\mathrm{ASLNR}(\nu_N) \\ &= \frac{(\sum_{i=1}^N \gamma(\phi_i)a(\phi_i))^2}{\sum_{i=1}^N \gamma^2(\phi_i)\sigma^2(\phi_i)}, \end{aligned} \tag{7.4}$$

is the slope-to-noise ratio of the observed sample. This quantity grows at a rate N as the number of pedigrees increases. The maximal value of SLNR_N is attained when there is equality in (7.3), which happens iff the optimal weight function (7.1) is used.

EXAMPLE 4 (Information loss due to heterogeneity). Suppose the population consists of two pedigree types ϕ_I and ϕ_{II}, with only ϕ_I linked to τ. Then $P_{\tau\phi_{II}}(\cdot)$ is uniform, and hence $a_{II} = a(\phi_{II}) = 0$. If $r_N = d\nu_N(\phi_I)$ is the proportion of linked families in $\{\mathcal{P}_i\}_{i=1}^N$, then

$$\begin{aligned} \mathrm{SLNR}_N &= N\frac{(\gamma_I a_I r_N)^2}{\gamma_I^2\sigma_I^2 r_N + \gamma_{II}^2\sigma_{II}^2(1-r_N)} \\ &= Nr_N\frac{a_I^2}{\sigma_I^2 + (\gamma_{II}/\gamma_I)^2\sigma_{II}^2(1-r_N)/r_N}, \end{aligned}$$

with $\gamma_I = \gamma(\phi_I)$ and so on. The optimal weighting function (7.1) reduces to $\gamma_{II} = 0$, that is, ignorance of the unlinked families. The corresponding optimal slope-to-noise ratio is $\mathrm{SLNR}_N^{\mathrm{opt}} = Nr_N a_I^2/\sigma_I^2$. In practice, one cannot distinguish the two subpopulations, and hence $\gamma_I = \gamma_{II}$. Therefore, the relative information loss due to heterogeneity can be quantified as

$$\frac{\mathrm{SLNR}_N}{\mathrm{SLNR}_N^{\mathrm{opt}}} = \frac{1}{1 + (\sigma_{II}/\sigma_I)^2(1-r_N)/r_N},$$

which is smaller, equal to or larger than r_N depending on whether σ_{II}^2/σ_I^2 is larger than, equal to or smaller than 1.

EXAMPLE 5 (Numerical values for nuclear families). Table 1 illustrates values of $\mathrm{ASLNR}(\delta_\phi)$ and the optimal weight function (7.1) for a number of nuclear families with binary phenotypes and a varying number of affected (unaffected) children when S_{pairs} [cf. (4.5)] is used as score function. Two dominant and two recessive models (with and without phenocopies) are included. As seen in the table, the presence of phenocopies reduces $\mathrm{ASLNR}(\delta_\phi)$ as well as the optimal weights. Further, $\mathrm{ASLNR}(\delta_\phi)$ varies more between families than do the optimal weights.

TABLE 1

Value of mean slope $a = a(\phi)$, local variance $\sigma^2 = \sigma^2(\phi)$, asymptotic slope-to-noise ratio $\mathrm{ASLNR}(\delta_\phi)$ *and optimal weight $\gamma = a(\phi)/\sigma^2(\phi)$ for a number of nuclear families with two parents (having unknown phenotypes) and k children, of which l are affected and $k - l$ unaffected. Two dominant models* $[g_0 = 0$ *or* 0.1, $(g_1, g_2) = (0.8, 0.9)]$ *and two recessive ones* $[g_0 = 0$ *or* 0.1, $(g_1, g_2) = (0.1, 0.9)]$ *are considered. Numbers for the models without phenocopies* $(g_0 = 0)$ *are at the upper part of the table. The disease allele frequency p equals* 0.1, *the map distance is measured in Morgans* $(\lambda = 1)$ *and the score function is* S_{pairs}

		Dominant models				Recessive models			
k	l	a	σ^2	ASLNR	γ	a	σ^2	ASLNR	γ
2	2	1.95	8.00	0.48	0.24	3.00	8.00	1.13	0.37
3	2	2.35	8.00	0.69	0.29	3.09	8.00	1.19	0.39
3	3	2.73	10.23	0.73	0.27	4.81	11.93	1.94	0.40
4	2	2.56	8.00	0.82	0.32	3.10	8.00	1.20	0.39
4	3	3.63	10.96	1.20	0.33	5.30	12.33	2.28	0.43
4	4	2.89	11.33	1.74	0.26	5.70	14.57	2.22	0.39
5	2	2.68	8.00	0.89	0.34	3.08	8.00	1.18	0.38
5	3	4.18	11.41	1.53	0.37	5.26	12.51	2.44	0.42
5	4	4.32	12.99	1.44	0.33	6.84	15.91	2.95	0.43
5	5	2.60	11.49	0.59	0.23	5.63	15.56	2.04	0.36
2	2	1.17	8.00	0.17	0.15	0.66	8.00	0.05	0.08
3	2	1.27	8.00	0.20	0.16	0.53	8.00	0.04	0.07
3	3	1.86	9.52	0.36	0.20	2.21	9.80	0.50	0.22
4	2	1.24	8.00	0.19	0.16	0.42	8.00	0.02	0.05
4	3	2.27	9.85	0.52	0.23	1.97	9.61	0.40	0.20
4	4	2.19	10.53	0.46	0.21	4.02	12.65	1.28	0.32
5	2	1.12	8.00	0.16	0.14	0.33	8.00	0.01	0.04
5	3	2.47	10.02	0.61	0.25	1.68	9.37	0.30	0.18
5	4	2.94	11.40	0.76	0.26	4.19	12.83	1.36	0.33
5	5	2.21	10.96	0.44	0.20	4.89	14.56	1.64	0.34

7.2. *Optimal score functions.* We will now investigate which score function maximizes $\mathrm{ASLNR}(\nu)$. Now $\mathrm{ASLNR}(\nu)$ can be written as

$$\mathrm{ASLNR}(\nu) = \frac{(\int \gamma(\phi)\sigma(\phi)\mathrm{ASLNR}(\delta_\phi)^{1/2}\, d\nu(\phi))^2}{\int \gamma^2(\phi)\sigma^2(\phi)\, d\nu(\phi)}, \tag{7.5}$$

and $\gamma(\phi)\sigma(\phi)$ can be viewed as an effective weight of pedigree type ϕ, which has been normalized for multiplicative scaling of the score function. Given $\gamma(\cdot)\sigma(\cdot)$, $\mathrm{ASLNR}(\nu)$ is optimized by separately optimizing $\mathrm{ASLNR}(\delta_\phi)$ *for each pedigree type* ϕ. Thus we may consider a fixed pedigree type.

When choosing score function for a pedigree of type ϕ, it will be convenient to restrict ourselves to the space $\mathbb{R}_0^M = \{b \in \mathbb{R}^M;\ \sum_{w \in \mathbb{Z}_2^m} b(w) = 0\}$, where $m = m(\phi)$ and $M = 2^m$. Observe that $S \in \mathbb{R}_0^M$ is equivalent to the average family

score being zero [cf. (8.1) below]. This is no essential restriction, since adding constants to the score function does not affect $\hat{\tau}_N$.

Let $b_\phi = \{b_\phi(w)\} \in \mathbb{R}_0^M$ and $B_\phi = \{B_\phi(w', w)\}$ be as in (6.8) and (6.10). Define the scalar product $\langle \cdot, \cdot \rangle_\phi : \mathbb{R}_0^M \times \mathbb{R}_0^M \to \mathbb{R}$ through

$$\langle S, b \rangle_\phi = S B_\phi b^T/4 = \sum_{w' \in \mathbb{Z}_2^m} \sum_{w \in \mathbb{Z}_2^m} S(w') B_\phi(w', w) b(w)/4. \tag{7.6}$$

This is possible, since by (6.9), the matrix B_ϕ is positive definite on $\mathbb{R}_0^M \times \mathbb{R}_0^M$. It follows from (6.7) and (6.9) that

$$\mathrm{ASLNR}(\delta_\phi) = 4\lambda \frac{\langle b_\phi B_\phi^{-1}, S \rangle_\phi^2}{\langle S, S \rangle_\phi}, \tag{7.7}$$

where B_ϕ^{-1} is the inverse of B_ϕ, viewed as an operator on $\mathbb{R}_0^M$. The optimal score function in $\mathbb{R}_0^M$ becomes

$$S_{\mathrm{opt}} \propto b_\phi B_\phi^{-1} = -P_{\phi\tau} A_\phi B_\phi^{-1}. \tag{7.8}$$

