Mathematical Statistics
Stockholm University

Strategies for Conditional Two-Locus Nonparametric Linkage Analysis

Lars Ängquist, Ola Hössjer, Leif Groop

Research Report 2006:10

ISSN 1650-0377

Postal address:

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

Internet:

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

Mathematical Statistics
Stockholm University
Research Report **2006:10**,
http://www.math.su.se/matstat

Strategies for Conditional Two-Locus Nonparametric Linkage Analysis

Lars Ängquist * Ola Hössjer † Leif Groop ‡

December 2006

Abstract

In this article we deal with two-locus nonparametric linkage (NPL) analysis, mainly in the context of conditional analysis. This means that one incorporates single-locus analysis information through conditioning when performing a two-locus analysis. Here we describe different strategies for using this approach

In Cox et al. (1999) they implemented this as follows: (i) Calculate the one-locus NPL process over the included genome region(s). (ii) Weight the individual pedigree NPL scores using a weighting function depending on the NPL scores for the corresponding pedigrees at specific conditioning loci. We generalize this by conditioning with respect to the inheritance vector rather than the NPL score and by separating between the case of known (predefined) and unknown (estimated) conditioning loci. In the latter case we choose conditioning locus, or loci, according to predefined criterions. The most general approach results in a random number of selected loci, depending on the results from the previous one-locus analysis.

Major topics in this article include discussions on optimal score functions with respect to the noncentrality parameter (NCP), and how to calculate adequate p-values and perform power calculations. We also discuss issues related to multiple tests which arise from the two-step

*Postal address: Mathematical Statistics, Lund University, Box 118, SE-221 00, Sweden. E-mail: Lars.Angquist@matstat.lu.se. Financial support from the Swedish Research Council.

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

‡Postal address: Department of Clinical Sciences, Diabetes and Endocrinology, Lund University, Malmö, SE-205 02, Sweden. E-mail: leif.groop@endo.mas.lu.se.

procedure with several conditioning loci as well as from the genome-wide tests.

KEY WORDS: Nonparametric linkage analysis, two-locus linkage analysis, conditional linkage analysis, score functions, conditioning loci, two-step procedure, noncentrality parameter, genome-wide significance and power calculations, ROC curves, Monte Carlo simulation.

Contents

1 Introduction

In this article we discuss genome-wide two-locus gene mapping in the context of conditional nonparametric linkage (NPL) analysis. Two locus analysis in general is motivated by the possibility to model and take advantage of gene-gene interaction. Conditional NPL analysis is motivated by the decreased computational complexity and multiple testing compared to unconditional two-locus analysis.

Here we discuss conditional procedures based on both known (predefined) and unknown (estimated) conditioning loci. In the latter case this leads to a two-step procedure: (i) Perform a single-locus genome scan. (ii) Use these results when choosing conditioning loci and perform a conditional two-locus analysis. Further, an optimal score function with respect to the noncentrality parameter (NCP), i.e. the expected NPL score under the alternative hypothesis H_1 for the underlying genetic disease model, is discussed. It turns out that this function might be defined in terms of the corresponding inheritance vector distribution. Moreover, significance calculations under both null and alternative hypotheses are included. For instance, the possibility of a multiple-step procedure, using several conditioning loci and score functions implies that the significance calculations are complicated by the possibly severe multiple testing.

Cox et al. (1999) suggested a conditional approach and defined a multiplicative two-locus score function aimed at detecting *epistasis* or *heterogeneity* when the two loci are nonsyntenic, i.e. located on different chromosomes. This approach was applied, for instance, by Schulze et al. (2004). A related conditional approach based on generalized estimating equations was described by Liang et al. (2001) and Chiu and Liang (2004).

In Dupuis et al. (1995) a two-locus score statistic was developed and used for both simultaneous and conditional search. Later, Tang and Siegmund (2002) discussed conditional search with respect to quantitative traits. Two-locus or multi-locus analysis in the form of a maximum lod score (MLS) approach has been discussed by Cordell et al. (1995), Farrall (1997) and Cordell et al. (2000). The two-marker mean test is outlined in Knapp et al. (1994) and Strauch et al. (2000) extended the score functions S_{pairs} and S_{all} to the two-locus case, through summation of one-locus scores, and implemented this approach into the program *GENEHUNTER-TWOLOCUS*. Recently, a conditional regression-based approach for affected sib-pairs in the context of syntenic disease loci is discussed by Barber et al. (2006).

Li and Reich (2000) presented a complete list of distinct two-locus disease models for binary phenotype-fully penetrant-biallelic disease loci. Further, discussion on two-locus disease models and gene-gene interaction may be found in MacLean et al. (1993), Kämpe (2001), Bengtsson (2001), Holmans (2002) and Ängquist et al. (2005).

Examples of review articles on approaches to two-locus and multi-locus analysis are Strauch et al. (2003) and Hoh and Ott (2003).

In *Section 2* we present some basic theory and introduce basic notation, including a description of two-locus genetic disease models. *Section 3* is devoted to the concept of nonparametric linkage analysis in the form of one-locus analysis, general (unconditional) two-locus analysis and various versions of conditional two-locus analysis. Subsequently, in *Section 4* the attention is brought to the concept of noncentrality parameters and its implication in the form of NCP-optimal score functions. In *Section 5* power calculations are performed and a concluding discussion is given in *Section 6*. Finally, technical details are referred to the *Appendices*.

2 Two-Locus Genetic Disease Models

Each NPL investigation is done in relation to a predefined autosomal genomic region Ω.[1] If C is the total number of chromosomes we define $\Omega = \cup_{i=1}^{C} c_i$, where c_i is the i^{th} chromosome which is of genetic map length $|c_i|$. The total length of Ω is therefore $|\Omega| = \sum_{i=1}^{C} |c_i|$. Given a specific locus x the corresponding chromosome where it is located is denoted $c(x)$.

We assume bi-allelic disease loci, i.e. only the disease and normal allelic variants, D and d respectively, are possible.

2.1 Definition

A two-locus genetic model consists of three parts:

(i) The disease allele-frequencies $p_1 = P(D_1)$ and $p_2 = P(D_2)$, where p_i is the probability of the disease allelic variant with respect to the i^{th} disease locus (i=1 or 2).

(ii) The penetrance matrix,

$$f = \begin{pmatrix} f_{00} & f_{01} & f_{02} \\ f_{10} & f_{11} & f_{12} \\ f_{20} & f_{21} & f_{22} \end{pmatrix}, \tag{1}$$

where f_{ij} refers to the probability of being affected if the corresponding genotypes contain i and j copies of D_1 and D_2 respectively.

[1] This is subsequently referred to as our *genome*.

(iii) The two disease loci, l_1 and l_2. Usually, they are assumed to be located on different chromosomes, i.e. $c(l_1) \neq c(l_2)$.

Oftenly, one assumes $f_{0j} \leq f_{1j} \leq f_{2j}$ and $f_{i0} \leq f_{i1} \leq f_{i2}$. If $f_{2j} = 1$ or $f_{i2} = 1$ we speak of a marginally fully penetrant model-structure and if $f_{0j} > 0$ or $f_{i0} > 0$ so called phenocopies[2] are allowed for marginally.

2.2 Examples

We will briefly describe four classes of two-locus genetic disease models adopting the notation of $g = (g_0, g_1, g_2)$ and $h = (h_0, h_1, h_2)$ for the marginal nonstandardized penetrance vectors of the 1^{st} and 2^{nd} disease locus:

(i) If the one-locus penetrances are combined in an additive fashion,

$$f_{ij} = g_i + h_j; \quad 0 \leq i, j \leq 2, \tag{2}$$

we speak of an *additive* two-locus genetic disease model.

(ii) If the two-locus penetrances instead are defined through products,

$$f_{ij} = g_i h_j; \quad 0 \leq i, j \leq 2, \tag{3}$$

we speak of a *multiplicative* two-locus genetic disease model.

(iii) If we consider a dual version of the multiplicative model where,

$$\forall i, j: \ (1 - f_{ij}) = (1 - g_i)(1 - h_j),$$

we get,

$$f_{ij} = g_i + h_j - g_i h_j; \quad 0 \leq i, j \leq 2, \tag{4}$$

which is a *heterogeneity* two-locus genetic disease model.

