A Unified Discussion on the Concept of Score Functions Used in the Context of Nonparametric Linkage Analysis

Lars Ängquist

Centre for Mathematical Sciences, Department of Mathematical Statistics, Lund University, Lund, Sweden.

Abstract: In this article we try to discuss nonparametric linkage (NPL) score functions within a broad and quite general framework. The main focus of the paper is the structure, derivation principles and interpretations of the score function entity itself. We define and discuss several families of one-locus score function definitions, i.e. the implicit, explicit and optimal ones. Some generalizations and comments to the two-locus, unconditional and conditional, cases are included as well.
Although this article mainly aims at serving as an overview, where the concept of score functions are put into a covering context, we generalize the noncentrality parameter (NCP) optimal score functions in Ängquist et al. (2007) to facilitate—through weighting—for incorporation of several plausible distinct genetic models. Since the genetic model itself most oftenly is to some extent unknown this facilitates weaker prior assumptions with respect to plausible true disease models without loosing the property of NCP-optimality.
Moreover, we discuss general assumptions and properties of score functions in the above sense. For instance, the concept of identical by descent (IBD) sharing structures and score function equivalence are discussed in some detail.

Keywords: nonparametric linkage analysis, allele sharing, genetic disease models, inheritance vectors, score functions, families of score function definitions, genetic models, NCP-optimality, IBD-sharing structures, equivalence of score functionse

1 Introduction

In *linkage analysis* (Ott, 1999) or, in a wider sense, *gene mapping* (Haines and Pericak-Vance, 2006; Siegmund and Yakir, 2007) one searches for disease loci along genetic regions of interest; in other words, through what we refer to as a *genome*. This is done by observing so called *genotypes* and *phenotypes* of a *pedigree set*, i.e. a set of multigenerational families, throughout the genome. The rationale for doing this is that, at a disease locus, the genotypes and phenotypes should generally show correlation of some strength on the individual level within the pedigree, where the actual strength depends on the structure of disease, i.e. the so called *genetic model*. Observed present correlations, measured through some kind of test statistic, suggests localizations of loci corresponding to underlying disease genes or, at least, it narrows down the interesting genome regions to neighbourhoods of the findings. The amount of trust put into such loci actually being disease-related are generally evaluated, in a standard sense, through statistical significance calculations; preferably corrected for the multiple testing throughout the genome. An example of a small pedigree set is given in Figure 1. To further reduce the size of a plausible region for an interesting disease finding, i.e. to use a fine-mapping technique, one may, for instance, use methods from the toolbox of *association analysis* (see Balding, 2006).

1.1 Basic notation and concepts

In practise the genotypes are observed as well-defined *allelic types* at polymorphic marker loci located along the genome of interest. Vaguely speaking, a marker locus might be seen as an, in some sense, observable short chromosomal segment and it is polymorphic if several types of genetic observations are possible, in the underlying population, with respect to this segment. Hence polymorphic markers correspond to genetic variation in the population.

Correspondence: Lars Ängquist, Centre for Mathematical Sciences, Department of Mathematical Statistics, Lund University, Lund, Sweden. Mariedalsvägen 33, S-21745, Malmö, Sweden. Tel: 0046-(0)70-3586673; Email: lars.angquist@matstat.lu.se.

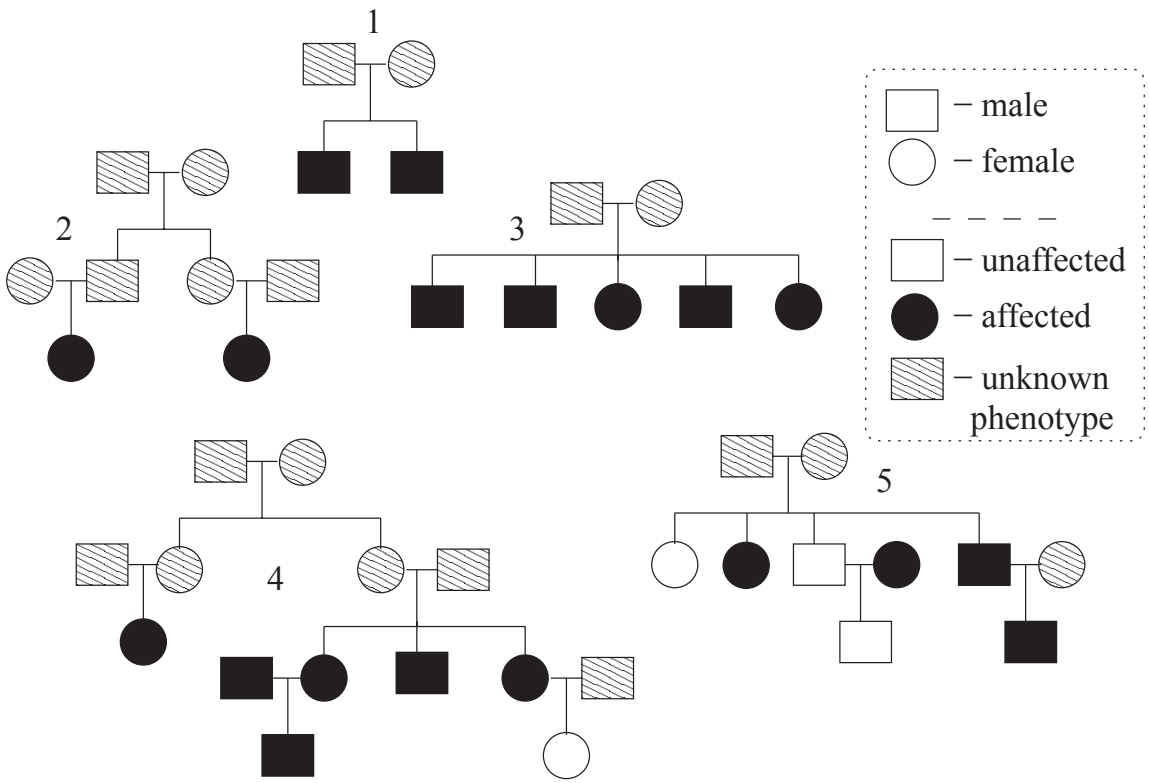

Figure 1. A pedigree set example consisting of 5 distinct pedigrees of different structures and phenotype settings.

Example 1 (Alleles and genotypes) *Consider a situation where we have a polymorphic locus with respect to three distinct possible allelic types-outcomes A, B and C within the population. Hence, at this locus, a specific individual will have any of the six consistent unordered genotypes; AA, AB, AC, BB, BC and CC, with certain probabilities jointly summing to one. For more information on, for instance, alleles, genotypes and genetic markers cf. Strachan and Read (2003).*

As a restriction or application, in *nonparametric linkage (NPL) analysis* (Whittemore and Halpern, 1994; Kruglyak et al. 1996; Ängquist, 2007) one searches for genetic linkage between disease and marker locus by observing and analyzing marker genotype data, without explicitly assuming a known genetic disease model. As noted above the linkage analysis approach may somewhat vaguely be described as analyzing the amount of dependence or correlation between genotypes and phenotypes among the observable individuals in the data set at hand, and hence in the nonparametric case one does not incorporate information on any disease loci in the standard analysis or search. Most oftenly such data is then taken to be representative of a homogeneous underlying population. Note that generally the phenotypes are assumed to be qualitative in the working form of indicators of *disease status*.