Thus we may define a "generalized Fisher information,"

$$I(\delta_\phi) := 4\lambda \frac{\langle b_\phi B_\phi^{-1}, b_\phi B_\phi^{-1} \rangle_\phi^2}{\langle b_\phi B_\phi^{-1}, b_\phi B_\phi^{-1} \rangle_\phi} = \lambda b_\phi B_\phi^{-1} b_\phi^T = \lambda P_{\phi\tau} A_\phi B_\phi^{-1} A_\phi P_{\phi\tau}^T \tag{7.9}$$

for pedigree type ϕ, which is the value of $\mathrm{ASLNR}(\delta_\phi)$ corresponding to the optimal score function (7.8). Notice that $\mathrm{ASLNR}(\delta_\phi)$ is unaffected if we replace the score function S with $S + \delta \mathbf{1}$ for any $\delta \in \mathbb{R}$ and $\mathbf{1} = (1, \ldots, 1)$. Thus it is no restriction to assume $S \in \mathbb{R}_0^M$. An asymptotic Fisher information

$$I(\nu) = \int I(\delta_\phi)\, d\nu(\phi)$$

is obtained by replacing $\mathrm{ASLNR}(\delta_\phi)$ by $I(\delta_\phi)$ in (7.2). It is the value of $\mathrm{ASLNR}(\nu)$ obtained when using an optimal score function (7.8) and optimal weights (7.1). It is easy to see that the optimal weight function is uniform ($\lambda(\phi) \propto 1$) when S_{opt} is used. For the given sample, we have

$$\mathrm{SLNR}_N \le I_N := N I(\nu_N) = \sum_{i=1}^N I(\delta_{\phi_i}), \tag{7.10}$$

with equality when S_{opt} and uniform weights are employed. The right-hand side of (7.10) quantifies the degree of information present in data. Of course, the genetic model must be known in order to reach this information upper bound. Since this is rarely the case for complex diseases, it is of interest to define asymptotic efficiencies. Let

$$\mathrm{eff}(\delta_\phi) := \frac{\mathrm{ASLNR}(\delta_\phi)}{I(\delta_\phi)}$$

denote the asymptotic efficiency of pedigree type ϕ. Asymptotic efficiency corresponding to ν is defined analogously and the efficiency of the given sample is

$$\text{eff}_N = \frac{\text{SLNR}_N}{I_N} = \sum_{i=1}^{N} r_i \,\text{eff}(\delta_{\phi_i}),$$

a weighted average of the individual pedigree efficiencies with weights $r_i = I(\delta_{\phi_i}) / \sum_{j=1}^{N} I(\delta_{\phi_j})$ proportional to the Fisher information.

8. Pointwise criteria. In this section, we discuss two pointwise criteria, and compare them with the asymptotic slope-to-noise ratio (7.4). Throughout this section, we assume that the family scores are centered, so that

$$E\big(\bar{Z}_i(t)|H_0\big) = 2^{-m_i} \sum_{w \in \mathbb{Z}_2^{m_i}} S(w) = 0. \tag{8.1}$$

The first criterion is the H_0-normalized *signal-to-noise ratio*,

$$\text{SNR}_N = \frac{E^2(Z_N(\tau)|H_1)}{\text{Var}(Z_N(t)|H_0)} = \frac{(\sum_{i=1}^{N} \gamma(\phi_i)\mu(\phi_i))^2}{\sum_{i=1}^{N} \gamma^2(\phi_i)\sigma^2_{H_0}(\phi_i)}, \tag{8.2}$$

where $\mu(\phi) = \mu_\tau(\phi)$ is the mean under H_1 at the disease locus for a pedigree of type ϕ and $\sigma^2_{H_0}(\phi) = \text{Var}_0(S)$ the variance under H_0. Comparing (8.2) with the definition of SLNR_N in (7.4), we have essentially replaced $a(\phi)$ and $\sigma^2(\phi)$ by $\mu(\phi)$ and $\sigma^2_{H_0}(\phi)$. In particular, this means that the optimal weight function (7.1) now becomes $\gamma(\phi) \propto \mu(\phi)/\sigma^2_{H_0}(\phi)$. In order to derive an optimal score function, a rewriting of the kind (7.5) reveals that it suffices to maximize

$$\text{SNR}(\delta_\phi) = \frac{\mu^2(\phi)}{\sigma^2_{H_0}(\phi)} = \frac{(S, 2^{m_i} P_{\phi\tau})^2_\phi}{(S,S)_\phi} = \frac{(S, 2^{m_i} P_{\phi\tau} - \mathbf{1})^2_\phi}{(S,S)_\phi} \tag{8.3}$$

for each pedigree type $\phi = \phi_i$, with $(S,b)_\phi = 2^{-m_i} \sum_w S(w)b(w)$ and $\mathbf{1} = (1, \ldots, 1)$ a row vector with ones only. In the last step of (8.3), we used (8.1). Thus, an optimal score function is $S \propto P_{\phi\tau} - 2^{-m_i}\mathbf{1}$. For estimation purposes, we may ignore the centering constant of the score function and simply put $S \propto P_{\phi\tau}$.

Our second pointwise criterion is

$$\overline{\text{SNR}}_N = \frac{(E^2(Z_N(\tau)|H_1)}{\text{Var}(Z_N(\tau)|H_1)} = \frac{(\sum_{i=1}^{N} \gamma(\phi_i)\mu(\phi_i))^2}{\sum_{i=1}^{N} \gamma^2(\phi_i)\sigma^2_{H_1}(\phi_i)}, \tag{8.4}$$

which differs from (8.2) in that we replaced $\sigma^2_{H_0}(\phi)$ by

$$\sigma^2_{H_1}(\phi) = \text{Var}\big(S\big(v_i(t)\big)|H_1\big) = \sum_{w \in \mathbb{Z}_2^{m_i}} S^2(w) P_{\phi\tau}(w) - \mu^2(\phi),$$

the variance of the family score under H_1 if $\phi = \phi_i$. Thus, in analogy with the SNR_N criterion, the optimal weight function is $\gamma(\phi) \propto \mu(\phi)/\sigma^2_{H_1}(\phi)$. In order to find the optimal score function, we must maximize $\overline{\mathrm{SNR}}(\delta_\phi) = \mu^2(\phi)/\sigma^2_{H_1}(\phi)$ for each $\phi = \phi_i$, which is equivalent to maximizing

$$\frac{\mu^2(\phi)}{\mu^2(\phi) + \sigma^2_{H_1}(\phi)} = \frac{[S, \mathbf{1}]^2_\phi}{[S, S]_\phi} = \frac{[S, \mathbf{1} - P^{-1}_{\phi\tau}/c]^2_\phi}{[S, S]_\phi}, \tag{8.5}$$

where $[S, b]_\phi = \sum_w S(w)b(w)P_{\phi\tau}(w)$, $P^{-1}_{\phi\tau}$ is a row vector obtained by pointwise inverting the elements of $P_{\phi\tau}$ and $c = 2^{-m_i}\sum_w P^{-1}_{\phi\tau}(w)$. The right-hand side of (8.5) follows from (8.1), which can be written as $[S, P^{-1}_{\phi\tau}]_\phi = 0$. Thus, it follows easily that $S \propto \mathbf{1} - P^{-1}_{\phi\tau}/c$ yields an optimal score function. In the estimation framework, this can be simplified to $S \propto -P^{-1}_{\phi\tau}$. Of course, this argument requires that $P_{\phi\tau}(w) \neq 0$ for all w, and in any case the optimal score function is very sensitive to small variations in $\{P_{\phi\tau}(w)\}$ when these are close to zero.

Both SNR_N and $\overline{\mathrm{SNR}}_N$ are related to the asymptotic pointwise power of the (idealized) test statistic $Z_N(\tau)$ [cf. Sham, Zhao and Curtis (1997) and Nilsson (1999)]. These authors also derive the optimal weighting schemes defined above.

9. Local specificity models. When mapping a disease susceptibility gene, a larger data set is needed if there is weak dependence between the phenotypes and inheritance vectors. The strength of this dependence is referred to as the specificity component of linkage analysis in Thompson (1997). In this section, we will consider local specificity models, corresponding to a weak genetic component at τ.