(iv) If the two-locus penetrances depends only on the total number of disease alleles,

$$f_{ij} = k_{i+j} \Rightarrow f = \begin{pmatrix} k_0 & k_1 & k_2 \\ k_1 & k_2 & k_3 \\ k_2 & k_3 & k_4 \end{pmatrix}, \tag{5}$$

for some function k_l, we speak of a *threshold* two-locus genetic disease model. For instance, if it is both sufficient and necessary to have at least two disease alleles to become affected, $k_0 = k_1 = 0$ and $k_2 = k_3 = k_4 = 1$.

[2] A *phenocopy* is an affected individual with 0 disease alleles present, i.e. with genotype dd at disease locus l.

3 Nonparametric Linkage Analysis

The basic tool of nonparametric linkage analysis is the *score function*, which quantifies compatibility between phenotype similarity and the inheritance pattern within a pedigree. For specialized discussions on this topic we refer to Whittemore and Halpern (1994), McPeek (1999), Sengul et al. (2001), Hössjer (2003), Lange and Lange (2004), Hössjer (2005b, 2005c) and Ängquist (2006).

3.1 One-Locus Analysis

Assume a pedigree set consisting of N pedigrees. The inheritance pattern of the k^{th} pedigree at locus x is determined by means of the *inheritance vector*,

$$v_k(x) = \Big(p_1(x), m_1(x), p_2(x), m_2(x), \ldots, p_{(n_k-f_k)}(x), m_{(n_k-f_k)}(x)\Big), \qquad (6)$$

see Donnelly (1983). In (6) n_k is the number of individuals, f_k the number of founders and $(n_k - f_k)$ the number of nonfounders of Pedigree k. Moreover, $p_i(x)$ and $m_i(x)$ equal 0 if the i^{th} nonfounder's paternal and maternal allele respectively, at locus x, originate from a grandfather and 1 if they originate from a grandmother. The number of vector positions equals the number of meioses $m_k = 2(n_k - f_k)$.

Define the *pedigree-specific NPL score* for the k^{th} pedigree at locus x as,

$$Z_k(x) = E\left[S\Big(v_k(x)\Big)\right] = \sum_{w \in \mathbb{V}_k} P_{v_k(x)}(w) S(w), \qquad (7)$$

where $P_{v_k(x)}(w) = P(v_k(x) = w \mid \text{MD})$ is the conditional probability of $v_k(x)$ given marker data MD, $\mathbb{V}_k$ is the full set of 2^{m_k} inheritance vectors and S is the one-locus score function for Pedigree k.

We assume that the score function is apriori standardized through,

$$S(w) \leftarrow \left(\frac{S(w) - E_{H_0}(S)}{\sigma_{H_0}(S)}\right),$$

where E_{H_0} and σ_{H_0} correspond to the expected value and standard deviation of S under the null hypothesis H_0, i.e. given that w is uniformly distributed over $\mathbb{V}_k$. From this follows,

$$\sum_{w \in \mathbb{V}_k} S_k(w) = 0 \quad \text{and} \quad 2^{-m_k} \sum_{w \in \mathbb{V}_k} S_k^2(w) = 1. \qquad (8)$$

The *NPL score* for the pedigree set (Kruglyak et al., 1996) is then a weighted linear combination,

$$Z(x) = \sum_{k=1}^{N} \gamma_k Z_k(x), \tag{9}$$

of the pedigree-specific NPL scores with weights γ_k satisfying,

$$\sum_{k=1}^{N} \gamma_k^2 = 1. \tag{10}$$

The weights may be chosen to depend on pedigree structure, size and phenotypes. A slightly different total NPL score, which differs from (9) for incomplete marker data, was introduced by Kong and Cox (1997).

Using the maximum of the *NPL score process* along Ω,

$$Z_{\max} = \sup_{x \in \Omega} Z(x), \tag{11}$$

as a test statistic makes it possible to test for the presence of any disease locus. The natural test hypotheses are,

$$\begin{cases} H_0: \text{ No disease locus on } \Omega, \\ H_1: \text{ At least one disease locus on } \Omega, \end{cases} \tag{12}$$

which leads to the genome-wide significance level,

$$\alpha(z) = P_{H_0}(Z_{\max} \geq z), \tag{13}$$

and power,

$$\beta(z) = P_{H_1}(Z_{\max} \geq z), \tag{14}$$

for a test that rejects H_0 when $Z_{\max} \geq z$.

3.2 Two-Locus Analysis

In the two-locus case we define an *unconditional* two-locus score function $S_k(w_1, w_2)$ for the k^{th} pedigree, which depends on inheritance vectors $w_1, w_2 \in \mathbb{V}_k$. In analogy with the one-locus case above, the scores are normalized using a two-locus generalization of (8) leading to,

$$\sum_{w_1, w_2} S_k(w_1, w_2) = 0 \quad \text{and} \quad 2^{-2m_k} \sum_{w_1, w_2} S_k^2(w_1, w_2) = 1. \tag{15}$$

Moreover, the pedigree-specific NPL score (7) and the pedigree set NPL score (9) are now generalized, being defined with respect to pairs of loci $(x, y) \in \Omega$, as

$$Z_k(x, y) = \sum_{w_1, w_2} P_{v_k(x,y)}(w_1, w_2) S_k(w_1, w_2) \tag{16}$$

and

$$Z(x,y) = \sum_{k=1}^{N} \gamma_k Z_k(x,y); \quad c(x) \neq c(y), \tag{17}$$

where γ_k are pedigree weights satisfying (10). As in the one-locus case, they may depend on pedigree structure and phenotypes. Further, $P_{v_k(x,y)}(w_1, w_2) = P(v_k(x) = w_1, v_k(y) = w_2|\text{MD})$. Since $c(x) \neq c(y)$ in (17) and inheritance at unlinked loci is independent, it follows that $P_{v_k(x,y)}(w_1, w_2) = P_{v_k(x)}(w_1)P_{v_k(y)}(w_2)$.

Using the null hypothesis in (12) we may, in analogy with (11) and (13), present the two-locus maximum NPL score and genome-wide significance level and power as,

$$Z_{\max,\text{tl}} = \sup_{\substack{x,y\in\Omega \\ c(x)\neq c(y)}} Z(x,y), \tag{18}$$

$$\alpha_{\text{tl}}(z) = P_{H_0}(Z_{\max,\text{tl}} \geq z), \tag{19}$$

and

$$\beta_{\text{tl}}(z) = P_{H_1}(Z_{\max,\text{tl}} \geq z), \tag{20}$$

where 'tl' is an abbreviation for 'two-locus'.

3.3 Conditional Two-Locus Analysis: Known Conditioning Locus

If we fix a *conditioning locus* y on chromosome $c(y)$ and use (16)-(17) with x varying through $\Omega_{c(y)} = \Omega \backslash c(y)$, we have a *conditional two-locus NPL analysis*.

This procedure leads to a more complicated and involved version of the normalization procedure in (15). Firstly, for the k^{th} pedigree, the centering is performed using the conditional constraints,

$$\sum_{w_1} S_k(w_1, w_2) = 0 \quad (\forall w_2), \tag{21}$$

where w_2 is associated with the conditioning locus y. Defining $\bar{S}_k^2(w_2) = 2^{-m_k} \sum_{w_1} S_k^2(w_1, w_2)$ we get, instead of (10), the constraint,

$$\sum_{k=1}^{N} \gamma_k^2 \sum_{w_2} P_{v_k(y)}(w_2)\ \bar{S}_k^2(w_2) = 1, \tag{22}$$

for the pedigree weights γ_k. A reasonable approach is to set γ_k identical for pedigrees with equal structure and phenotypes. The conditional variance $\bar{S}_k^2(w_2)$ quantifies how variable S_k is at the first locus given inheritance vector w_2 at the second locus. Notice that $P_{v_k(y)}(w_2)$ allows for imperfect data at y.