In this context the prime quantity of central importance to the actual statistical analysis-procedure is the process of *inheritance of alleles*. Each individual inherits two alleles, i.e. a genotype, at each chromosomal locus; one from the father and one from the mother. The inherited alleles themselves originates from either the corresponding grandfather or the grandmother and this leads to the following statement:

For a single pedigree, at locus x, the inheritance process may be totally described by the binary—zero-one—*inheritance vector* (Donnelly, 1983),

$$v(x) = (p_1, m_1, p_2, m_2, ..., p_{n-f}, m_{n-f}) \qquad (1)$$

where p_i and m_i correspond to the ith nonfounder's paternal and maternal allele respectively, i.e. each value is connected to one of the $m = 2(n - f)$ specific meioses.[1] Note that, for instance, one may

[1]A *nonfounder* (*founder*) has both (has not any) of its parents included in the pedigree. The rationale for the number of meioses being $m = 2(n - f)$ is that, which follows from above, each of the $n - f$ nonfounders corresponds to two meioses. Here n and f are the total number of individuals and the number of founders in the pedigree respectively.

in practise let 0 and 1 correspond to inheriting grandpaternal and grandmaternal alleles respectively.

Example 2 (Founders and nonfounders) *Consider the pedigrees in Figure 2. In both cases the parents constitutes the set of founders, whereas the siblings are the nonfounders.*

In the same manner as (1), but somewhat more obscure, one may summarize inheritance through *IBD-sharing structures*, where IBD means *identical-by-descent*. Two alleles are IBD if they are both ancestrally inherited from the same unique founder allele[2] with respect to the corresponding pedigree. Basically, forming IBD-sharing structures means grouping the elements of the set of all the 2^m possible inheritance vectors, $\mathbb{V}$, according to some pedigree-relational symmetry rules, into distinct IBD-groups. Such symmetry rules are, at least in principle, to some extent subjective. A commonly accepted example is that inheritance vectors will fall into the same group if they correspond to, i.e. one gets the same inheritance structure, permuting the inheritance of two siblings (with corresponding offsprings). Most oftenly, this is accepted even if the siblings being of distinct sexes. For more information, see Example 3 and Appendix A.

To numerically facilitate analysis of inheritance and phenotype-genotype dependence one may introduce a *score function*. Expressed in general terms this is just a function S giving a (numerical) score $S(v)$ to each possible inheritance vector $v \in \mathbb{V}$, i.e. it serves as a representation of the quantification of phenotype-genotype correlation.[3] Normally one searches for inheritance-wise deviations in the form of increased allele-sharing among affecteds,[4] since this indicates presence of genetic linkage between the marker and disease loci. As a consequence one aims at giving inheritance vectors consistent with such increased sharing high scores. On the other hand vectors being nonconsistent, in this sense, are then given low scores.

[2]In other words, they are both inherited instances of $g_i \in \mathbb{G}$, for some $i = 1, 2, \ldots, 2f$, where $\mathbb{G}$ is the set of all $2f$ founder alleles and f is the number of founders.

[3]One may note that this notion of a score function may be seen as adopting a data-mining perspective where such functions are used for scoring patterns (Hand et al. 2001). In this case one observes and scores inheritance patterns.

[4]Or more generally within phenotype-groups.

Assumption 1 *We assume that score functions are invariant within IBD-sharing structures. Explicitly, this implies that each inheritance vector v corresponding to a specific structure A produces the same output (score), i.e.*

$$S(v) = S(w);\ \forall v, w \in \mathbb{V}_A,$$

where $\mathbb{V}_A$ is the equivalence class including all inheritance vectors corresponding to structure A.

Hence considering a pedigree with m corresponding meioses leads to 2^m possible scores,

$$\mathbb{S} = \{S(v_1), S(v_2), \ldots, S(v_{2^m}),\} = \{s_1, s_2, \ldots, s_{2^m}\}, \quad (2)$$

assuming some order of inheritance vectors.[5] In this setting some scores will according to symmetry, and in some cases by—explicitly or implicitly—definition, be numerically equal. Using the context of IBD-sharing structures one may reformulate (2) as

$$\mathbb{S} = \{s_1, s_2, \ldots, s_n\}, \quad (3)$$

where the index corresponds to the by-score-ordered set of IBD-sharing structures, i.e. a natural restriction (order) is given by assuming $s_1 < s_2 < \ldots < s_n$. Quite naturally one may note that $n \leq 2^{m-f}$.

Remark 1 *In fact one may instantly note that $n \leq 2^{m-f}$, where f is the number of founders, which follows from what is generally referred to as 'founder couple reduction' (Kruglyak* et al. *1996; Gudbjartsson* et al. *2000). This is an inheritance symmetry property originating from the uncertainty of founder phases (ambiguity of inheritance vector interpretation).*

Example 3 (Founder couple reduction) *For an affected sib-pair (ASP), see Figure 2, one may illustrate Remark 1 through the following example: Let the parents have genotypes {A, B} and {C, D}. Oftenly the inheritance vector is defined with each position corresponding to a well-defined paternal or maternal allele of a specified nonfounder; see (1). This is then also reflected in the ordering of the alleles (including the founder*

[5]For instance using the standard decimal interpretation (conversion) of the binary zero-one inheritance vector with length m.

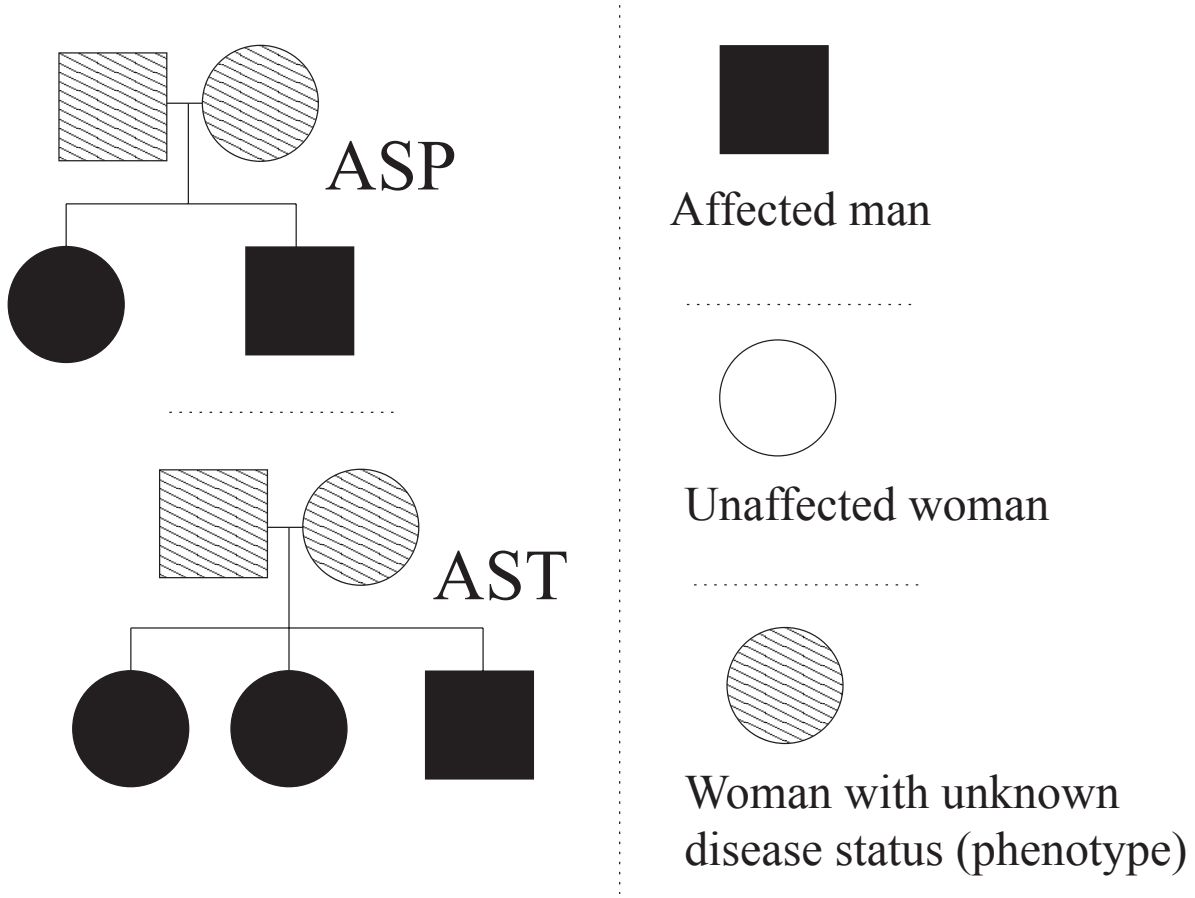

Figure 2. The pedigree structures corresponding to affected sib-pair (ASP) and affected sib-trio (AST) pedigrees.

alleles) in the sense that, for instance, the left allele (A and C) correspond to paternal inheritance and the right allele (B and D) to maternal inheritance. Since most likely[6] *the ordering (phases) of the founder alleles is unknown we, in these cases, do not really know which of the following ordered founder-genotypes that is truth-valid:*

AB/CD, AB/DC, BA/CD, BA/DC.