Following Whittemore (1996), it is possible to define a one-parameter family of genetic models $\{P_{\phi\tau\varepsilon}(\cdot)\}_{\varepsilon\geq 0}$, where

$$P_{\phi\tau\varepsilon}(w) := P\big(v_i(\tau) = w | Y_i, \varepsilon\big) = 2^{-m}\big(1 + \varepsilon^k S_{\mathrm{loc}}(w)/k!\big) + o(\varepsilon^k) \tag{9.1}$$

as $\varepsilon \to 0$, k is a fixed positive integer and $m = m(\phi)$ is the number of meioses corresponding to a pedigree of type $\phi = \phi_i$. The scalar parameter ε measures the strength of the genetic component. The smaller ε is, the less information about the inheritance vector is contained in Y_i. The score function $S_{\mathrm{loc}} \in \mathbb{R}^M$, $M = 2^m$, must satisfy $\sum_w S_{\mathrm{loc}}(w) = 0$ because of the constraint $\sum_w P_{\phi\tau\varepsilon}(w) = 1$. Notice that $S_{\mathrm{loc}}(w)$ may be interpreted as a likelihood score function $d^k \log P_{\phi\tau\varepsilon}(w)/d\varepsilon^k|_{\varepsilon=0}$, and hence it can be shown to be locally optimal as $\varepsilon \to 0$ in a pointwise sense when testing H_0 versus $H_1(t)$ for a fixed t [cf. Cox and Hinkley (1974)].

One way of achieving (9.1) is to assume that the disease allele frequency is kept fixed while the vector of penetrance parameters $\psi = \psi(\varepsilon)$ varies with ε.

For instance, for binary phenotypes, as in Example 1, one can vary the penetrance probabilites (g_0, g_1, g_2) with ε in such a way that the prevalence $K_p = P(Y_{ik} = 1)$ is kept fixed. It is shown in McPeek (1999) and Hössjer (2001b) that (9.1) holds, with $k = 1$ for inbred pedigrees (containing loops) and $k = 2$ for outbred pedigrees (with no loops). For instance, in the additive case ($g_1 = (g_0 + g_2)/2$),

$$S_{\text{loc}} \propto S_{\text{pairs}} - E_0(S_{\text{pairs}})$$

for outbred pedigrees with $\bar{\mathcal{P}}_i$ consisting of affecteds only, with S_{pairs} as defined in (4.5). As another illustration, consider the Gaussian model in Example 2 and assume for simplicity that there are no covariates. Then the vector of penetrance parameters is $\psi = (m_0, m_1, m_2, \sigma^2)$. Assume further that the residual variance σ^2 is fixed while (m_0, m_1, m_2) varies with ε, given the constraint that the overall population mean $m = E(Y_{ik})$ in (4.7) is kept fixed. For outbred pedigrees and additive models ($m_1 = (m_0 + m_2)/2$) it turns out that (9.1) holds with $k = 2$ and

$$S_{\text{loc}} \propto S_{\text{add}} - E_0(S_{\text{add}}),$$

with S_{add} as defined in (4.6).

Local expansions (9.1) can also be performed for rare disease models. Then the vector of penetrance parameters is kept fixed whereas the disease allele frequency $p = \varepsilon$ tends to zero. Examples of locally optimal score functions obtained in this case are S_{robdom} [McPeek (1999)] and its generalization to arbitrary phenotype models [Hössjer (2001b)].

Consider now a fixed pedigree of type ϕ. Define quantities $a_\varepsilon(\phi)$, $\sigma_\varepsilon^2(\phi)$, $B_{\phi\varepsilon}$ and $\langle \cdot, \cdot \rangle_{\phi\varepsilon}$ as in (6.7), (6.9), (6.10) and (7.6) for model $P_{\phi\tau\varepsilon}$. Then introduce [provided the kth derivative of the remainder term in (9.1) is negligible as $\varepsilon \to 0$]

$$\text{SLE}(\delta_\phi) = \frac{(d^k a_\varepsilon(\phi)/d\varepsilon^k|_{\varepsilon=0})^2}{\sigma_0^2(\phi)} = \lambda \frac{\langle S, S_{\text{loc}} \rangle_{\phi 0}^2}{\langle S, S \rangle_{\phi 0}}, \tag{9.2}$$

as the *slope efficacy* of the score function S for pedigree type ϕ, with k as in (9.1). In the second step of (9.2) we used the fact that $B_{\phi 0} = -2^{-m+1} A_\phi$, where $m = m(\phi)$, so that $\langle S, b \rangle_{\phi 0} = 2^{-m-1} S(-A_\phi) b^T$ and $d^k a_\varepsilon(\phi)/d\varepsilon^k = 2\lambda \langle S, S_{\text{loc}} \rangle_{\phi 0}$.

Let $\text{ASLNR}_\varepsilon(\delta_\phi)$ be the asymptotic slope-to-noise ratio (7.7) when the specificity parameter is ε. Naturally, $\text{ASLNR}_0(\delta_\phi) = 0$, since $\varepsilon = 0$ in (9.1) results in a uniform inheritance distribution and then estimation of τ makes no sense. If we Taylor expand $\text{ASLNR}_\varepsilon(\delta_\phi)$ w.r.t. ε we obtain

$$\text{ASLNR}_\varepsilon(\delta_\phi) = \frac{\text{SLE}(\delta_\phi)}{k!} \varepsilon^k + o(\varepsilon^k). \tag{9.3}$$

Hence the slope efficacy essentially determines $\text{ASLNR}_\varepsilon(\delta_\phi)$ for small ε. It is easy to see that $\text{SLE}(\delta_\phi)$ is maximized by taking $S = S_{\text{loc}}$ in (9.2). Thus S_{loc} is locally

optimal, even in our estimation framework. This can also be seen by replacing B_ϕ by $B_{\phi\varepsilon}$ in (7.8). Then the optimal score function S_{opt} satisfies

$$\begin{aligned} S_{\text{opt}} &\propto -(P_{\phi\tau\varepsilon} - 2^{-m}\mathbf{1})A_\phi B_{\phi\varepsilon}^{-1} \\ &\propto \big(S_{\text{loc}} + o(1)\big)A_\phi\big(A_\phi + o(1)\big)^{-1} = S_{\text{loc}} + o(1) \end{aligned}$$

as $\varepsilon \to 0$.

Local Fisher information and efficiencies are defined analogously as the corresponding "fixed ε"-quantities in Section 7.2. In particular, the local efficiency of the score function S for pedigree type ϕ becomes

$$\text{leff}(\delta_\phi) := \frac{\langle S, S_{\text{loc}}\rangle_{\phi 0}^2}{\langle S, S\rangle_{\phi 0}\langle S_{\text{loc}}, S_{\text{loc}}\rangle_{\phi 0}},$$

which can be interpreted as the square of the correlation coefficient between S and S_{loc} in the metric induced by $\langle\cdot,\cdot\rangle_{\phi 0}$.

Notice that the two pointwise criteria introduced in Section 8 become equivalent as $\varepsilon \to 0$, since in the limit the two variance normalizations in (8.2) and (8.4) are the same. The pointwise efficacy for pedigree type ϕ becomes

$$\mathcal{E}(\delta_\phi) = \frac{(d^k\mu_{\tau\varepsilon}(\phi)/d\varepsilon^k|_{\varepsilon=0})^2}{\sigma_{H_0}^2(\phi)} = \frac{(S, S_{\text{loc}})_\phi^2}{(S, S)_\phi}, \tag{9.4}$$

where $\mu_{\tau\varepsilon}(\phi)$ is the value of $\mu_\tau(\phi)$ for model $P_{\phi\tau\varepsilon}$. Compared to (9.2), we have thus essentially replaced the scalar product $\langle\cdot,\cdot\rangle_{\phi 0}$ by $(\cdot,\cdot)_\phi$. Let $\text{SNR}_\varepsilon(\delta_\phi)$ be the signal-to-noise ratio (8.3) when the specificity parameter is ε. Then a Taylor expansion (9.3) holds with $\text{SNR}_\varepsilon(\delta_\phi)$ and $\mathcal{E}(\delta_\phi)$ in place of $\text{ASLNR}_\varepsilon(\delta_\phi)$ and $\text{SLE}(\delta_\phi)$. Thus it follows, in view of (9.4), that S_{loc} is locally optimal also in terms of the pointwise SNR-criterion. This suggests that the information loss inherent in the pointwise SNR-criterion is less crucial for small ε.