Example 1 Assuming perfect marker data, the conditional procedure described in Cox et al. (1999) may be recognized as a special case of the general approach. Since $S_k\left[v_k(x)\right]$ and $S_k\left[v_k(y)\right]$ are both known for perfect marker data, putting,

$$S_k(w_1, w_2) = S_k(w_1) f\left[S_k(w_2)\right], \tag{23}$$

we may rewrite (17) as,

$$Z(x, y) = \sum_{k=1}^{N} \gamma_k Z_k(x) f\Big[Z_k(y)\Big], \tag{24}$$

with normalization $\sum_{k=1}^{N} \gamma_k^2 f\Big[Z_k(y)\Big]^2 = 1$. This is a multiplicative two-locus score which is analogous to a one-locus NPL score process along $\Omega_{c(y)}$, weighting pedigrees according to a combination of pedigree-specific weights γ_k and a function of one-locus NPL scores at locus y.[3]

Remark 1 The function f in (24) may be defined in various ways (see e.g. Cox et al., 1999; Ängquist, 2001),

$$\begin{array}{ll} 1.\ f^z_{\text{prop}+}(Z) = |Z|\ I(Z \geq z), & 2.\ f^z_{\text{prop}-}(Z) = |Z|\ I(Z \leq z), \\ 3.\ f^z_{\text{epi}}(Z) = I(Z \geq z), & 4.\ f^z_{\text{het}}(Z) = I(Z \leq z), \\ 5.\ f^z_{\text{dist}+}(Z) = I\left[F(Z) \geq z\right], & 6.\ f^z_{\text{dist}-}(Z) = I\left[F(Z) \leq z\right], \end{array} \tag{25}$$

where $I(A)$ is the indicator function of the event A, F is the null distribution function of the pedigree-specific NPL score and z is a predefined threshold. Since Functions 5-6 in (25) take the whole distribution of Z into account, the threshold z is comparable for pedigrees of various form and structure. For homogeneous pedigree sets, i.e. when all pedigree structures and phenotypic settings are equal, this is equivalent to using Functions 3-4.

The maximum conditional two-locus NPL score is,

$$Z_{\max,y} = \sup_{x \in \Omega_{c(y)}} Z(x, y). \tag{26}$$

To be able to later define significance levels we split our null hypothesis (12) into two parts as

$$\begin{array}{ll} H_0^{c(y)}: & \text{No disease locus on Chromosome } c(y), \\ \bar{H}_0^{c(y)}: & \text{No disease locus outside Chromosome } c(y). \end{array} \tag{27}$$

The conditional significance level, given marker data $\text{MD}_{c(y)}$ on Chromosome $c(y)$, is

$$\alpha_y(z) = P_{\bar{H}_0^{c(y)}}(Z_{\max,y} \geq z | \text{MD}_{c(y)}), \tag{28}$$

[3]For imperfect data, (24) is not directly equivalent to using (17) with $S_k(w_1, w_2) = S_k(w_1) f\left[S_k(w_2)\right]$.

and the conditional power

$$\beta_y(z) = P_{\bar{H}_1^{c(y)}}(Z_{\max,y} \geq z | \mathrm{MD}_{c(y)}), \tag{29}$$

where $\bar{H}_1^{c(y)}$ is the alternative hypothesis,

$$\bar{H}_1^{c(y)}: \quad \text{At least one disease locus outside Chromosome } c(y), \tag{30}$$

corresponding to the lower part of (27).

In (27), we have replaced H_0 by the less restrictive null hypothesis $\bar{H}_0^{c(y)}$. In this setting y is allowed to be, or being linked to, a disease locus. The underlying assumption for this argument can be formalized as:

Assumption 1 *(i) $MD_{c(y)}$ is independent of phenotypes under $H_0^{c(y)}$.*

(ii) $MD_{c(y_1)}, MD_{c(y_2)}, \ldots, MD_{c(y_k)}$ are conditionally independent given phenotypes if $k \geq 2$, all $c(y_j)$ are different, and at most one $H_0^{c(y_j)}$ is not valid.

As a consequence, $\alpha_y(z)$ in (28) can be calculated in the same way as a one-locus significance level $\alpha(z)$ along $\Omega_{c(y)}$.

3.4 Conditional Two-Locus Analysis: Unknown Conditioning Loci

When we do not have, or assume, explicit knowledge of an obvious conditioning locus, such as a known disease locus, we may randomly select interesting loci according to some predefined one-locus criterion. This is the motivation for the two-step procedure described below.

Since we possibly deal with multiple conditioning loci, we replace the less restrictive null hypothesis (27) by (12).

Selecting Conditioning Loci

Define the chromosome-wise NPL score maximum,

$$Z_{\max,c} = \sup_{x \in c} Z(x),$$

and the corresponding chromosomal significance level,

$$\alpha_c(z) = P_{H_0^c}(Z_{\max,c} \geq z),$$

where H_0^c is the upper part of (27). A general one-locus NPL score-dependent selection criterion for conditioning chromosomes may be formulated as

$$\mathbb{C} = \{c \ ; \ \ Z_{\max,c} \geq z_c\}, \tag{31}$$

where z_c is a given threshold for Chromosome c.

Further, denote the random positions of the chromosome-wise NPL score maximums as

$$y_c = \arg\max_{x \in c} Z(x). \tag{32}$$

Using (31)-(32) the conditioning loci are selected as,

$$\mathbb{Y} = \{y_c \ ; \ \ c \in \mathbb{C}\}, \tag{33}$$

which guarantees that they are all located on different chromosomes.[4]

Example 2 Two possible choices of z_c in (31) are (i) equal thresholds $z_c = z$ and (ii) genetic length-dependent thresholds, e.g. $z_c = \alpha_c^{-1}(\delta)$. The tuning constants z and δ reflect the number of conditioning loci that the investigator is willing to use. Their choice is a compromise between finding true interactions on one hand and avoiding severe multiple testing on the other hand.

Remark 2 If there is prior evidence that a disease locus exists somewhere along Chromosome c one may use,

$$\mathbb{Y} = \arg\max_{x \in c} Z(x), \tag{34}$$

and H_0 may in this case be replaced by the weaker $\bar{H}_0^c$.

Combining Conditional NPL Scores

The most straightforward generalization of (26) to several conditioning loci is to consider a test statistic,

$$Z_{\max,\mathbb{Y}} = \max_{y \in \mathbb{Y}} \max_{x \in \Omega_{c(y)}} Z(x,y) = \max_{c \in \mathbb{C}} Z_{\max,y_c}. \tag{35}$$

However, we will use a more refined approach based on conditional two-locus p-values,

$$p_c = \alpha_{y_c}(Z_{\max,y_c}), \tag{36}$$

[4] This restriction may be relaxed, requiring only a certain minimum map distance L between all pairs of syntenic conditioning loci.

with corresponding set of conditioning loci $\mathbb{Y}$ selected through (31). Instead of (35), we then use the minimum p-value,

$$p_{\min} = \begin{cases} \min_{c \in \mathbb{C}} p_c & \text{if } \mathbb{C} \neq \emptyset, \\ 1 & \text{if } \mathbb{C} = \emptyset, \end{cases} \tag{37}$$

as test statistic and reject H_0 whenever $p_{\min}$ is smaller than or equal to a given threshold u. As opposed to (35), (37) takes into account varying chromosome lengths and the inheritance vectors at $y \in \mathbb{Y}$.

Using (37) we get a genome-wide, or global, significance level

$$\alpha_{\mathbb{Y}}(u) = P_{H_0}(p_{\min} \leq u), \tag{38}$$

and power

$$\beta_{\mathbb{Y}}(u) = P_{H_1}(p_{\min} \leq u). \tag{39}$$

The probability of including a conditioning locus from Chromosome c in $\mathbb{Y}$ is $\lambda_c = \alpha_c(z_c)$. Since each p_c has a uniform distribution on $(0, 1)$,[5] a simple Bonferroni upper bound for the significance level is,

$$\alpha_{\mathbb{Y}}(u) \leq (\sum_{c=1}^{C} \lambda_c) u.$$

Hence, $\sum_c \lambda_c$ can be viewed as a crude measure of the *effective number* of conditioning loci used. In particular, if $z_c = \alpha_c^{-1}(\delta)$, the effective number of of conditioning loci is $C\delta$.