The implication of this is that all inheritance vectors related through transformations between these founder-genotypes are inheritance-wise evidentially equivalent. (Hence giving rise to equivalent IBD-sharing structures; see Appendix A.)

For further information on equivalent IBD-sharing structures consider Appendix A.

[6]According to the fact that the pedigree construction excluded farther (earlier) generations.

1.2 Aims and scope

Our primary goal with this paper is to, in such a generally accessible way as possible, formalize and discuss the structure of nonparametric linkage score functions. Oftenly, in published works, these functions are either directly applied using some of the standard instances or derived in an ad hoc or highly theoretical, or non-intuitive, fashion.

Having this in mind, the text to follow is not a complete summary of suggested and published score function variants, or the most theoretical exposition out there. Rather, it aims at being a review-like overview discussing the underlying structure, contexts of derivations and interpretations (and to some extent performance) of certain families of NPL score functions.

In *Section 2* three distinct such families—the implicit, the explicit and the optimality-based one—are introduced and discussed, whereas *Section 3* gives a new generalization of an existing optimality-based function. A small simulation study with respect to five distinct score functions of

various types is performed in *Section 4*. The two appendices, *Appendix A* and *Appendix B*, discusses equivalence-properties with respect to structure and standardization of score functions respectively.

2 Approaches to Score Function Definitions

For an underlying disease to be genetically inheritable, i.e. to include a *genetic component*, some kind of correlation between the phenotype and the disease genotypes must exist. This is usually described by means of a *genetic model* λ. One may note that λ usually, at least to some extent, is unknown so, if needed, it is estimated prior to analysis using so called *segregation analysis* (Khoury et al. 1993; Haines and Pericak-Vance, 2006). The complete, possibly multilocus, genetic model may be summarized as,

$$\lambda = (p, f, l), \tag{4}$$

where p is the set of *disease allele frequencies*, f is the set of *penetrance values*, describing the link between phenotypes and disease genotypes, and l defines the *disease loci positions*.

Now, to define a score function one basically has to instantiate the numerical scores corresponding to (2) or (3). This may be done in several distinct ways, which is furtherly discussed below. What truly is the core question with respect to such definitions is the evidential performance of the corresponding score function. (Most likely in the form of *statistical power* calculations.) One may note that the relative performance of different score functions depends on the underlying genetic model λ and the combined present pedigree-structure of the pedigree set.

A score function performing well under a wide range of different $\lambda \in \Lambda$, where Λ is the set of all possible disease models, is termed a *robust* score function. The best score function with respect to a criterion C and disease model λ is called an *optimal* score function $S_{\text{opt}} = S^C(v|\lambda)$.

2.1 Implicitly defined versions

Vaguely speaking, as noted above, at a true disease locus, the IBD-sharing within phenotypes should be expected to increase. This makes it possible to define functions, depending on pedigree IBD-sharing only, meeting this requirement (property). Since such functions implicitly instantiate (2) and (3) through the higher-level sharing-based function definition we call them *implicitly defined score functions*. Next, we will note on two distinct such definitions.

2.1.1 Traditional score functions

Firstly, S_{pairs} (Weeks and Lange, 1988) is based on IBD-sharing among all pairs of affected individuals in the pedigree,

$$S_{\text{pairs}}(v) = \sum_{(a_i, a_j) \in \mathbb{A}} \text{IBD}(a_i, a_j), \tag{5}$$

where $i < j$, $\mathbb{A}$ is the set of affecteds in the pedigree[7] and IBD(x, y) is the number of alleles shared IBD between individuals x and y.

Secondly, S_{all} (Whittemore and Halpern, 1994) is based on the simultaneous IBD-sharing among all the affecteds in the pedigree,

$$S_{\text{all}}(v) = \frac{1}{2^{|\mathbb{A}|}} \sum_{h \in \mathbb{H}} \prod_{i=1}^{2f} b_i(h)!, \tag{6}$$

where $|\mathbb{A}|$ is the number of affecteds, $\mathbb{H}$ is a set containing the elements corresponding to all ways of selecting one allele from each affected, $2f$ is the number of founder alleles in the pedigree and $b_i(h)$ is the number of times the ith founder allele is present in selection $h \in \mathbb{H}$.[8]

Example 4 (Score function S_{all}) *For an ASP (Fig. 2) one may examplify (6) as follows: Let the parents have genotypes {A, B} and {C, D}. Then, if the affected siblings inherit {A, C} and {A, D} respectively we have the following h-selection possibilities ($\forall h \in \mathbb{H}$):*

$$h_1 = \{A, A\}; h_2 = \{A, D\}; h_3 = \{C, A\}; h_4 = \{C, D\},$$

where for instance, treating A as the 1*st founder allele,* $\prod_{i=1}^{4} b_i(h_1) = 2!0!0!0! = 2$.

Remark 2 *Both (5) and (6) give high (low) scores to excess (low) IBD-sharing. The difference lies in that the latter one, relatively seen, upweight increased sharing of specific founder alleles within*

[7] Including the ordered affected individuals $a_1, a_2, \ldots, a_{|\mathbb{A}|}$, where $|\mathbb{A}|$ is the number of affecteds or, equivalently, the *cardinality* of the set $\mathbb{A}$.

[8] Each specific selection h consists of $|\mathbb{A}|$ alleles that may be grouped according to their ancestral history, i.e. each allele is a copy of one of the $2f$ founder alleles. The link to the number of members in the ith group is $b_i(h)$, i.e. where $b_i(h)$ is the number of g_i alleles; see Footnote 2.

large groups of individuals, thus reflecting a higher degree of belief in such inheritance evidence.

2.1.2 Extended score functions

Both functions (5) and (6) are defined, given the inheritance vector v, with respect to the set of affecteds $\mathbb{A}$ only, which might be notationally pointed out as $S(v) = S(v \mid \mathbb{A})$. Henceforth we refer to such score functions as *traditional* score functions. In fact a vast majority of the most commonly used functions are of this kind.