10. Outlook.

10.1. *Incomplete marker information.* Theorem 2 has potential applications to planning how many markers are needed for estimating the disease locus. Suppose we wish to use an equally spaced grid of fully polymorphic markers with grid size δ. Then, with $A_N = \text{ASLNR}(\nu_N)$, a rough upper bound for δ is $\text{SLNR}_N^{-1} = (NA_N)^{-1}$, which is of the order N^{-1}, with the constant of proportionality depending on the informativeness of the pedigrees in the population. In order to find this constant, one may consider an asymptotic scenerio with $\delta = cN^{-1}$, as is done by Kong and Wright (1994) for backcrosses. Another robust approach is to use the generalized estimating equations method of Liang, Huang and Beaty (2000) and Liang, Chiu and Beaty (2001). Their method yields $\sqrt{N}$-consistent

estimators of τ when the population is a mixture of finitely many types and the marker genes are fully polymorphic (but their number does not grow with N). However, it is still seen as an open problem to find consistent estimators of τ for arbitrary pedigree structures and varying degree of heterozygosity of the markers.

10.2. *Sampling criteria.* Define CELOD(t) as the expected value of the lod score in (4.3), conditional on observed phenotypes. This quantity is discussed, for example, in Sections 5.10 and 9.7 of Ott (1999), as an important criterion for planning linkage studies when phenotypes have been observed. In our context of perfect marker information $\text{CELOD}(t) = \sum_{i=1}^{N} \mu_t(\phi_i)$, with $\mu_t(\phi)$ as defined below (L2), using the lod score function (4.4). It follows from Proposition 1 that $\mu(\phi) = \mu_\tau(\phi) > \mu_t(\phi)$ for any $t \neq \tau$, and thus (L2) holds for lod score analysis when the assumed parametric model is true, since $\gamma(\cdot) \equiv 1$. Hence, the maximal conditional expected lod score is given by $\text{MCELOD} = \text{CELOD}(\tau) = \sum_{i=1}^{N} \mu(\phi_i)$, motivating that $\mu(\phi)$ is a useful sampling criterion for detecting linkage. However, a pointwise sampling criterion more related to the power to detect linkage is

$$\text{SNR}(\delta_\phi) = \frac{(\mu(\phi) - \mu_{H_0}(\phi))^2}{\sigma^2_{H_0}(\phi)}. \tag{10.1}$$

This agrees with (8.3), apart from the fact that we no longer assume $\mu_{H_0}(\phi) = E_0(S_\phi) = 0$. In fact, Jensen's inequality implies that for the lod score function $\mu_{H_0}(\phi) = 2^{-m} \sum_w \log_{10}(2^m P_{\phi\tau}(w)) < 0$, unless $P_{\phi\tau}(w) \equiv 2^{-m}$.

An example of (10.1) was furnished by Liang, Huang and Beaty (2000) and Liang, Chiu and Beaty (2001). Extending work of Risch and Zhang (1995, 1996), they considered family scores conditionally on phenotypes for, for example, sib pairs. Using the number of alleles shared IBD by the sib pair as score function, they proposed $|\mu(\phi) - 1|$ as sampling criterion, both for quantitative and binary phenotypes. Here $\phi = (Y_3, Y_4)$ is the type of the pedigree, as described in Example 3. Since $\mu_{H_0}(\phi) = 1$ and $\sigma^2_{H_0}(\phi) = 1/2$, we obtain $|\mu(\phi) - 1| = \sqrt{\text{SNR}(\delta_\phi)/2}$, proving that $|\mu(\phi) - 1|$ is equivalent to $\text{SNR}(\delta_\phi)$ as sampling criterion.

In this paper, we have seen that $\text{ASLNR}(\delta_\phi) = a^2(\phi)/\sigma^2(\phi)$ is a natural performance criterion for a pedigree of type ϕ, in terms of accuracy to locate τ, rather than $\text{SNR}(\delta_\phi)$ which is tailored for (pointwise) ability to *detect* linkage. Naturally, the two criteria are related. For instance, for sib pairs we have $a(\phi) = 4\lambda(\mu(\phi) - 1)$ [cf. Proposition 1 in Liang, Chiu and Beaty (2001)]. Hence, in this case $|\mu(\phi) - 1|$ is equivalent to using $a^2(\phi)$ as sampling criterion. Therefore, $a^2(\phi)/\sigma^2(\phi)$ can be viewed as a normalized version of $a^2(\phi)$, where also the amount of local fluctuations of the family scores around τ is accounted for. Notice that negative values of $a(\phi)$ need not contradict (L3), even for a population with only one pedigree type. This is so since the weight functions $\lambda(\cdot)$ may take on

negative values. In fact, the optimal weight function (7.1) stipulates that $a(\phi)$ and $\lambda(\phi)$ should have the same sign.

It is an interesting research topic to analyze further which sampling designs can be derived using $\mathrm{SNR}(\delta_\phi)$ or $\mathrm{ASLNR}(\delta_\phi)$. Both criteria are valid for arbitrary pedigree structures and handle lod socres, NPL and QTL score functions within a unified framework.

10.3. *Allowing for interference.* In (3.7), we assumed Haldane's map function $h \to \theta_h$, corresponding to no chiasma interference. Zhao and Speed (1996) showed that any valid map function can arise by modeling the crossover process of a fixed meiosis as a thinning (with probability $1/2$) of an underlying chiasma process, the latter modeled as a stationary renewal process with interarrival distribution F. If $F(0) = 0$ we must have $\theta_0 = 0$ and $d\theta_h/dh|_{h=0} = \lambda$, which gives a local Markov property $P(v_{ij}(\tau + h) = 1 - w | v_{ij}(\tau) = w) = \lambda h + o(h)$, $w = 0, 1$. If further Y_{ij} and $v_{ij(-\tau)} = \{v_{ij}(t);\ t \neq \tau\}$ are conditionally independent given $v_{ij}(\tau)$, we conjecture that the asymptotics of Theorem 2 carry over, at the expense of more technical arguments. The reason is that the *rescaled* sequence of crossovers $\{T_{Nj}\}_{j \neq 0}$ (obtained from crossovers at $\tau + T_{Nj}N^{-1}$) asymptotically follows a Poisson process. Thus our asymptotic theory is robust w.r.t. choice of map function.

10.4. *Linked trait loci.* Suppose that the second locus in Example 4 is linked to the first, with position $\tau' \in [0, l]$. The mean value function $t \to \mu_t(\phi_{II})$ is then peaked at τ' rather than τ. The question is in what way this affects the asymptotic behavior of $\hat{\tau}_N$, the estimator of τ. Let $a_{II} = d\mu_t(\phi_{II})/dt|_{t=\tau}$ and $\sigma^2_{II} = \lim_{h \to 0} h^{-1}(\mathrm{Var}(\bar{Z}_i(\tau + h) - \bar{Z}(\tau)))$, for any $i \in II$. Then we conjecture that Theorem 2 can be extended to

$$(rN)\frac{a^2}{\sigma^2}\big(\hat{\tau}_N(u) - \tau\big) \xrightarrow{\mathcal{L}} \arg\max_u \tilde{Z},$$

where $r = \lim r_N$ is the limiting proportion of pedigrees from I, $a = \gamma_I a_I$, $\sigma^2 = \gamma_I^2\sigma_I^2 + (1-r)\gamma_{II}^2\sigma_{II}^2/r$ and $s \to \tilde{Z}(s)$ is a compound Poisson process with variance function $|s|$ and mean value function $-|s| + \varepsilon s$, where $\varepsilon = (1-r)\gamma_{II}a_{II}/(r\gamma_I a_I)$, provided $|\varepsilon| < 1$ is assumed. Thus $\hat{\tau}_N$ is still an N-consistent estimator of τ, though with a nonvanishing asymptotic bias.

APPENDIX: PROOFS

A.1. Regularity conditions of Theorem 1.

(G1) $\liminf_{N \to \infty} \inf_{0 \leq t \leq l;\, |t-\tau| \geq \delta} N^{-1/2}(E(Z_N(\tau)) - E(Z_N(t))) > 0$ for each $\delta > 0$.

(G2) $\lim_{N \to \infty} \limsup_{t \to \tau} N^{-1/2}|E(Z_N(\tau)) - E(Z_N(t) - a|t - \tau|^\alpha)|/|t - \tau|^\alpha = 0$, for some constants $\alpha, a > 0$.

(G3) There exists a process $W(\cdot)$ on $D(-\infty,\infty)$, with $E(W(s)) = 0$ and $\mathrm{Var}(W(s)) = |s|^{2\beta}$ for some constant $1/2 \le \beta \le 1$, such that

$$P\big(\arg\max\big(\tilde{Z}(\cdot)\big) \text{ is bounded}\big) = 1,$$

with $\tilde{Z}(\cdot)$ as defined in (5.2).

(G4) Let $\hat{s}_{NL}$ and $\hat{s}_L$ denote the arg max, restricted to $[-L, L]$, of $\tilde{Z}_N(\cdot)$ and $\tilde{Z}(\cdot)$, respectively. Then $\hat{s}_{NL} \to_{\mathcal{L}} \hat{s}_L$ as $N \to \infty$ for each $0 < L < \infty$.