4 Noncentrality Parameters

A quantity of great importance is the *noncentrality parameter (NCP)* which measures the expected NPL score, at the disease locus or loci, under an alternative hypothesis. For one-, two- and conditional two-locus NPL scores we define,

$$\begin{aligned} \text{NCP}_{l_1} &= E_{H_1}\Big[Z(l)\Big], \\ \text{NCP}_{l_1,l_2} &= E_{H_1}\Big[Z(l_1,l_2)\Big], \\ \text{NCP}_{l_1|l_2} &= E_{H_1}\Big[Z(l_1,l_2)|\text{MD}_{c(l_2)}\Big]. \end{aligned} \tag{40}$$

Notice that the first two quantities in (40) are constants, whereas the third is a random variable since we condition on marker data from Chromosome $c(l_2)$.

[5] Ignoring discreteness effects of the null distribution of $Z_{\max,c}$.

Assume a homogeneous pedigree set consisting of N pedigrees, perfect data, and equal pedigree weights ($\gamma_k = 1/\sqrt{N}$). Then

$$\begin{aligned} \mathrm{NCP}_{l_1} &= A\ \sqrt{N}, \\ \mathrm{NCP}_{l_1,l_2} &= B\ \sqrt{N}, \\ \frac{1}{\sqrt{N}}\ \mathrm{NCP}_{l_1|l_2} &\xrightarrow{\mathbb{P}} D \quad \text{as } N \to \infty, \end{aligned} \tag{41}$$

where A and B are the NCPs for a single pedigree and $\xrightarrow{\mathbb{P}}$ denotes convergence in probability. A natural interpretation of D is an average NCP per family for conditional two-locus analysis.

To derive expressions for the NCPs in (41) we define the joint inheritance vector distribution at the two disease loci,

$$P(w_1, w_2) = P\Big(v(l_1) = w_1, v(l_2) = w_2 | Y, H_1\Big), \tag{42}$$

with corresponding marginal distributions, $P_1(w_1)$ and $P_2(w_2)$, and common phenotype vector Y. Calculation of this distribution is outlined in Appendix 6.3. The following theorem presents maximal NCPs and corresponding optimal score functions. A proof is given in Appendix 6.3.

Theorem 1 *For a homogeneous pedigree set, the maximum NCPs are*

$$\begin{aligned} A^2 &= 2^m \sum_{w_1} P_1^2(w_1) - 1, \\ B^2 &= 2^{2m} \sum_{w_1,w_2} P^2(w_1, w_2) - 1, \\ D^2 &= 2^m \sum_{w_1,w_2} \frac{P^2(w_1, w_2)}{P_2(w_2)} - 1. \end{aligned} \tag{43}$$

The maxima in (43) are attained for score functions,

$$\begin{aligned} S(w_1) &\propto P_1(w_1) - 2^{-m}, \\ S(w_1, w_2) &\propto P(w_1, w_2) - 2^{-2m}, \\ S(w_1, w_2) &\propto P(w_1|w_2) - 2^{-m}. \end{aligned} \tag{44}$$

■

5 Results

Throughout this section, we will use homogeneous pedigree sets taken from Figure 1.

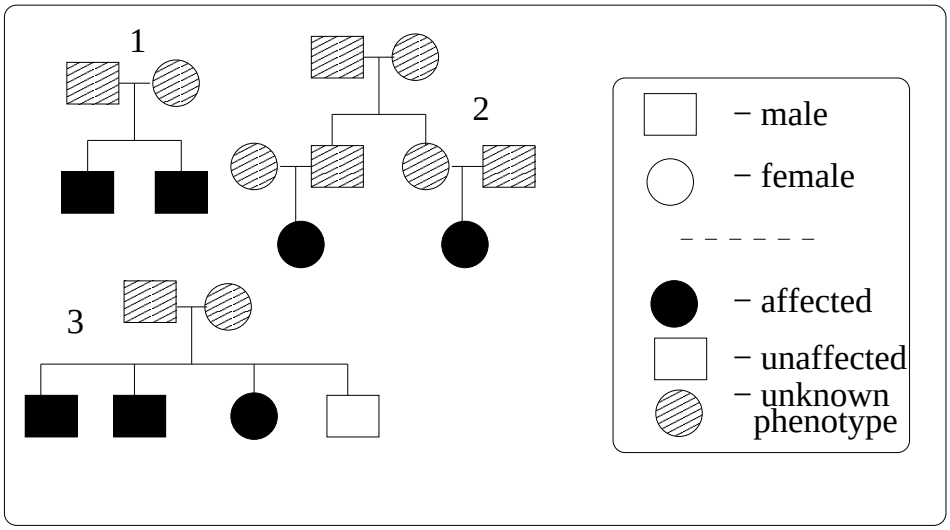

Figure 1: The pedigrees used in NCP and power calculations.

5.1 NCP Calculations

We calculate maximal NCP parameters in (43) for several distinct genetic models under the constraint of a constant disease prevalence K. For more details on NCP calculations, see Appendix 6.3. In Figure 2-5 all the results are displayed for various types of disease models.

In all simulations we report NCP as function of the *displacement*,

$$d = \max_{i,j} f_{ij} - \min_{i,j} f_{ij} = f_{22} - f_{00},$$

which quantifies the strength of the genetic model. We have $d = d(x)$, where x is the penetrance parameter defined in Appendix 6.3.

From Figures 2-5, we notice that:: (i) In all cases, $B \geq D \geq A$ in (43). (ii) For multiplicative models, as expected, $A{=}D$. (iii) For additive and heterogeneity models, D is almost consistently larger than A, though much closer to A than B. (iv) For threshold models, D tends to be closer to B than A.

This implies that among our set of models the threshold-type (followed by the heterogeneity-type) seems to be most suitable for conditional two-locus analyses, see Section 6.3 for further discussion.

To explore the influence of allele frequencies, we perform in Figure 6 conditional two-locus maximum NCP calculations with respect to Pedigree 1 under 10 different choices of $p = p_1 = p_2$.