In Ängquist (2006) several extensions to traditional score functions are given. Now, assume a traditional instance S and let S' denote a corresponding *extended* version. A first-level extension is to combine information from both phenotype groups (affecteds as well as unaffecteds) through

$$S' = S'(v) = S'(v \mid \mathbb{A} \cup \mathbb{U}\mathbb{A}) = S(v \mid \mathbb{A}) + S(v \mid \mathbb{U}\mathbb{A}). \quad (7)$$

This aims at additionally searching for unusual IBD-sharing within the set of unaffecteds $\mathbb{U}\mathbb{A}$. Note that $S(v \mid \mathbb{U}\mathbb{A})$ in practise means, given inheritance vector v, applying the traditional score function S to the same pedigree set, in the standard way using the same function-definition, but with phenotypes interchanged between affecteds and unaffecteds.[9]

Example 5 (Extended score functions; phenotype-switching) *Consider the pedigree consisting of two parents (unknown phenotypes) and four siblings (A, B, C and D) in Figure 3. When calculating $S(v \mid \mathbb{A})$ this is done with respect to Siblings A and D. After the phenotype-switching process displayed in Figure 3, $S(v \mid \mathbb{U}\mathbb{A})$ is calculated using Siblings B and C. Note that the actual score function algorithm, for instance underlying (5) or (6), is the same in both cases.*

A second-order extension may be formulated as

$$\begin{aligned} S' = S'(v) &= S'(v \mid \mathbb{A} \cup \mathbb{A}\mathbb{U} \cup \mathbb{U}\mathbb{P}) \\ &= [S(v \mid \mathbb{A}) - S(v \mid \mathbb{A} \cup \mathbb{A}\mathbb{U} \cup \mathbb{U}\mathbb{P}) \\ &\quad + [S(v \mid \mathbb{U}\mathbb{A}) - S(v \mid \mathbb{A} \cup \mathbb{U}\mathbb{A} \cup \mathbb{U}\mathbb{P})] \quad (8) \\ &= S(v \mid \mathbb{A}) + S(v \mid \mathbb{U}\mathbb{A}) \\ &\quad - 2S(v \mid \mathbb{A} \cup \mathbb{U}\mathbb{A} \cup \mathbb{U}\mathbb{P}), \end{aligned}$$

where $\mathbb{U}\mathbb{P}$ denotes the set of individuals with unknown phenotype. Here one additionally corrects for the *overall* sharing within the pedigree, i.e. it compares the IBD-sharing (through the traditional function S) within phenotype-groups to what is jointly given on the pedigree-level.

Remark 3 *An intuitive critiscism to extensions as (7) and (8) might be that unaffecteds is not to the same extent as affecteds a secure (final) phenotype, since in time such individuals might turn into affecteds.*[10] *However, this could be solved by letting ambiguous cases, according to some criterion, being labelled as having affection status unknown, i.e. as $\mathbb{U}\mathbb{P}$-individuals. Further, given a well-defined probability model for (possible) affection time one may weight the analysis (scores) with respect to this model.*

Remark 4 *Another objection against usage of unaffecteds in this way may be raised if a disease is not purely caused by gene mutations, but rather through a combination of genetic and environmental factors. In this case the unaffecteds in the pedigree are not obviously good representatives of the normal population in terms of genetic composition. It is then logically possible that such unaffecteds still share a common genetic background with the affected relatives, but lack certain environmental factors; or lack some trigger events in their health history. Hence, their role in gene mapping is in this case of secondary importance.*

2.2 Explicitly defined versions

It is perfectly possible not to use a closed definition or high-level algorithm when calculating the vector of scores constituting the corresponding score function. We refer to such cases as *explicitly defined score functions*.

The construction of an explicit score function reduces to (explicitly) distributing scores to all present IBD-sharing structures, thus reflecting numerically the assumed connection between these sharing structures and evidence for a present disease locus. For instance, such an approach might be interesting if one can show by some real examples, or a priori assume, that certain combination of inheritance vector states are impossible or unlikely.

[9]Generally $S(v|\mathbb{X})$ means applying the traditional score function S, given v, to the arbitrary pedigree subgroup $\mathbb{X}$. Also note that $\mathbb{A} \cup \mathbb{U}\mathbb{A}$ equals the subgroup of all individuals in the pedigree with *known* phenotype.

[10]This might be the case for, in some relative (to the disease) sense, *young* individuals. These persons might be interpreted as what we later refer to as *ambiguous* cases.

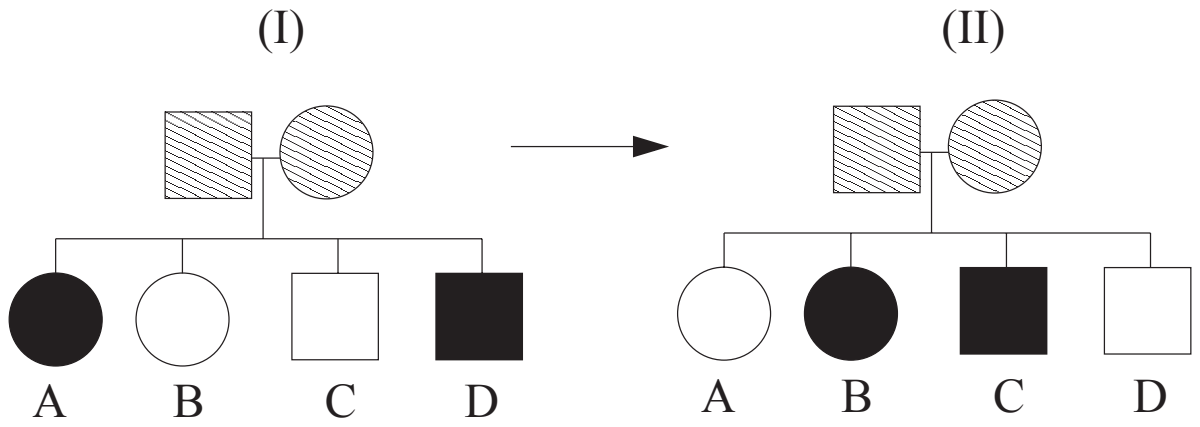

Figure 3. A pedigree consisting of 4 siblings (two affecteds, two unaffecteds). The two distinct cases (left to right) display the corresponding phenotype-switching process involved in the definition of extended score functions.

Example 6 (Explicit ASP-definition) *Once more, consider an ASP. Here three IBD-sharing structures (with scores s_1, s_2 and s_3) are possible corresponding to the sib-pair sharing 0, 1 and 2 alleles IBD respectively. Arbitrarily fixing s_1 and s_3 with $s_1 < s_3$ the closure of an explicit definition is reflected by the choice of s_2 with the restriction of $s_1 \leq s_2 \leq s_3$; see Section 2.4 and Appendix B.*

Explicit definitions, so to speak, implicitly make some (though quite vague) assumptions on the type of underlying disease structure. In this sense they are more strongly directed towards certain disease models than implicit definitions, but much less so than the family of definitions described below in Section 2.3. There explicit assumptions on true (plausible) genetic disease models λ under corresponding alternative hypotheses H_1 are made.

2.3 Optimality defined versions

If having an explicit algorithm (as for implicitly defined versions) but where this algorithm is formulated with respect to, in some sense, an optimality criterion C, we say that we deal with *C-optimal* score functions.

Given a disease model λ, define the expected score at the disease locus under this model as

$$E(S \mid \lambda) = \sum_{w \in \mathbb{V}} S(w) P(w \mid \lambda), \qquad (9)$$

where $P(w|\lambda)$ is the *inheritance distribution* under disease model λ. The expected value in (9) is referred to as the *noncentrality parameter (NCP)*. It is showed in Ängquist et al. (2007) (based on results given in Hössjer, 2005) that optimal score functions with respect to (maximization of) NCPs may be expressed as

$$S(w) \propto P(w \mid \lambda) - 2^{-m}, \qquad (10)$$

with m equaling the number of meioses. This approach might be interpreted as basing the scores on the difference between inheritance vector-probabilities under the null and alternative hypothesis in all cases.[11] The rationale for being interested in NCPs are that this concept is closely linked, but not equivalent, to statistical power (Feingold et al. 1993).