(G5) Let $Z_N^0(t) = Z_N(t) - E(Z_N(t))$. Then, for each $\varepsilon > 0$,

$$\lim_{L\to\infty} \limsup_{N\to\infty} P\bigg(\sup_{s\in\mathbb{S}_N, |s|\ge L} N^{\beta d}\big(|Z_N^0(\tau + sN^{-d}) - Z_N^0(\tau)|/|s|^{\alpha}\big) \ge \varepsilon \bigg) = 0,$$

where $\mathbb{S}_N = N^d([0, l] - \tau)$.

Condition (G2) concerns the local scaling of the mean function of Z_N. The local scaling of the variance function is not stated explicitly, but is implicit in (G3). The N^d-consistency of $\hat{\tau}_N$ will follow from (G1), (G2) and (G5). Condition (G5) was used by Anevski and Hössjer (2002a) in the context of functional estimation under order restrictions.

PROOF OF THEOREM 1. Let $\hat{s}_N$ denote the arg max in (5.3) and $\hat{s}$ the arg max of $\tilde{Z}(\cdot)$ on $(-\infty,\infty)$. Our objective is to prove $\hat{s}_N \to_{\mathcal{L}} \hat{s}$ as $N \to \infty$. In view of (G4), it suffices to verify that

$$\lim_{L\to\infty} P(\hat{s}_L \ne \hat{s}) = 0 \tag{A.1}$$

and

$$\lim_{L\to\infty} \limsup_{N\to\infty} P(\hat{s}_{NL} \ne \hat{s}_N) = 0. \tag{A.2}$$

Formula (A.1) follows immediately from (G3). Further, (G1) and (G2) imply that $\eta_0 = -\limsup_{N\to\infty} \sup_{s\in\tilde{\mathbb{S}}_N\setminus\{0\}} E(\tilde{Z}_N(s))/|s|^{\alpha}$ satisfies $0 < \eta_0 < 1$. Combining this with (G5), we obtain

$$\lim_{L\to\infty} \limsup_{N\to\infty} P\big(\tilde{Z}_N(s) \le -\eta|s|^{\alpha} \text{ for all } s \in \tilde{\mathbb{S}}_N, |s| \ge L\big) = 0,$$

for any $0 < \eta < \eta_0$, and this implies (A.2). □

PROOF OF PROPOSITION 1. Notice first that $\sigma^2(\phi) > 0$ follows immediately from (6.9), since at least one term in this double sum is guaranteed to be positive. In order to prove (6.14) and (6.15), we will use the following algebraic result: Given two real-valued sequences $\{x_i\}_{i=1}^{2^m}$ and $\{y_i\}_{i=1}^{2^m}$, we have

$$\sum_{i=1}^{2^m} x_i y_i = \sum_{i=1}^{2^m} x_{(i)} y_{(j_i)} \le \sum_{i=1}^{2^m} x_{(i)} y_{(i)}, \tag{A.3}$$

where $\{x_{(i)}\}$ and $\{y_{(i)}\}$ are the order statistics of the respective sequences. Further, divide $\{1, \dots, 2^m\}$ into K disjoint groups $I_1, \dots, I_K$ such that $x_{(i)}$ is constant when $i \in I_k$. Then equality holds in (A.3) iff j_i and i belong to the same I_k for all i.

Rewrite $a(\phi)$ as $a(\phi) = \sum_{j=1}^m (\sum_w P_{\phi\tau}(w) S_\phi(w) - \sum_w P_{\phi\tau}(w) S_\phi(w + e_j))$. We establish (6.15) by indentifying $\{P_{\phi\tau}(w)\}_w$ and $\{S_\phi(w)\}_w$ with the x- and y-sequences in (A.3). Since $\{P_{\phi\tau}(w)\}_w$ is not constant $K \geq 2$. Suppose $a(\phi) = 0$. Then we must have equality in (A.3) "m times," meaning that $w \in I_k \Rightarrow w + e_j \in I_k$, $j = 1, \dots, m$. By induction this implies $I_k = \{1, \dots, 2^m\}$, contradicting the fact that $K \geq 2$. Formula (6.14) is proved using

$$\begin{aligned}
\mu_t(\phi) &= \sum_w S_\phi(w) P_{\phi t}(w) \\
&= \sum_w S_\phi(w) \sum_{w'} P_{\phi\tau}(w') Q_h(w', w) \\
&= \sum_{w''} \theta_h^{|w''|} (1 - \theta_h)^{m - |w''|} \sum_w S_\phi(w) P_{\phi\tau}(w + w'') \\
&< \mu_\tau(\phi) \sum_{w''} \theta_h^{|w''|} (1 - \theta_h)^{m - |w''|} = \mu_\tau(\phi),
\end{aligned}$$

where $w'' = w' - w$ and $h = |t - \tau|$. The strict inequality follows since $\sum_w S_\phi(w) P_{\phi\tau}(w + w'') < \mu_\tau(\phi)$ for at least some w'', arguing as for $a(\phi)$ above. □

LEMMA 1. *The functions*

$$a(\cdot), \sigma^2(\cdot) \text{ and } \sigma_{H_0}^2(\cdot) \text{ are continuous on } \Phi, \tag{A.4}$$

$$\{\mu_t(\cdot)\}_{0 \leq t \leq l} \text{ are equicontinuous on } \Phi, \tag{A.5}$$

and further,

$$\tilde{\nu}_N \xrightarrow{\mathcal{L}} \tilde{\nu} \qquad \text{as } N \to \infty. \tag{A.6}$$

Let $\phi = \phi_i$ *and* $m = m(\phi)$. *Define* $a_h(\phi) = h^{-1}(E(\bar{Z}_i(\tau) - \bar{Z}_i(\tau + h))$, $\sigma_h^2(\phi) = h^{-1} \operatorname{Var}(\bar{Z}_i(\tau + h) - \bar{Z}_i(t))$, $\bar{\mu}(\phi) = \sup_{0 \leq t \leq l} |\mu_\tau(\phi) - \mu_t(\phi)|$, $\Delta_{\phi h} = h^{-1}(Q_{\phi 0} - Q_{\phi h}) + \lambda A_\phi$ *and* $\|\Delta_{\phi h}\| = \max_{w,w'} |\Delta_{\phi h}(w, w')|$. [*The symbols* $a_h(\phi)$ *and* $\sigma_h^2(\phi)$ *are different from* $a_\varepsilon(\phi)$ *and* $\sigma_\varepsilon^2(\phi)$ *in Section* 9.] *Then, there exist positive constants* $C_0(m), C_1(m), \dots$ *such that*

$$\|\Delta_{\phi h}\| \leq C_0(m) h, \tag{A.7}$$

$$|a_h(\phi) - a(\phi)| \leq C_1(m) h \sigma_{H_0}(\phi), \tag{A.8}$$

$$|\sigma_h^2(\phi) - \sigma^2(\phi)| \leq C_2(m) h \sigma_{H_0}^2(\phi), \tag{A.9}$$

$$|a(\phi)| \leq C_3(m) \sigma_{H_0}(\phi), \tag{A.10}$$

$$\bar{\mu}(\phi) \le C_4(m)\sigma_{H_0}(\phi), \tag{A.11}$$

$$\sigma^2(\phi) \le C_5(m)\sigma^2_{H_0}(\phi). \tag{A.12}$$

PROOF. Without loss of generality, we may assume (throughout the proof) that the score function is centered so that $E_0(S) = 0$. We first prove the (equi)continuity of the functions in (A.4) and (A.5). Let m be the value of that (L5), (2.5) and (3.1) imply that $\phi \to P_{\phi\tau}(w)$ is continuous on Φ_k for each $w \in \mathbb{Z}_2^m$. Therefore $\phi \to B_\phi$ is continuous on Φ_k as well, and so (L4) gives continuity of $\sigma^2(\cdot)$, since $\sigma^2(\phi) = S_\phi B_\phi S_\phi^T$. The remaining functions in (A.4) are handled similarly.

Now (3.6), the fact that $Q_{\phi h}$ is constant on Φ_k and the just verified continuity of $\phi \to P_{\phi\tau}(w)$ imply that $\{\phi \to P_{\phi t}(w)\}_{0\le t\le l}$ is an equicontinuous family of functions on Φ_k for each $w \in \mathbb{Z}_2^m$. Since $\mu_t(\phi) = \sum_w S_\phi(w) P_{\phi t}(w)$, (A.5) follows.