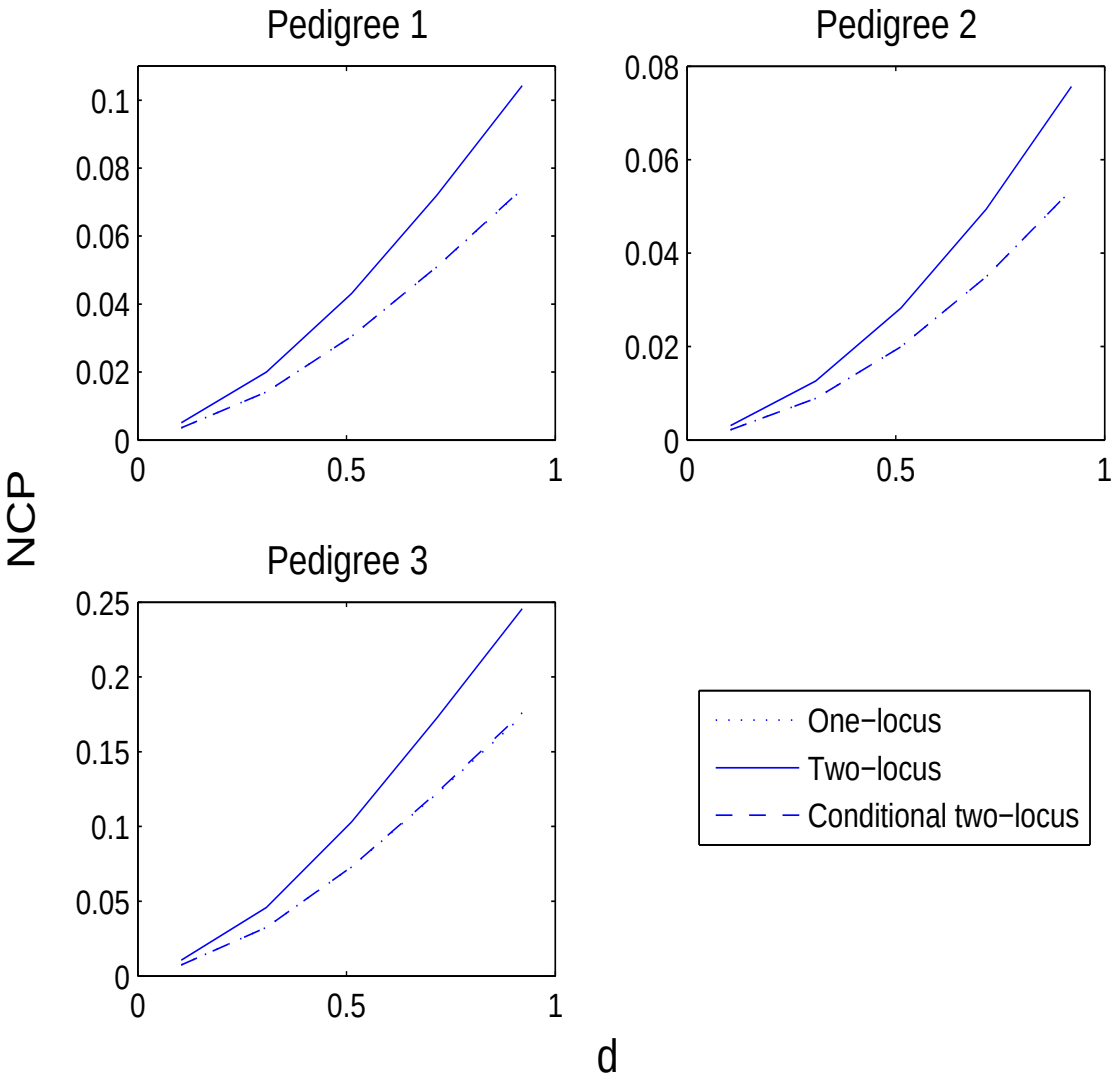

Figure 2: Maximum one-, two- and conditional two-locus NCP calculations using Pedigrees 1-3, equal disease allele frequency at both disease loci ($p = p_1 = p_2 = 0.02$), a constant disease prevalence K=0.1 and symmetric *additive* two-locus disease penetrance models (f^1) and displacement $d = d(x)$ with $x = 0 : 0.05 : 0.2$. $N = 1$, i.e. NCP= A, B and D in (41) respectively.

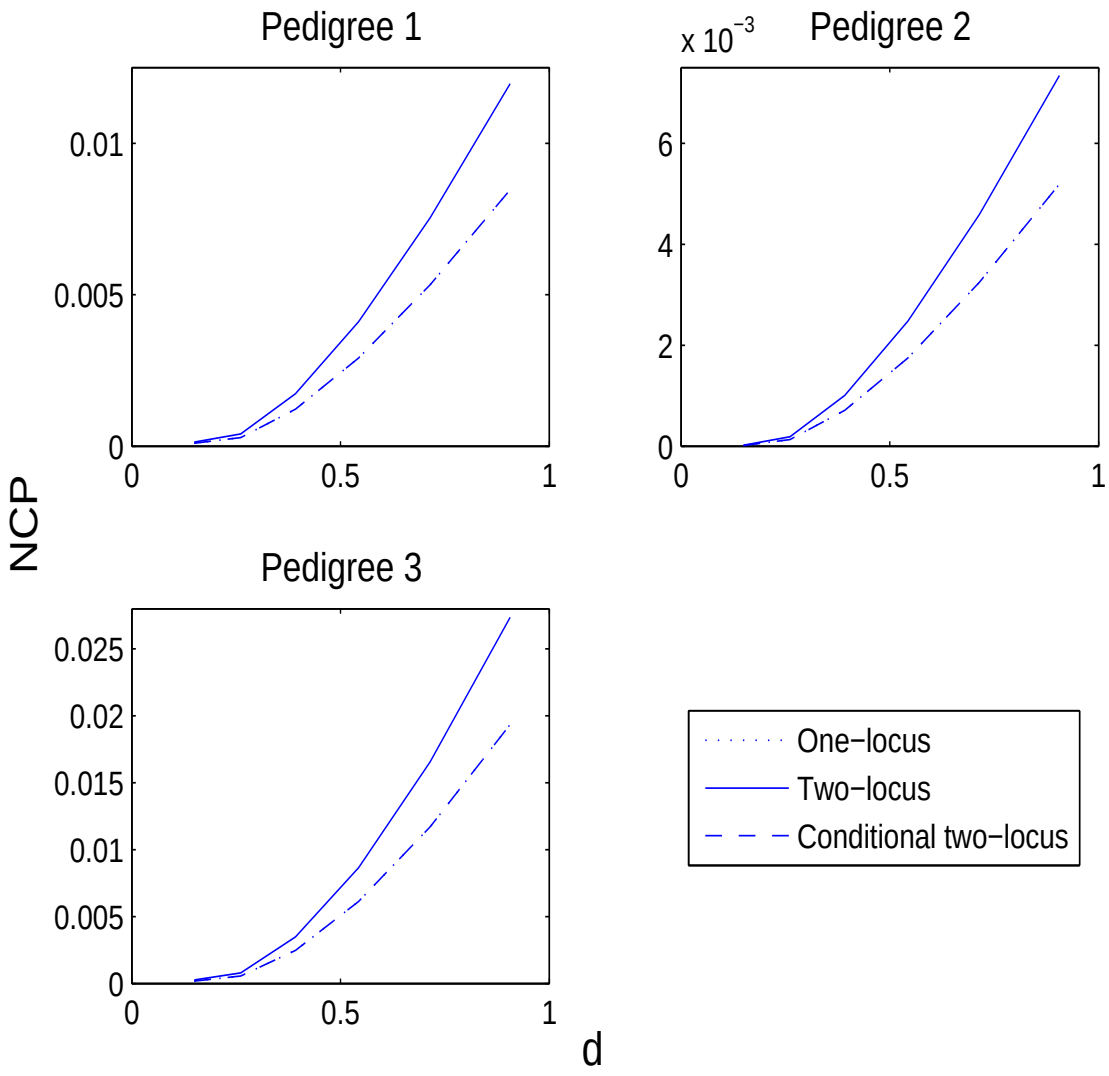

Figure 3: Maximum one-, two- and conditional two-locus NCP calculations using symmetric *multiplicative* two-locus disease penetrance models (f^2) and displacement $d = d(x)$ with $x = 0.20 : 0.05 : 0.45$. For details, see Figure 2.