Hence one may note that the optimal score function (10) depends on the true genetic model and should be interpreted as, in this sense, the best possible result that the investigator might expect when the genetic model is correctly specified. In practice though, the genetic model is often unknown. Then in a natural way, for each choice of score function and for a range of different genetic models, (10) facilitates comparisons with optimality, leading to a quantification of the apparent loss of information. The optimal score function might also serve as a form of explicit score function with respect to certain assumptions or prior information.

Further, in Hössjer (2003) *locally most powerful* tests are outlined using specific parametric models (in the form of exponential expansions) for the inheritance distribution under alternatives.

[11] Note that $P(w|H_0) = 2^{-m}$ for all $w \in \mathbb{V}$.

Consider also the discussions in Whittemore (1996), Kong and Cox (1997), Nicolae (1999) and McPeek (1999).

2.4 Equivalent score functions

As a way of enhancing interpretation one usually uses *standardized* versions of the score functions. Standardization is performed through

$$S(v) \leftarrow \left[\frac{S(v) - \mu}{\sigma} \right], \qquad (11)$$

where, for a pedigree with m meioses,

$$\begin{cases} \mu = E(S \mid H_0) = \sum_i 2^{-m} S(w_i) \\ \sigma^2 = V(S \mid H_0) = \sum_i 2^{-m} S(w_i)^2 - \mu^2 \end{cases},$$

are the mean and variance of S prior to standardization; under the null hypothesis H_0 of no linkage and where summation is over all the 2^m distinct elements $w \in \mathbb{V}$.[12]

Remark 5 *Note that S on the right-hand side in (11) is referred to as an 'unstandardized' score function, whereas S on the left-hand side is a 'standardized' score function.*

Equipped with the concept of standardization one may define *equivalent* (unstandardized) score functions. In order to define this concept in a clear and straighforward manner we need the following additional assumption.

Assumption 2 *We assume that there is a general agreement on the order of the IBD-sharing structures, i.e. that* s_i $(\forall i)$ *in (3) correspond to the same structure regardless of which score function you choose.*

If two unstandardized score functions through standardization are transformed to equal[13] standardized score functions they are referred to as being equivalent. For more detailed information and corresponding equivalence-criterions, see Appendix B.

Example 7 (Equivalence of S_{pairs} and S_{all} for ASPs) *For an ASP the score functions* S_{pairs} *and* S_{all}*, defined in (5) and (6) respectively, are equivalent. This follows since they both lead to the distinct standardized numerical scores* $\{-\sqrt{2}, 0, \sqrt{2}\}$*. Adopting the approach in (2) these scores correspond to, in turn, 4, 8 and 4 distinct inheritance vectors related to the ASP sharing 0, 1 and 2 alleles IBD respectively. (Here we have* $m = 4$ *meioses and* $2^m = 16$ *unique inheritance vectors.) Alternatively, one may use (3) where* $\mathbb{S}$ *only contain these three scores (structures), which are then attained with probabilities 0.25, 0.50 and 0.25 respectively under* H_0*.*

One may also note that actual numerical standardized scores corresponding to a specific score function (or several equivalent ones) are dependent on the score distribution $P(s|H_0)$ under the null hypothesis H_0, which is given by the actual pedigree structure and phenotype setting.[14]

2.5 Real studies and data

Note that throughout this article we try to discuss score functions without explicitly mentioning the actual test statistics they are used in connection with when facing real and imperfect marker data MD.[15]

An exception is the use of standardization through (11) which implicitly refer to the practise of the 'NPL score' test statistic (Kruglyak et al. 1996; Ängquist, 2007).

$$Z(x) = E(S[v(x)] \mid \text{MD}), \qquad (12)$$

where the expected value, at locus x, is taken over $P\,[v(x)|\text{MD}]$ which is the inheritance distribution given the observed marker data.[16]

Given imperfect data the variance of the NPL score $V(Z) \leq V(S)$, hence if decreasing leading to conservative procedures assuming $V(Z) = V(S) = 1$. In order to increase the actual variance in data, hence reducing the conservativeness, one usually bases real studies on so called *multipoint analysis*, where all inheritance information from the surrounding chromosome is used when calculating the inheritance distribution at a locus. Here the calculations are preferably performed using *Hidden Markov Models (HMMs)* through the *Lander-Green-Kruglyak algorithm*; see Lander and Green

[12]Note that we end up with the standardized properties $E(S|H_0) = 0$ and $V(S|H_0) = 1$.

[13]Two score functions S^1 and S^2 (unstandardized or standardized) are equal if, using the formulation of (3), $s_i^1 = s_i^2$ for all i.

[14]This follows since these settings uniquely define the standardization parameters μ and σ in (11).

[15]In other words, when the complete inheritance process over corresponding loci is not known with probability one.

[16]Note that (12) refer to a single pedigree (*pedigree-specific NPL score*; see Ängquist (2007) for more information). For a full pedigree set one uses pedigree-weighted sums with respect to such present scores.

(1987), Kruglyak et al. (1995) and the expositional review in Ziegler and Koenig (2006).[17] Actually, the complete marker data assumption seems fairly realistic when all pedigree members are genotyped with a density of SNP markers of at least, say, 0.1 cM.

Replacing σ^2 in (11) with $V(Z)$ at each loci leads to the interpretation of the standardized score as a *common statistical score function* based on the derivative of a corresponding likelihood function (see Kong and Cox, 1997).[18]

However, note that although the choice of test statistic and possible standardization procedure are important from a testing and statistical significance perspective it is not particularily essential for the present discussion. Moreover, generally the interpretations and relative performances of the different score function variants will not change when dealing with imperfect data, hence this matter is only noted on in this specific subsection.

2.6 Two-locus score functions

One may generalize the one-locus procedure above in order to simultaneously, or sequentially, search for two distinct disease loci on the genome. The former case is referred to as an *unconditional* analysis, whereas the latter case is a *conditional* analysis performed conditioning on some kind of genetic information at one, or several, conditioning loci. One may generally use the same basic score function definitions in both cases, taking into account that the standardizations will differ.

Implicitly defined score functions may in some cases be relatively easily generalized to the two-locus case, but in some cases the corresponding score-algorithm will be refrainingly more complex. As a positive example, one may generalize (5) into a two-locus score function. In Ängquist et al. (2007) the following, quite general, formulation is given

$$S_{\text{pairs}}(w_1, w_2) = \sum_{i<j} [\text{IBD}_{i,j}(w_1) + \text{IBD}_{i,j}(w_2)]^k, \quad (13)$$

where $\text{IBD}_{i,j}$ (w_i) equals $\text{IBD}(a_i, a_j)$ in (5) with respect to inheritance vector w_i, related to the ith (disease or marker) loci, and $\{a_i, a_j\} \in \mathbb{A}$.

For $k > 1$ (13) may be thought of as trying to capture epistatic joint pairwise IBD-sharing within a pedigree. The case $k = 1$ of (13) corresponds to the additive score function used in Strauch et al. (2000),

$$\begin{aligned} S_{pairs}(w_1, w_2) &= \sum_{i<j} [\text{IBD}_{i,j}(w_1) + \text{IBD}_{i,j}(w_2)] \\ &= \sum_{i<j} \text{IBD}_{i,j}(w_1) + \sum_{i<j} \text{IBD}_{i,j}(w_2) \\ &= S_{\text{pairs}}(w_1) + S_{\text{pairs}}(w_2), \end{aligned}$$

which these authors also implemented into the analysis program GENEHUNTER-TWOLOCUS. In the applications of Ängquist et al. (2007) the case $k = 2$ is used, which shows close to NCP-optimal performance for the one-parameter genetic disease model families used in their simulations.