In order to prove (A.6), notice first that (L1) and the fact that $m(\cdot)$ is constant on each Φ_k proves $\bar{\nu}_N \to_{\mathcal{L}} \bar{\nu}$. This, together with the continuity $\phi \to P_{\phi\tau}(w)$ proves (A.6).

If follows from (3.8) and the definition of A_ϕ that

$$\Delta_{\phi h}(w, w') = \begin{cases} h^{-1}\theta_h^{|w'-w|}(1-\theta_h)^{m-|w'-w|}, & |w'-w| \ge 2, \\ h^{-1}\big(\lambda h - \theta_h(1-\theta_h)^{m-1}\big), & |w'-w| = 1, \\ h^{-1}\big(1-(1-\theta_h)^m - m\lambda h\big), & w' = w. \end{cases}$$

Using the facts that $0 \le \theta_h \le 1/2$, $\theta_h \le \lambda h$, $|1 - e^{-x} - x| \le x^2/2$, $|(1-x)^{m-1} - 1| \le (m-1)x$ and $(1-x)^{m-1} - 1 + mx| \le m(m-1)x^2/2$ when $x > 0$, it follows after some elementary calculations that (A.7) holds with $C_0(m) = \lambda^2 m(m+1)/2$.

To verify (A.8), we notice that $a_h(\phi) = h^{-1}S(Q_{\phi 0} - Q_{\phi h})P_{\phi\tau}^T$. Together with (6.7), this implies that $a_h(\phi) - a(\phi) = S\Delta_{\phi h}P_{\phi\tau}^T$. Then, it follows from $\sum_w P_{\phi\tau}(w) = 1$, the Cauchy–Schwarz inequality and (A.7) that

$$\begin{aligned} |a_h(\phi) - a(\phi)| &\le \|\Delta_{\phi h}\| \sum_w |S(w)| \le 2^m \|\Delta_{\phi h}\| \sigma_{H_0}(\phi) \\ &\le 2^m C_0(m) h \sigma_{H_0}(\phi). \end{aligned}$$

In order to prove (A.9), put $\delta_i = \bar{Z}_i(\tau + h) - \bar{Z}_i(\tau)$ and write

$$\sigma_h^2(\phi) - \sigma^2(\phi) = \big(h^{-1}E(\delta_i^2) - \sigma^2(\phi)\big) - h^{-1}E^2(\delta_i) =: i - ii.$$

Use (6.9), (3.8), condition on $\bar{Z}_i(\tau)$ when evaluating $E(\delta_i^2)$ and $E(\delta_i)$ and use the Markov property of $v_i(\cdot)$ in the right direction from τ, to find that

$$\begin{aligned} i = {} & h^{-1}\big(\theta_h(1-\theta_h)^{m-1} - \lambda h\big) \sum_w P_{\phi\tau}(w) \sum_{w';|w'-w|=1} \big(S(w') - S(w)\big)^2 \\ & + h^{-1} \sum_{k=2}^m \theta_h^k (1-\theta_h)^{m-k} \sum_w P_{\phi\tau}(w) \sum_{w';|w'-w|=k} \big(S(w') - S(w)\big)^2. \end{aligned}$$

Use the facts that $|\theta_h(1-\theta_h)^{m-1} - \lambda h| \le m(\lambda h)^2$ and $\sum_{k=2}^m \theta_h^k (1-\theta_h)^{m-k} \le 2\theta_h^2 \le 2(\lambda h)^2$, since $\theta_h < 0.5$. This yields

$$|i| \le (m+2)\lambda^2 h \sum_w P_{\phi\tau}(w) \sum_{w'} \big(S(w') - S(w)\big)^2$$
$$\le 2(m+2)(2^m+1)2^m \lambda^2 h \sigma^2_{H_0}(\phi) := C_{21}(m) h \sigma^2_{H_0}(\phi),$$

where in the last step we used $(S(w') - S(w))^2 \le 2(S^2(w') + S^2(w))$ and $\sum_w P_{\phi\tau}(w) = 1$. A similar analysis proves $|ii| \le C_{22}(m) h \sigma^2_{H_0}(\phi)$, and thus (A.9) follows, with $C_2(m) = C_{21}(m) + C_{22}(m)$.

Next, (A.10) follows from (6.7) and the Cauchy–Schwarz formula, since $|a(\phi)| \le 2m\lambda \sum_w |S(w)| \le 2m\lambda 2^m \sigma_{H_0}(\phi)$. Formula (A.12) is proved similarly, using (6.9). To verify (A.11), notice that $\mu_\tau(\phi) - \mu_t(\phi) = S(Q_{\phi 0} - Q_{\phi|t-\tau|}) P^T_{\phi\tau}$. Therefore $|\mu_\tau(\phi) - \mu_t(\phi)| \le \sum_w |S(w)| \le 2^m \sigma_{H_0}(\phi)$, making use of $|Q_{\phi 0}(w,w') - Q_{\phi h}(w,w')| \le 1$. From this (A.11) follows. □

A.2. Additional regularity condition for Theorem 2. Let $F_X = \tilde{\mu} \circ X^{-1}$ be the distribution of $X(\xi)$ in (6.16). Decompose F_X into a discrete and continuous part according to $F_X = \sum_{k=1}^\infty \varepsilon_k \delta_{x_k} + (1-u)G$ where $\varepsilon_1 \ge \varepsilon_2 \ge \cdots$ are the sizes of the atoms of F_X, $u = \sum_{k=1}^\infty \varepsilon_k$ and G is continuous. Let $C_k = X^{-1}(x_k)$ and $C = \bigcup_{k=1}^\infty C_k$. Then impose the following:

(L8) As $N \to \infty$, $\tilde{\nu}_N(C) \to \tilde{\nu}(C)$ and $\tilde{\nu}_N(C_k) \to \tilde{\nu}(C_k)$ for each k.

Notice that (A.6) only implies $\limsup_N \tilde{\nu}_N(C_k) \le \tilde{\nu}(C_k)$, since each C_k is closed. The extra conditions in (L8) are needed to ensure that vertical ties for $\tilde{Z}_N(\cdot)$ and $\tilde{Z}(\cdot)$ occur asymptotically with the same probability.

PROOF OF THEOREM 2. We will prove (6.21) by applying Theorem 1 with $\alpha = 1$, $\beta = 1/2$ and a and σ^2 as in (6.12) and (6.13). For this we need to establish (G1)–(G5). Formula (6.20) will be proved in conjunction with (G4). Since $m(\cdot)$ is bounded, we let $C_k = \max_m C_k(m)$ denote the maximum of the constants $C_k(m)$ appearing in Lemma 1.

We first establish (G1). Write

$$N^{-1/2}\big(E\big(Z_N(\tau) - Z_N(t)\big)\big) = \int \gamma(\phi)\big(\mu_\tau(\phi) - \mu_t(\phi)\big)\, d\nu(\phi)$$
$$+ \int \gamma(\phi)\big(\mu_\tau(\phi) - \mu_t(\phi)\big)\, d(\nu_N - \nu)(\phi)$$
$$=: I + II.$$

The integrand of II is bounded by $|\gamma(\phi)|\bar{\mu}(\phi)$ in absolute value, uniformly in t. Notice further that $\int \gamma^2(\phi)\sigma^2_{H_0}(\phi)\, d\nu(\phi) < \infty$ because of (L7) and Theorem 5.3 in Billingsley (1968). Since the square of the integrand in II is bounded

by $C_4^2\gamma^2(\phi)\sigma_{H_0}^2(\phi)$ [cf. (A.11)], the equicontinuity formula (A.5), (L6) and Theorems 5.1 and 5.4 in Billingsley (1968) prove that $\lim_{N\to\infty}\sup_{0\le t\le l}|II| = 0$. Since $\mu_t(\phi) = -SQ_{\phi,|t-\tau|}P_{\phi\tau}$, it is easy to see that $\mu_t(\phi)$ is continuous in t. Now the integrands of I and II are the same, so the discussion above concerning upper bounds of the integrand implies that I is continuous in t because of dominated convergence. Since $I > 0$ for each fixed $t \neq \tau$ because of (L1), it follows that $\inf_{t;|t-\tau|\ge\delta} I = 0$ for each $\delta > 0$, and this completes the proof of (G1).