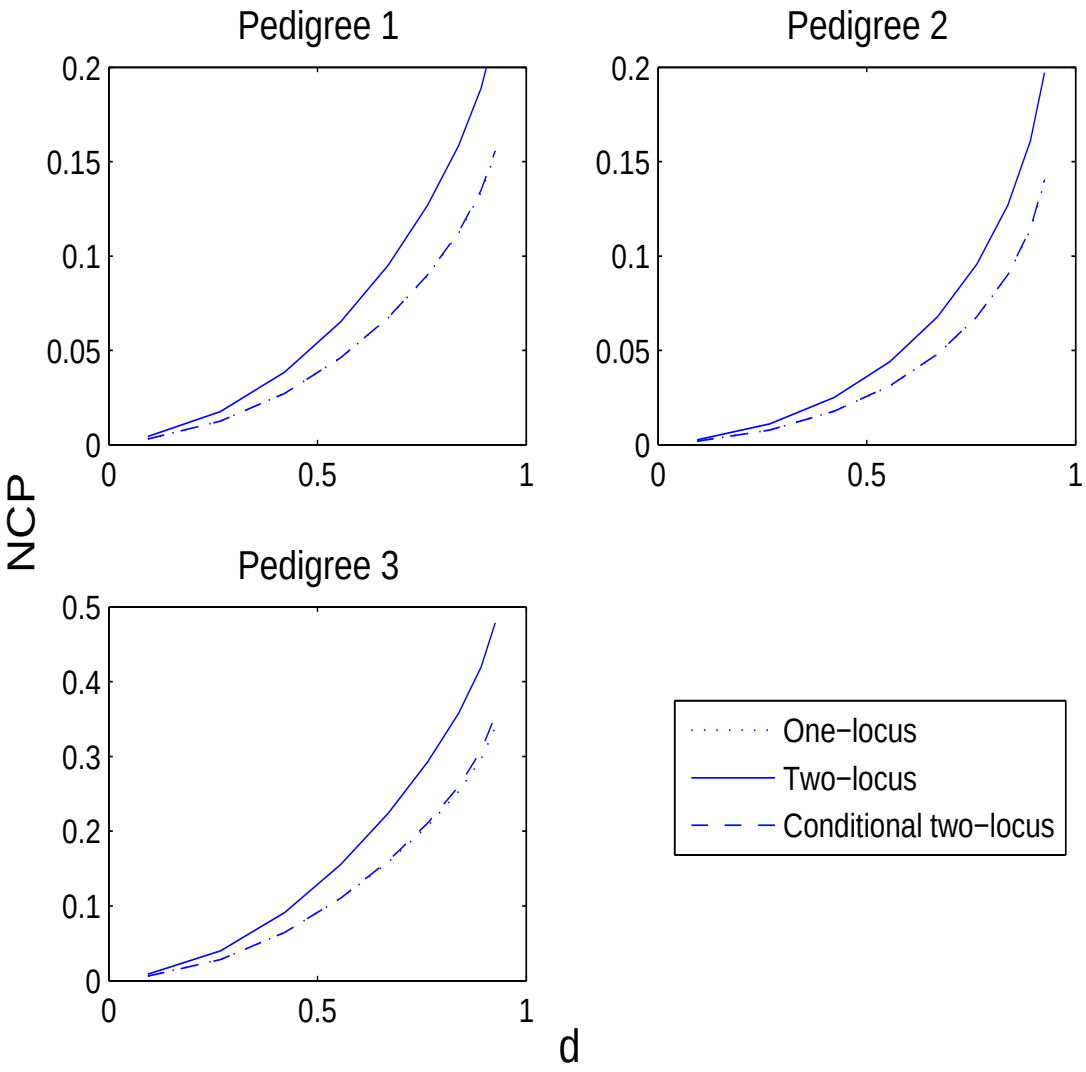

Figure 4: Maximum one-, two- and conditional two-locus NCP calculations using symmetric *heterogeneity* two-locus disease penetrance models (f^3) and displacement $d = d(x)$ with $x = 0 : 0.05 : 0.4$. For details, see Figure 2.

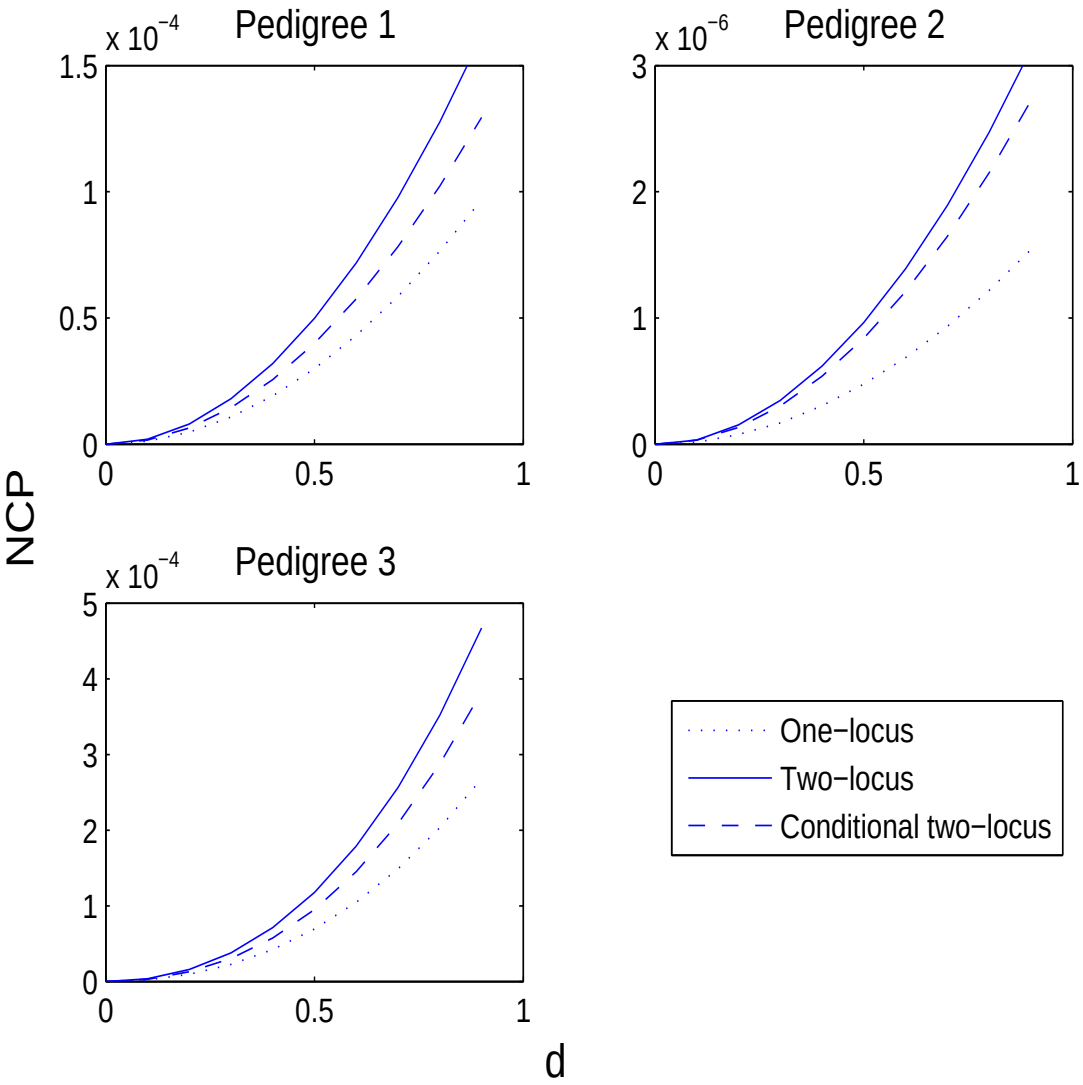

Figure 5: Maximum one-, two- and conditional two-locus NCP calculations using symmetric *threshold* two-locus disease penetrance models (f^4) displacement $d = d(x)$ with $x = 0 : 0.05 : 0.45$. For details, see Figure 2.

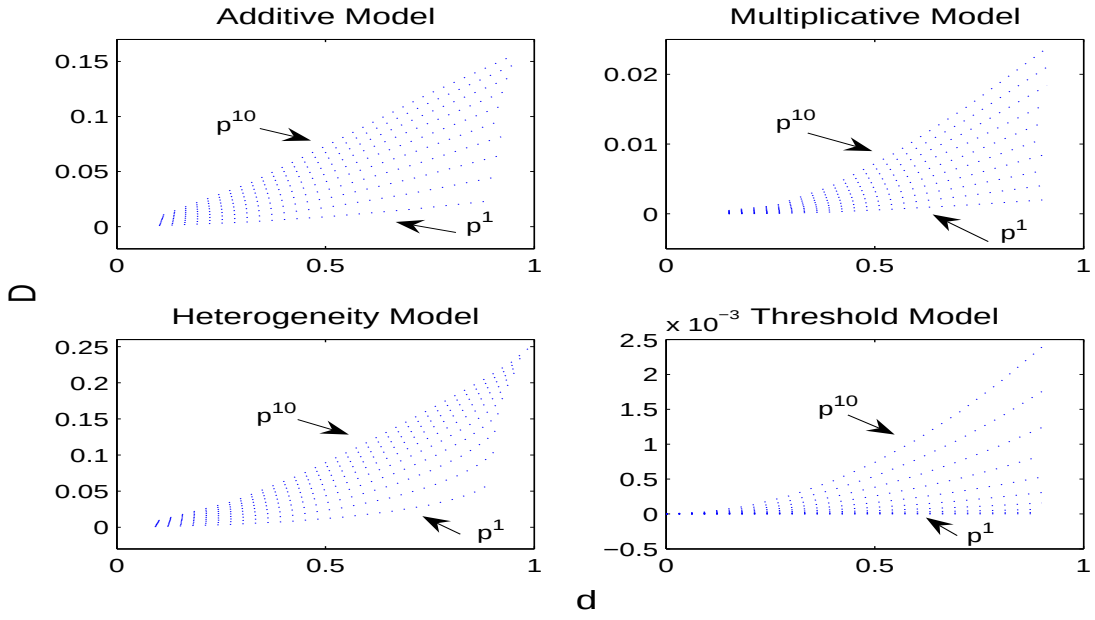

Figure 6: Maximum conditional two-locus asymptotic NCP for Pedigree 1, assuming $N = 1$, a fixed conditioning locus l_2, $p = p_1 = p_2$ with $(p^1, p^2, \ldots, p^{10}) = (0.005, 0.010, \ldots, 0.050)$, a constant disease prevalence $K = 0.1$, four distinct symmetric two-locus disease penetrance models $(f^1 - f^4)$ and displacements $d = d(x)$.