Example 8 (ASP score matrix) *For ASPs one might summarize a two-locus score function completely using a* 3×3 *score matrix.*[19] *Letting*

$$S(i, j) = S(IBD_1 = i, IBD_2 = j),$$

where $IBD_k = l$ *means that the ASP shares l alleles IBD at the kth (marker or disease) locus, leads to the general score matrix*

$$\mathbb{S} = \begin{bmatrix} S(0,0) & S(0,1) & S(0,2) \\ S(1,0) & S(1,1) & S(1,2) \\ S(2,0) & S(2,1) & S(2,2) \end{bmatrix} \quad (14)$$

Several instances and substructures of (14) are given, implemented and discussed in Ängquist et al. (2005).

Two-locus explicitly defined score functions are concept-wise straightforward generalizations of one-locus ones. Moreover, the NCP-optimal score function (10) of Ängquist et al. (2007), for unconditional and conditional two-locus analysis respectively, may be generalized to

[17] A textbook on HMMs is Cappé et al. (2005).

[18] On *standard score functions* see e.g. Clayton and Hills (1993) or Garthwaite et al. (1995). A specialized monograph on the theory and philosophy of the *likelihood approach* is Edwards (1992).

[19] Note that this is possible according to our assumption of scoring all inheritance vectors leading to similar IBD-sharing structures equally. In this case, at each locus, the *three* distinct IBD-sharing structures correspond to the number of alleles (0, 1 or 2) shared IBD by the affected sib-pair.

$$\begin{cases} S(w_1, w_2) \propto P(w_1, w_2) - 2^{-2m}, \\ S(w_1 \mid w_2) \propto P(w_1 \mid w_2) - 2^{-m}. \end{cases} \quad (15)$$

Note that the interpretation of these scores as being proportional to probability-based differences with respect to the null and (assumed) alternative hypotheses still hold true.

3 Generalization to the Optimality-Based Definition Approach

In some cases where the true disease model λ is fully or partially unknown the usage of the NCP-optimal score function (10) based on an estimate (or assumption) $\hat{\lambda}$ may be considered to lack robustness and applicability. In order to reduce unnecessary usage-avoidance we will next try to further generalize this approach, hence increasing its practical usefulness. More explicitly, our suggested method is adapted to include prior, assumed or intuitive, information on λ in a more direct sense than what is available through using explicit versions, but still avoiding the assumption of a *single* plausible genetic disease model.

3.1 Algorithm

Begin with choosing d distinct genetic disease models

$$\{\lambda_1, \lambda_2, ..., \lambda_d\} \in \Lambda,$$

with inheritance distributions under corresponding alternatives

$$P_i = P(w \mid \lambda_i); i = 1, 2, ..., d.$$

Now, a simple generalization to the previous score in (10) is given by

$$S(w) \propto \frac{\sum_{i=1}^{d} [P(w \mid \lambda_i) - 2^{-m}]}{d}, \quad (16)$$

where d in the denominator in principle is unnecessary (according to the standardization) but makes comparisons between (10) and (16) possible in a natural way.

A further generalization arises if adopting a *Bayesian perspective* with respect to the prior distribution of possible disease models.[20] Fix d and let $\pi = (\pi_1, \pi_2, \ldots, \pi_d)$, with $\Sigma_{i=1}^{d} \pi_i = 1$, be the vector of prior probabilities corresponding to the d distinct disease models. This leads to (16) being generalized into

$$S(w) \propto \sum_{i=1}^{d} \pi_i [P(w \mid \lambda_i) - 2^{-m}]. \quad (17)$$

One may note that (16) is the special case of (17) where $\pi = (1/d, 1/d, \ldots, 1/d)$ and that (10) correspond to $d = 1$ and hence $\pi = \pi_1 = 1$ for a single disease model λ_1. Finally, observe that the NCP-optimality property (Ängquist et al. 2007) is kept if (in a somewhat abstract sense) π, given the present knowledge-base, is the true probability distribution with respect to the present genetic disease model-ambiguity.

4 A Small Simulation Study

For illustrational purposes we include a small-scale simulation analyses in this section. We perform power calculations for various settings and present them through *ROC-curves*, i.e. as plots with significance levels versus power with respect to a set of underlying score thresholds (Selin, 1965; Bradley, 1996). The results are given, and graphically displayed, in Figures 4–6.

4.1 Simulation set-up

Consider a pedigree consisting of two parents of unknown phenotypes and M siblings. For instance, Pedigree 3 in Figure 1 is such a pedigree with $M = 5$. We construct three *homogeneous* pedigree sets, i.e. a set consisting of pedigrees with similar structure and phenotype setting only, based on three distinct such pedigrees with $M = 6$: (i) Pedigree 1 consisting of 4 affected and 2 unaffected siblings. (ii) Pedigree 2 consisting of 3 affected and 3 unaffected siblings. (iii) Pedigree 3 consisting of 2 affected and 4 unaffected siblings. The number of pedigrees in each pedigree set is put to $N = 15$.

Further, for each case we use a genome consisting of a single chromosome of length $G = 4$ Morgans, $J = 2000$ simulations and score thresholds ranging from $T = 3$ to $T = 10$. The analyses are made with respect to five distinct score functions: The $S_1 = S_{\text{pairs}}$ function in (5), the $S_2 = S_{\text{all}}$ function in (6), the extended version, S_3, using (7) for S_{pairs}, the extended version, S_4, using (7) for

[20] Subjective or empirically objective perspective; for concepts see e.g. Winkler (1972) and Gelman et al. (2004).

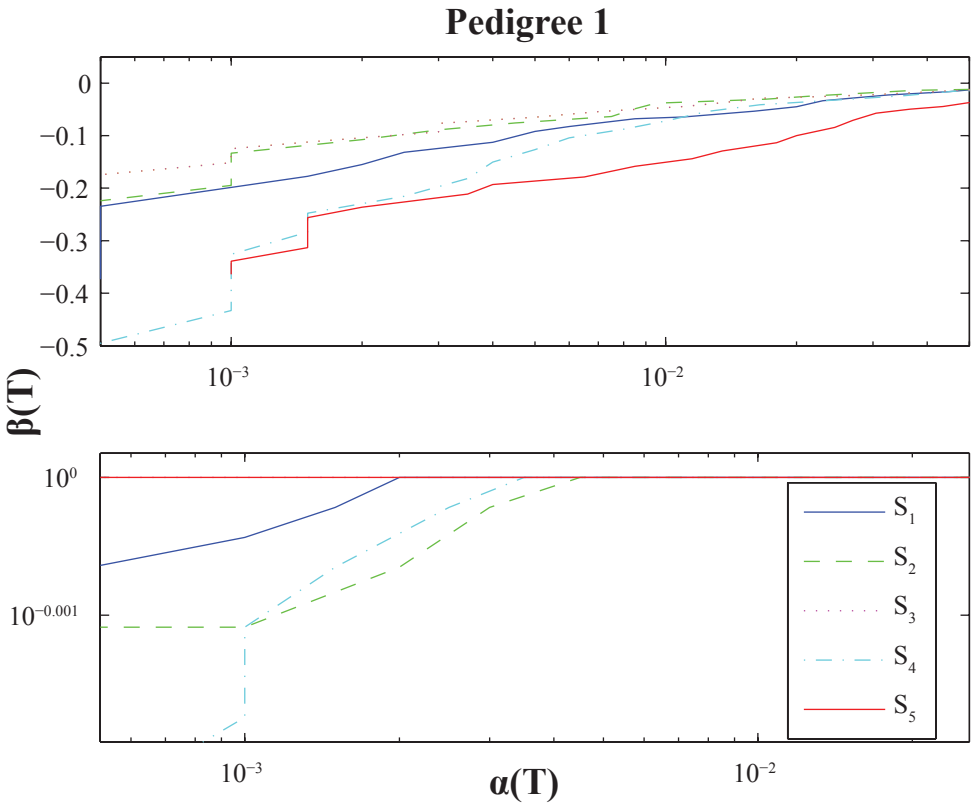

Figure 4. Power calculations for Pedigree 1 and score functions S_1-S_5. Presented as ROC-curves with significance levels $\alpha(T)$ vs. powers $\beta(T)$ for score thresholds T. (Logarithmic X/Y-scales.) Upper and lower panel uses penetrance vectors $f = (0.02, 0.2, 0.8)$ and $f = (0.02, 0.8, 0.8)$ respectively.