In order to prove (G2), put $a_{Nh} = h^{-1}N^{-1/2}E(Z_N(\tau) - Z_N(\tau+h))$. Then

$$a_{Nh} - a = \int \gamma(\phi)\big(a_h(\phi) - a(\phi)\big)\,d\nu_N(\phi) + \int \gamma(\phi)a(\phi)\,d(\nu_N - \nu)(\phi) =: III + IV.$$

Notice that the second power of the integrand in IV is less than $C_3^2\gamma^2(\phi)\sigma_{H_0}^2(\phi)$ because of (A.10). Then (A.4), (L6), (L7) and Theorems 5.1 and 5.4 of Billingsley (1968) imply $\lim_{N\to\infty} IV = 0$. Next, $|III| \le C_1 h\int|\gamma(\phi)|\sigma_{H_0}(\phi)\,d\nu_N(\phi)$ because of (A.8). Therefore $\lim_{N\to\infty}\limsup_{h\to 0}|III| = 0$ follows from (L7), and this establishes (G2). We next prove (G4) and (6.20). It follows from (2.7), (5.2) and (6.18), that

$$\tilde{Z}_N(s) = z(s;\boldsymbol{\omega}_N), \tag{A.13}$$

with $\boldsymbol{\omega}_N = (\boldsymbol{T}_N, \boldsymbol{\xi}_N) = (\{T_{Nj}\}_{j\neq 0}, \{\xi_{Nj}\}_{j\neq 0})$. Here $\tau + T_{Nj}N^{-1}$ is the jth $(-j$th) crossover for *any* meiosis taking place in $\mathcal{P}_1,\dots,\mathcal{P}_N$ to the right (left) of τ as $j > 0$ ($j < 0$). Formally and w.l.o.g., we assume that $\tau + T_{Nj}N^{-1}$ is defined also outside $[0,l]$. Further, $\xi_{Nj} = (\phi_{Nj}, w_{Nj}, w'_{Nj})$ means that the crossover at $\tau + T_{Nj}N^{-1}$ occurs for a pedigree of type ϕ_{Nj} so that the inheritance vector changes from w_{Nj} to w'_{Nj} (in the direction from τ).

Let $\arg\max_{uL} z$ denote the $\arg\max_u$-functional applied to z, when restricted to $\mathrm{supp}(z)\cap[-L,L]$. Define the mapping $\Gamma_{uL}:\Omega\to\mathbb{R}$ according to $\Gamma_{uL}(\boldsymbol{\omega}) = \arg\max_{uL} z(\cdot;\boldsymbol{\omega})$. Our objective is to prove that

$$\hat{s}_{NL}(u) \xrightarrow{\mathcal{L}} \hat{s}_L(u) \tag{A.14}$$

as $N\to\infty$, where $\hat{s}_{NL}(u) = \Gamma_{uL}(\boldsymbol{\omega}_N)$, $\hat{s}_L(u) = \Gamma_{uL}(\boldsymbol{\omega})$ and $\boldsymbol{\omega}\sim\check{\nu}$.

The rationale for (A.14) is that $\boldsymbol{\omega}_N$ is close to $\boldsymbol{\omega}$ in distribution. By means of a coupling argument, our first step in proving (A.14) is to replace $\boldsymbol{\omega}_N$ with another random element of Ω. Put $\tilde{\boldsymbol{\omega}}_N = (\boldsymbol{T}, \tilde{\boldsymbol{\xi}}_N)$, where the components of $\tilde{\boldsymbol{\xi}}_N = \{\tilde{\xi}_{Nj}\}_{j\neq 0} = \{(\phi_{Nj}, \tilde{w}_{Nj}, \tilde{w}'_{Nj})\}_{j\neq 0}$ are i.i.d. with marginal distribution $\tilde{\nu}_N$. We may now couple $\boldsymbol{\xi}_N$ with $\tilde{\boldsymbol{\xi}}_N$ (i.e., choose a version of $\boldsymbol{\xi}_N$ with its prescribed distribution) as follows: Let $\{i_j\}_{j\neq 0}$ be i.i.d. random variables with the uniform distribution on $\{1,\dots,N\}$, such that the crossover at $\tau + T_{Nj}N^{-1}$ occurs in $\mathcal{P}_{i_j}$, that is, $\phi_{Nj} = \phi_{i_j}$. If $j > 0$, we let k be the largest positive integer less than j such that $i_k = i_j$, provided such an integer exists. Then define, by induction w.r.t. $j > 0$,

$$(w_{Nj}, w'_{Nj}) = \begin{cases} (\tilde{w}_{Nj}, \tilde{w}'_{Nj}), & \text{if } i_b \neq i_j,\ b = 1,\dots,j-1,\\ (w'_{Nk}, w'_{Nj}), & \text{otherwise,}\end{cases}$$

where, in the latter case, w'_{Nj} is chosen uniformly among the $m(\phi_{Nj})$ inheritance vectors at Hamming distance 1 from w'_{Nk}. The definition of (w_{Nj}, w'_{Nj}) for $j < 0$ is analogous. It is clear that, for any positive integer K,

$$P\big((w_{Nj}, w'_{Nj}) = (\tilde{w}_{Nj}, \tilde{w}'_{Nj}),\ j = -K, \dots, -1, 1, \dots, K\big) \to 1 \tag{A.15}$$

as $N \to \infty$. Notice next that $\{T_{Nj}\}_{j=1}^{\infty}$ and $\{T_{Nj}\}_{j=-1}^{-\infty}$ evolve as two independent Poisson processes with intensity $\bar{\lambda}_N$. (We assume here that T_{Nj} is defined in this way when $\tau + T_{Nj}N^{-1} \notin [0, l]$.) Therefore, we can choose a version of $\boldsymbol{T}$ such that $\boldsymbol{T}_N = (\bar{\lambda}_N/\bar{\lambda})\boldsymbol{T}$, and hence

$$\Gamma_{uL}(\boldsymbol{\omega}_N) = (\bar{\lambda}_N/\bar{\lambda})\Gamma_{uL}\big((\boldsymbol{T}, \boldsymbol{\xi}_N)\big). \tag{A.16}$$

Now $\bar{\lambda}_N \to \bar{\lambda}$ follows from (L1), and hence (A.15) and (A.16) together imply that $P(|\Gamma_{uL}(\boldsymbol{\omega}_N) - \Gamma_{uL}(\tilde{\boldsymbol{\omega}}_N)| > \varepsilon) \to 0$ for any $\varepsilon > 0$ as $N \to \infty$. Thus (A.14) will follow if we prove

$$\Gamma_{uL}(\tilde{\boldsymbol{\omega}}_N) \xrightarrow{\mathcal{L}} \Gamma_{uL}(\boldsymbol{\omega}). \tag{A.17}$$

Fix $0 < \varepsilon < 1$. There exists a compact set $K \subset \Xi$ with $\tilde{\nu}(K) \geq 1 - \varepsilon$. In view of (L4), (L6) and (6.16), we can pick some $\delta > 0$ such that $\tilde{d}(\xi_1, \xi_2) \leq \delta$ and $\xi_1 \in K$ imply $|X(\xi_2) - X(\xi_1)| \leq \varepsilon$. Now $\sum_{k=1}^{\infty} |\tilde{\nu}_N(C_k) - \tilde{\nu}(C_k)| \leq \varepsilon$ for all N large enough, because of (L8). This and (A.6) imply that we may construct a coupling between two random variables $\tilde{\xi}_N \sim \tilde{\nu}_N$ and $\xi \sim \tilde{\nu}$ such that for all N large enough $P(\tilde{d}(\xi, \tilde{\xi}_N) \leq \delta) \geq 1 - \varepsilon$ and $P((\xi, \tilde{\xi}_N) \in \bigcup_{k=1}^{\infty}((C_k \times U_k) \cup (U_k \times C_k))) \leq \varepsilon$, where $U_k = \Xi \setminus C_k$. Thus, with probability at least $1 - 3\varepsilon$, it holds that $\tilde{\xi}_N$ and ξ both belong to either the same C_k or to $U = \Xi \setminus C$, where $C = \bigcup_1^{\infty} C_k$. In the former case $\tilde{\xi}_N = \xi$, and in the latter case $|X(\tilde{\xi}_N) - X(\tilde{\xi})| \leq \varepsilon$.

Next we couple the i.i.d. sequences $\boldsymbol{\xi}$ and $\tilde{\boldsymbol{\xi}}_N$ by coupling the individual components ξ_j and $\tilde{\xi}_{Nj}$ as described above. Let A_j be the event that a coupling exists between ξ_j and $\tilde{\xi}_{Nj}$. Then $\{A_j\}$ are independent with $P(A_j) \geq 1 - 3\varepsilon$.