The average conditional noncentrality parameter for a homogeneous pedigree set with perfect marker data and uniform weights $\gamma_k = c$ can be written as

$$D = \frac{\sum_{w_2} \mathrm{NCP}_{l_1|l_2}(w_2)\bar{S}(w_2)P_2(w_2)}{\sqrt{\sum_{w_2} \bar{S}^2(w_2)P_2(w_2)}}, \tag{45}$$

where $\bar{S}^2(w_2) = 2^{-m}\sum_{w_1} S^2(w_1, w_2)$ is the conditional variance and

$$\mathrm{NCP}_{l_1|l_2}(w_2) = \sum_{w_1} S(w_1, w_2)P(w_1|w_2)/\bar{S}(w_2)$$

the conditional noncentrality parameter for one pedigree when $v(l_2) = w_2$. Notice that any multiplicative constant of S cancels in (45). This does not violate (22), since c can be varied freely. For the optimal score function (44) (with proportionality constant removed), we get an optimal conditional noncentrality parameter

$$\mathrm{NCP}_{l_1|l_2}(w_2) = 2^{m/2}\sqrt{\sum_{w_1} [P(w_1|w_2) - 2^{-m}]^2} \tag{46}$$

and

$$D^2 = \sum_{w_2} \mathrm{NCP}^2_{l_1|l_2}(w_2)P_2(w_2).$$

Hence the conditional noncentrality parameter quantifies how much $P(\cdot|w_2)$ deviates from a uniform distribution and D^2 averages the squared conditional noncentrality parameter with respect to P_2.

Figure 7 displays (46) for Pedigree 1 when IBD(w_2), the number of alleles shared identical-by-descent by the affected sib-pair, equals 0, 1 or 2. For additive and heterogeneity models, $\mathrm{NCP}_{l_1|l_2}(w_2)$ decreases with IBD(w_2), whereas the opposite is true for threshold models. For multiplicative models, $P(w_1, w_2) = P_1(w_1)P_2(w_2)$ when there are no unaffecteds in the pedigree. Hence $P(w_1|w_2) = P(w_1)$ and $\mathrm{NCP}_{l_1|l_2}(w_2)$ is independent of w_2. Note that the conditional NCP, given fixed prevalence K, is increasing with allele frequency p, which is consistent with Figure 6.

5.2 Power and Significance

Although the NCP is related to power the two concepts are not equivalent due to nonnormality of linkage scores and multiple testing. For this reason we also discuss power in this section.

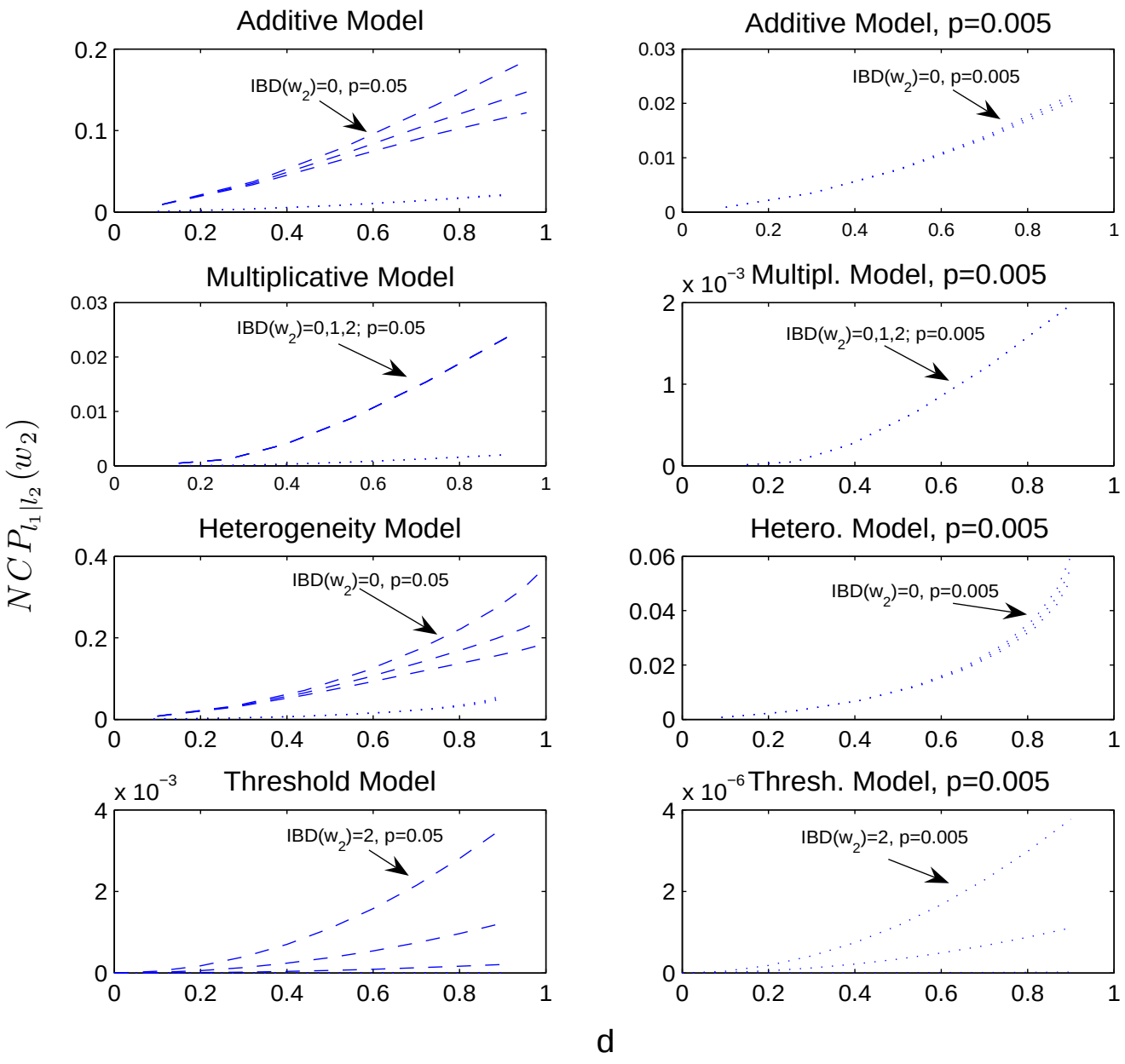

Figure 7: Optimal conditional noncentrality parameters $\mathrm{NCP}_{l_1|l_2}(w_2)$ for Pedigree 1. We assume a fixed conditioning locus l_2, $p = p_1 = p_2 \in \{0.005, 0.05\}$ (dotted and dashed lines), disease prevalence $K = 0.1$, four symmetric two-locus penetrance models ($f^1 - f^4$) and displacements $d = d(x)$.

Methods of Calculation

The significance level and power can be calculated using either: (i) *Analytical approximations* based on Gaussian extreme value theory (Lander and Botstein, 1989; Feingold, 1993; Feingold et al., 1993; Lander and Kruglyak, 1995). (ii) *Simulations* (Boehnke, 1986; Ploughman and Boehnke, 1989; Ott, 1989; Terwilliger et al., 1993).