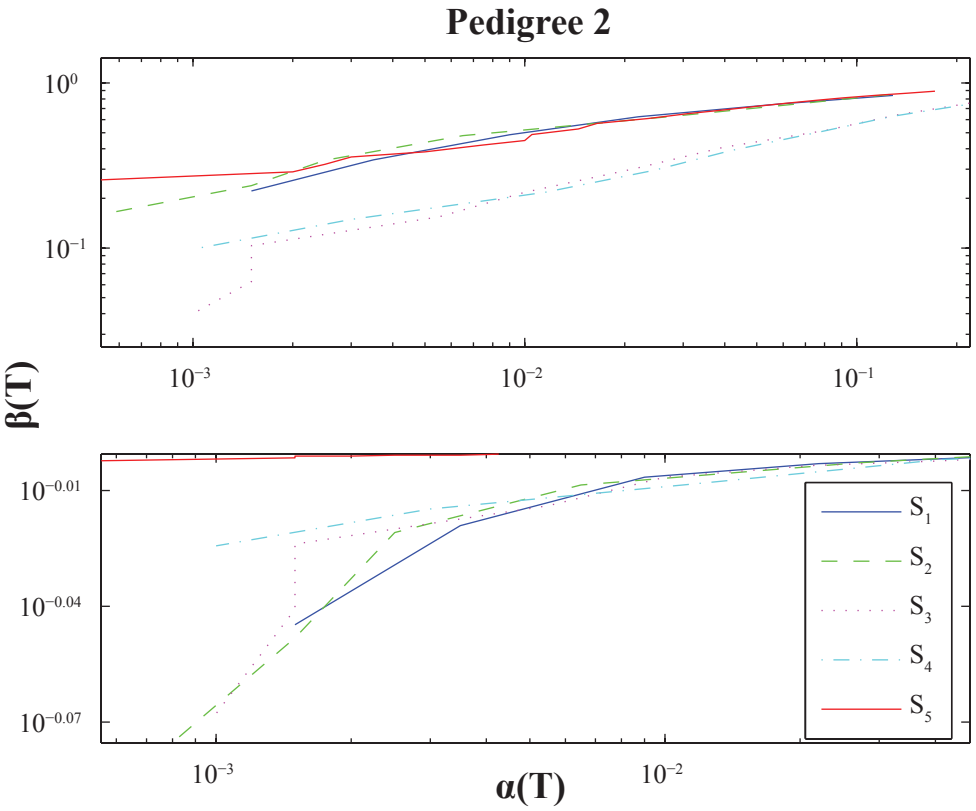

Figure 5. Power calculations for Pedigree 2. See caption of Figure 4.

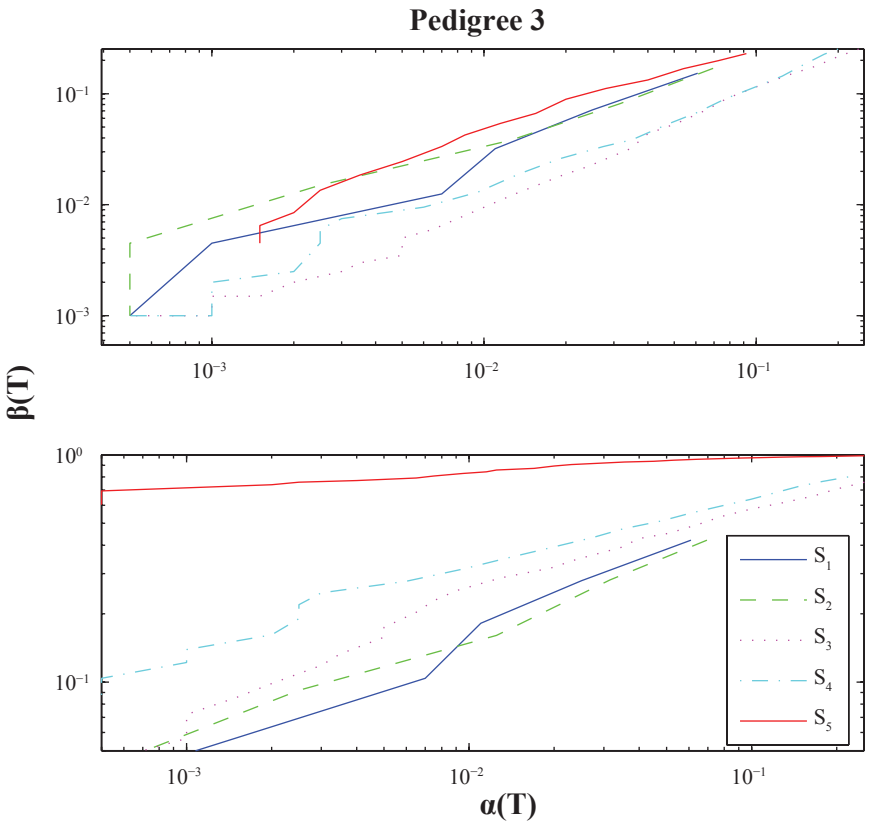

Figure 6. Power calculations for Pedigree 3. See caption of Figure 4.

S_{all}, the NCP-optimal score function S_5 in (10). All score functions are standardized through (11) and calculations are performed using the NPL score approach (12).

Finally, we used two genetic models, λ_1 and λ_2, where both correspond to disease allele frequency $p = 0.01$, but with distinct penetrance vectors,

$$f = (f_0, f_1, f_2) = (0.02, 0.20, 0.80) \text{ and } (0.02, 0.80, 0.80)$$

respectively. Here f_i denotes the probability for an individual, having a disease genotype consisting of i disease alleles and $2 - i$ normal alleles, of being affected.

4.2 Results and discussion

It is quite hard to draw very certain conclusions from such a small study, once more note that this section is in some sense a side-track, but a few general observations of some interest may be stated: (i) S_2 performs better than S_1 for Pedigree 1, whereas the opposite is true for Pedigree 2–3 under λ_2. In other words their relative performance is affected by the pedigree structure as noted above. (ii) The extended versions S_3 and S_4 often outperforms the traditional (nonextended) versions S_1 and S_2. These extensions seem somewhat more favourable for Pedigree 3 than for Pedigree 1, which seems reasonable since the latter pedigree has a structure more directed towards unaffected individuals. They also seem more advantageous under λ_2 than for λ_1, which might be explained by the latter model having more IBD-sharing discrimination power within the subgroup of unaffecteds; according to a higher disease penetrance for disease heterozygotes. (iii) The NCP-optimal score function S_5 is performance-wise much better under λ_2. Probably mainly follows from similar reasoning as given in the last sentence under (ii).

Acknowledgements

I send my best regards to Professor Ola Hössjer for prior co-authorship, discussions and ideas that strongly affected my appreciation and views of the concepts constituting this article. Thank you!

I am grateful also towards two anonymous reviewers for several insightful comments and suggestions.

References

Ängquist, L., Anevski, D. and Luthman, H. *Unconditional two-locus non-parametric linkage analysis: On composite null hypotheses with and without gene-gene interaction* (Tech. Rep. No. 2005:28). Lund: Department of Mathematical Statistics, Lund University.

Ängquist, L. 2006, June. *Some notes on the choice of score function in nonparametric linkage analysis.* (Free download from homepage: 'http://www.maths.lth.se/matstat/staff/larsa/').