Let $M \in \mathrm{Po}(2L/(\bar{\lambda}\kappa_1))$ denote the number of jumps T_j within $[-L, L]$, and $j_1, \dots, j_M$ the corresponding indices. It is clear that M is independent of $(\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{\xi}}_N)$. Let D_L denote the arg max of $z(s; \boldsymbol{\omega})$ restricted to $[-L, L]$, and define M' as the number of jumps T_j contained in $\bar{D}_L$, the convex hull of the closure of D_L, plus any of the two endpoints $-L$ and L that belong to $\bar{D}_L$. Put $\Lambda_L = 0$ if $D_L = [-L, L]$, and otherwise

$$\Lambda_L = \sup_{|s| \leq L} z(s; \boldsymbol{\omega}) - \sup_{s \in [-L, L] \setminus D_L} z(s; \boldsymbol{\omega}).$$

Condition now on M and assume that $\bigcap_{k=1}^{M} A_{j_k}$ holds. If $M' > 2$, we must have $M' \geq 4$, and then the sum of the jumps $X_j = X(\xi_j)$ corresponding to the $M' - 2$ inner points of $\bar{D}_L$ is zero. It is possible to show that with probability 1, all of

the corresponding $M' - 2$ crossovers ξ_j lie in C. Hence, by construction of the coupling, $\tilde{\xi}_{Nj} = \xi_j$ for all these $M' - 2$ crossovers. Now

$$P\left(\{\Lambda_L = 0\} \cap \{\Gamma_{uL}(\boldsymbol{\omega}) \neq \Gamma_{uL}(\tilde{\boldsymbol{\omega}}_N)\} \Big| M, \bigcap_{k=1}^{M} A_{j_k}\right) = 0. \tag{A.18}$$

This is clear if $M = 0$, since then $\Gamma_{uL}(\boldsymbol{\omega}) = \Gamma_{uL}(\tilde{\boldsymbol{\omega}}_N) = (2u - 1)L$. If $M > 0$, $\{\Lambda_L = 0\}$ implies that all M jumps X_{j_k} equal zero. This has probability zero if $F(\{0\}) = 0$. Otherwise, $C_k = \{0\}$ for some k, and then by the construction of the coupling $\Gamma_{uL}(\boldsymbol{\omega}) = \Gamma_{uL}(\tilde{\boldsymbol{\omega}}_N) = (2u - 1)L$. Now (A.18) and the construction of the coupling imply

$$\begin{aligned} &P\left(\Gamma_{uL}(\boldsymbol{\omega}) \neq \Gamma_{uL}(\tilde{\boldsymbol{\omega}}_N) \Big| M, \bigcap_{k=1}^{M} A_{j_k}\right) \\ &\quad \leq P\left(0 < \Lambda_L \leq \sum_{k=1}^{M} |X_{Nj_k} - X_{j_k}| \Big| M, \bigcap_{k=1}^{M} A_{j_k}\right) \\ &\quad \leq P\left(0 < \Lambda_L \leq M\varepsilon \Big| M, \bigcap_{k=1}^{M} A_{j_k}\right), \end{aligned}$$

where $X_{Nj} = X(\tilde{\xi}_{Nj})$, and thus $|X_{Nj} - X_j| \leq \varepsilon$ given A_j. This implies that $P(\Gamma_{uL}(\boldsymbol{\omega}) \neq \Gamma_{uL}(\tilde{\boldsymbol{\omega}}_N)|M) \leq (1 - P(\bigcap_1^M A_{j_k})) + P(0 < \Lambda_L \leq M\varepsilon|M)$. Let K be a fixed positive number. After averaging out M and using $P(A_{j_k}) \geq 1 - 3\varepsilon$, we find that

$$P\big(\Gamma_{uL}(\boldsymbol{\omega}) \neq \Gamma_{uL}(\tilde{\boldsymbol{\omega}}_N)\big) \leq 3E(M)\varepsilon + P(0 < \Lambda_L \leq K\varepsilon) + P(M > K).$$

This implies (A.17), since we may first choose K as large as we please, and then ε arbitrarily small.

In order to prove (6.20), notice that $\boldsymbol{\omega} \to z(\cdot; \boldsymbol{\omega})$ is a continuous functional from Ω to $D(-\infty, \infty)$. But $\boldsymbol{\omega}_N \to_{\mathcal{L}} \boldsymbol{\omega}$ follows from (A.15), $\boldsymbol{T}_N = (\bar{\lambda}_N/\bar{\lambda})\boldsymbol{T}$ and (A.18), and hence the continuous mapping theorem implies (6.20).

For the proof of (G5) we refer to Hössjer (2001a). The main idea is to write $\mathbb{S}_N \setminus (-L, L)$ as a union of intervals of polynomially increasing size, and the supremum is first taken over each such subinterval separately. The proof of (G3), finally, is very similar to that of (G5). □

Acknowledgment. The author thanks two anonymous referees for a number of valuable suggestions. In particular, they pointed out previous work in the genetics literature on N-consistent estimators and confidence regions.

REFERENCES

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

ANEVSKI, D. and HÖSSJER, O. (2002a). A general asymptotic scheme for inference under order restrictions. Unpublished.

ANEVSKI, D. and HÖSSJER, O. (2002b). Monotone regression and density function estimation at a point of discontinuity. *J. Nonparametr. Statist.* **14** 279–294.

ARCONES, M. A. (1994). Distributional convergence of M-estimators under unusual rates. *Statist. Probab. Lett.* **21** 271–280.

ARCONES, M. A. (1998). L_p-estimators as estimates of a parameter of location for a sharp-pointed symmetric density. *Scand. J. Statist.* **25** 693–715.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

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 weight pairwise correlation statistic. *Genetic Epidemiology* **11** 189–200.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

DARVASI, A. and SOLLER, M. (1997). A simple method to calculate resolving power and confidence interval of QTL map location. *Behavior Genetics* **27** 125–132.

DARVASI, A., WEINREB, A., MINKE, V., WELLER, J. I. and SOLLER, M. (1993). Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* **134** 943–951.

DUDOIT, S. and SPEED, T. P. (1999). A score test for linkage using identity by descent data from sibships. *Ann. Statist.* **27** 943–986.

DÜMBGEN, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *Ann. Statist.* **19** 1471–1495.

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

HÖSSJER, O. (2001a). Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. Technical Report 2001:16, Centre for Mathematics, Lund Univ.

HÖSSJER, O. (2001b). Determining inheritance distributions via stochastic penetrances. Technical Report 2001:17, Centre for Mathematics, Lund Univ.

KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191–219.

KONG, A. and WRIGHT, F. (1994). Asymptotic theory for gene mapping. *Proc. Natl. Acad. Sci. U.S.A.* **91** 9705–9709.

KRUGLYAK, L., DALY, M. J., REEVE-DALY, M. P. and LANDER, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American J. Human Genetics* **58** 1347–1363.

KRUGLYAK, L. and LANDER, E. S. (1995). High-resolution genetic mapping of complex traits. *American J. Human Genetics* **56** 1212–1223.

LIANG, K.-Y., CHIU, Y.-F. and BEATY, T. H. (2001). A robust identity-by-descent procedure using affected sib pairs: Multipoint mapping for complex diseases. *Human Heredity* **51** 64–78.

LIANG, K.-Y., HUANG, C.-Y. and BEATY, T. H. (2000). A unified sampling approach multipoint analysis of qualitative and quantitative traits in sib pairs. *American J. Human Genetics* **66** 1631–1641.

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

NILSSON, S. (1999). Two contributions to genetic linkage analysis. Licentiate thesis, Chalmers Univ. Technology, Gothenburg, Sweden.

OTT, J. (1999). *Analysis of Human Genetic Linkage*, 3rd ed. Johns Hopkins Univ. Press.

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

RISCH, N. and ZHANG, H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: Sampling considerations. *American J. Human Genetics* **58** 836–843.

SHAM, P., ZHAO, J. H. and CURTIS, D. (1997). Optimal weighting scheme for affected sib-pair analysis of sibship data. *Ann. Human Genetics* **61** 61–69.

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

THOMPSON, E. A. (1997). Conditional gene identity in affected individuals. In *Genetic Mapping of Disease Genes* (I.-H. Pawlowitzki, J. H. Edwards and E. A. Thompson, eds.) 137–146. Academic Press, San Diego.

WHITTEMORE, A. S. (1996). Genome scanning for linkage: An overview. *American J. Human Genetics* **59** 704–716.

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

ZHAO, H. and SPEED, T. P. (1996). On genetic map functions. *Genetics* **142** 1369–1377.

DEPARTMENT OF MATHEMATICS
STOCKHOLM UNIVERSITY
S-10691 STOCKHOLM
SWEDEN
E-MAIL: ola@math.su.se