Ängquist, L. 2007. *Pointwise and genomewide significance calculations in gene mapping through nonparametric linkage analysis: Theory, algorithms and applications* (Doctoral Thesis No. 2006:15). Lund: Department of Mathematical Statistics, Lund University.

Ängquist, L., Hössjer, O. and Groop, L. 2007. *Strategies for conditional two-locus nonparametric linkage analysis* (Tech. Rep. No. 2007:1). Lund: Department of Mathematical Statistics, Lund University. (Accepted for publication by 'Human Heredity'; will most likely appear in forthcoming 2008, 66:2 issue).

Balding, D.J. 2006. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics,* 7:781–91.

Bradley, A.P. 1996. ROC curves and the χ^2 test. *Pattern Recognition Letters,* 17:287–94.

Cappé, O., Moulines, E. and Rydén, T. 2005. *Inference in hidden Markov models* [Springer Series in Statistics]. New York: Springer.

Clayton, D. and Hills, M. 1993. *Statistical models in epidemiology.* Oxford: Oxford University Press.

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

Edwards, A.W.F. 1992. *Likelihood: Expanded edition* (Second Edition ed.). New York: John Hopkins University Press.

Feingold, E., Brown, P.O. and Siegmund, D. 1993. Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *American Journal of Human Genetics,* 53:234–51.

Garthwaite, P.H., Jolliffe, I.T. and Jones, B. 1995. *Statistical inference.* London: Prentice Hall.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. 2004. *Bayesian data analysis* (Second Edition ed.) [Texts in Statistical Science]. Boca Raton (Florida): Chapman *k* and Hall/CRC.

Gudbjartsson, D.F., Jonasson, K., Frigge, M. and Kong, A. 2000. ALLEGRO, a new computer program for multipoint linkage analysis. *Nature Genetics,* 25:12–3.

Haines, J.L. and Pericak-Vance, M.A. Eds. 2006. *Genetic analysis of complex disease.* New York: Wiley-Liss.

Hand, D.J., Mannila, H. and Smyth, P. 2001. *Principles of data mining.* Cambridge, Massachusetts: The MIT Press.

Hössjer, O. 2003. Determining inheritance distributions via stochastic penetrances. *Journal of the American Statistical Association,* 98:1035–51.

Hössjer, O. 2005. Information and effective number of meioses in linkage analysis. *Journal of Mathematical Biology,* 50(2):208–32.

Khoury, M.J., Beaty, T.H. and Cohen, B.C. 1993. *Fundamentals of genetic epidemiology* [Monographs un Epidemiology and Biostatistics, Volume 22]. New York and Oxford: Oxford University Press.

Kong, A. and Cox, N. 1997. Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics,* 61:1179–88.

Kruglyak, L., Daly, M.J. and Lander, E.S. 1995. Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *American Journal of Human Genetics,* 56:519–27.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. 1996. Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics,* 58:1347–63.

Lander, E.S. and Green, P. 1987. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America,* 85:2363–7.

McPeek, M.S. 1999. Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology,* 16:225–49.

Nicolae, D.L. 1999, Jun. *Allele sharing models in gene mapping: A likelihood approach* (Doctoral Thesis). Chicago: Department of Statistics, University of Chicago.

Ott, J. 1999. *Analysis of human genetic linkage* (Third ed.). New York: The John Hopkins University Press.

Selin, I. 1965. *Detection theory* [The RAND Corporation]. Princeton, New Jersey: Princeton University Press.

Siegmund, D. and Yakir, B. 2007. *The statistics of gene mapping* [Statistics for Biology and Health]. New York: Springer.

Strachan, T. and Read, A.P. 2003. *Human molecular genetics* (Third ed.). London and New York: Garland Science.

Strauch, K., Fimmers, R., Kurz, T., Deichmann, K.A., Wienker, T.F. and Baur, M.P. 2000. Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: Application to mite sensitization. *American Journal of Human Genetics,* 66:1945–57.

Weeks, D.E. and Lange, K. 1988. The affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics,* 42:315–26.

Whittemore, A.S. 1996. Genome scanning for linkage: An overview. *American Journal of Human Genetics,* 59:704–16.

Whittemore, A.S. and Halpern, J. 1994. A class of tests for linkage using affected pedigree members. *Biometrics,* 50:118–27.

Winkler, R.L. 1972. *An introduction to Bayesian inference and decision.* New York: Holt, Rinehart and Winston, Inc.

Ziegler, A. and Koenig, I.R. 2006. *A statistical approach to genetic epidemiology: Concepts and applications.* Weinheim: Wiley-WCH.

Appendix
Some Technicalities

A Equivalence regarding IBD-sharing Structures

Vaguely speaking equivalent IBD-sharing structures means that these structures correspond to structurally similar (exchangeable) genetic inheritance. Consider two structures, A and B say. In its simplest form this above property correspond to cases where A is transformed to B if the inheritances of two individuals, of similar pedigree structure positions, are switched (permuted).

Example 9 (Equivalent AST-structures) *Assume an affected sib-trio (AST) pedigree as shown in Figure 2. Order the three affected siblings as Sibling 1, 2 and 3 respectively. Quite naturally, let IBD(i, j) = k mean that the ith and jth sibling share k alleles IBD.*[21]

If $IBD(1, 2) = 2$ *and* $IBD(1, 3) = 0$*, implicitly* $IBD(2, 3) = 0$.[22] *Now, permute Sibling 1 and Sibling 3. This leads to* $IBD(1, 2) = 0$ *and* $IBD(1, 3) = 0$*, and* $IBD(2, 3) = 2$*. These two cases, i.e. the corresponding inheritance vectors, are assumed to produce equivalent IBD-sharing structures.*

B Criterion regarding unstandardized score function equivalence

Formulating criterions for producing equivalent (equivalence classes of) score functions turns out to be most straightforwardly achieved using the general score function definition (3).

When $n = 2$, given an underlying score distribution under H_0, there is only one type of standardized score function definition, i.e. regardless of the instantiation of the unstandardized scores, s_1 and s_2, one ends up with the same standardized function. Explicitly, all score functions are equivalent.

In the same manner, when $n = 3$, for each value (constant) $c \in [0, \infty]$ the relation

$$s_3 - s_2 = c(s_2 - s_1) \qquad (18)$$

defines a single specific equivalence class.[23] The meaning of this is that all unstandardized score functions fulfilling the above relation with the same c are forced to become equivalent.

Example 10 (Equivalence class of S_{pairs} and S_{all}) *For an ASP, as noted above,* S_{pairs} *and* S_{all} *are equivalent. The whole corresponding equivalence class, including these two instances, are defined through using* $c = 1$ *in (18). This class is constituted by all symmetric score functions, i.e. such functions obeying* $s_3 - s_2 = s_2 - s_1$.

Finally consider an arbitrary $n > 3$. Here, each ordered vector of constants $(c_1, c_2, \ldots, c_{n-1})$ where all $c_k \in [0, 1]$, with natural restricting constraint $\sum_{k=1}^{n-1} c_k = 1$, and the vector moreover fulfills

$$\left[\frac{s_{k+1} - s_k}{s_n - s_1} \right] = c_k; \;\; k = 1, 2, \ldots, n-1,$$

then defines an equivalence class of (unstandardized) score functions in the same sense as above.

[21] Formally, the range of indices are $i, j = 1,2,3$ and $k = 0,1,2$.

[22] Since 1 and 2 have ancestrally the same (founder) genotypes according to full IBD-sharing.

[23] The extreme cases $c = 0$ and $c = \infty$ in (18) calls for specific interpretations. In the former case $s_3 = s_2$ and in the latter case $s_2 = s_1$. In both cases, as always, the specific numerical scores are then defined through the properties of standardization.