The advantage of (i) is fast computations and available explicit expressions. However, it is still only an approximative procedure which, even in the modified versions correcting for nonnormality (Tang and Siegmund, 2001; Ängquist and Hössjer, 2005), may give biased results. Related approaches are, for example, described in Hernández et al. (2005) and Bacanu (2005). On the other hand, (ii) is more adjustable to complicated situations and do not give biased results in the limit of large Monte Carlo samples. Its drawback is rather the computational burden. Modified simulation algorithms have been suggested in order to deal with this problem, such as importance sampling (Malley et al., 2002; Ängquist and Hössjer, 2004) and the fast but slightly biased replicate-pool method (Song et al., 2004; Wigginton and Abecasis, 2006).

For conditional two-locus analysis with known conditioning locus, another possibility is to use *permutation testing* when marker data from the pedigrees are fully, or close to, exchangeable. One may note that the procedure outlined in Cox et al. (1999), using our general two-locus score function framework, may be generalized to permuting inheritance vectors rather than one-locus NPL scores at the conditioning loci.[6] With unknown conditioning loci, there is an additional level of uncertainty regarding the actual set of conditioning loci and their inheritance vectors. It seems difficult to adjust analytical approximations, permutation testing and fast simulation procedures to this in proper and convenient ways. Hence we use direct Monte Carlo simulation based on J replicates in all simulations. For instance, estimates $\hat{\alpha}_{\mathbb{Y}}(u)$ and $\hat{\beta}_{\mathbb{Y}}(u)$ of the significance level and power in Section 3.4 are $\sum_{j=1}^{J} I(p_{\min}^{j} \leq u)/J$, where $p_{\min}^{j}$ is the minimum p-value for the j^{th} replicate. For all J replicates we simulate marker data along all chromosomes conditional on phenotypes under H_0 and H_1 respectively.

We present calculations using so called *receiver operating characteristic (ROC)* curves (Selin, 1965; Bradley, 1996) by plotting power against significance level for various thresholds.

[6] Explicitly, the original set of inheritance vectors $v_k(y)$ at the conditioning loci y is being replaced by the permuted counterpart $v_{\pi_k}(y)$, where $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$ is a permutation of $(1, 2, \ldots, N)$.

Monte Carlo Simulations

The power calculations below are performed using score functions S_{pairs} (Whittemore and Halpern, 1994) in the one-locus case, and S^{2loc}_{pairs} and S^{Cox}_{pairs}, with epistatic weights $f(Z) = f^0_{\text{epi}}(Z)$, in the conditional two-locus case, see (24) and Appendix 6.3. In addition we also included the NCP-optimal score function S_{opt} of (44), in the one- and conditional two-locus simulations.

We consider homogeneous pedigree sets with equal pedigree weights $\gamma_k = 1/\sqrt{N}$, four chromosomes of equal length 1.5 M with disease loci l_1 and l_2 located in the middle of the first two chromsomes. Further, we use symmetric additive (f^1), multiplicative (f^2), heterogeneity (f^3) and threshold (f^4) models, setting the prevalence to K=0.01, the maximum penetrance to f_{22}=0.99 and the disease-allele frequencies $p = p_1 = p_2$ so that f_{00} attains its minimum value 0.

Our simulated one-locus results are summarized in Figures 8-9. The performance of S_{opt} and S_{pairs} is quite similar, in some cases practically identical. This may be explained as follows: (i) These examples involve small pedigrees, almost only consisting of affected individuals. In such cases, S_{opt} and S_{pairs} are usually quite similar. (ii) The NCP-criterion maximizes the expected NPL score at the disease locus. This is not necessarily the optimal approach with respect to maximizing power.

Generally, for small pedigrees S_{pairs} is often close to optimal. For larger pedigrees S_{opt} often outperforms S_{pairs} to an extent depending on the genetic model.

Next, we compare conditional two-locus power calculations based on a *single* conditioning locus $y = l_2$ in (26) with a single estimated locus $y = \hat{l}_2$ in (34) from Chromosome $c(l_2)$. The results are displayed in Figures 10-13.

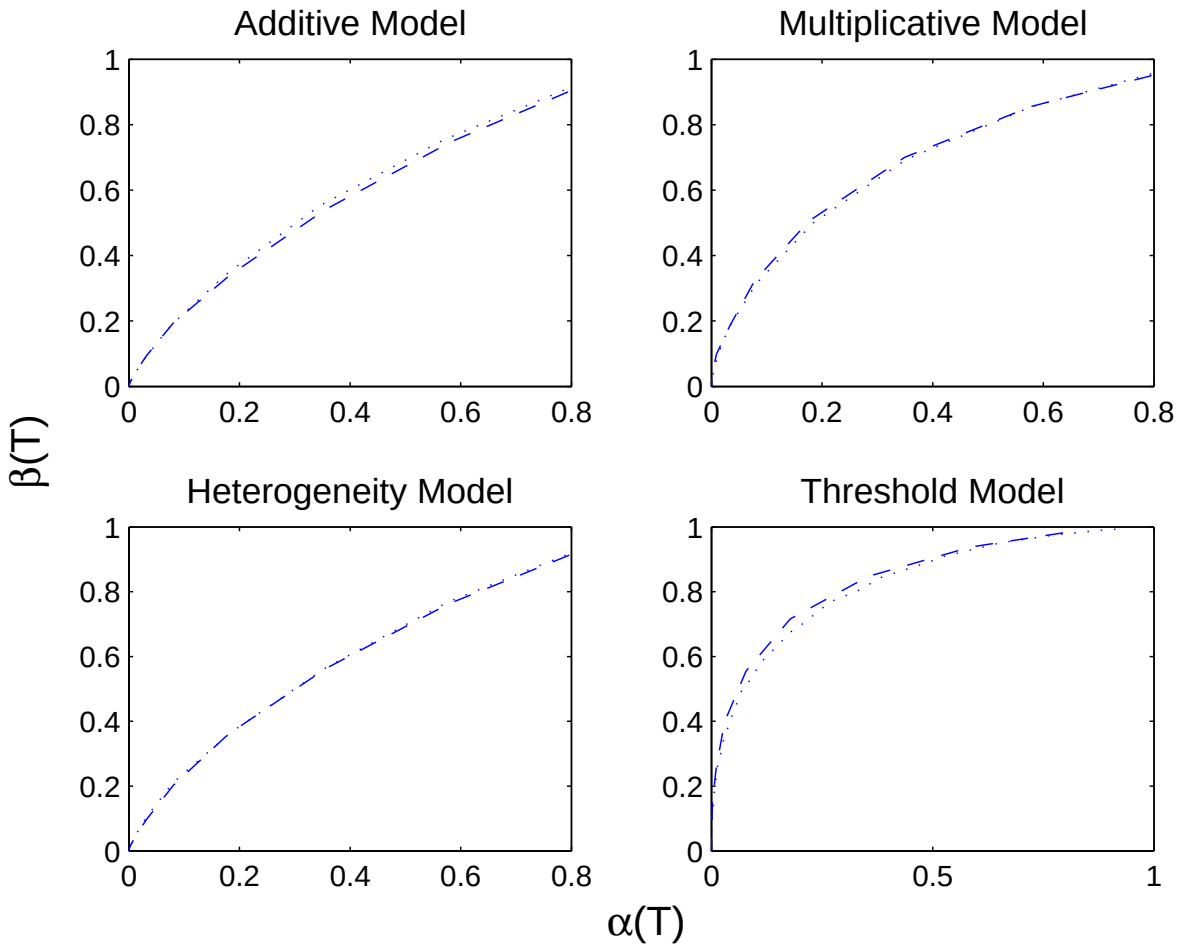

Figure 8: One-locus ROC-curves, using Pedigree 1 and $K = 0.01$ for: (i) An additive model, f^1, with $p = p_1 = p_2 = 0.0075$. (ii) A multiplicative model, f^2, with $p = p_1 = p_2 = 0.08$. (iii) A heterogeneity model, f^3, with $p = p_1 = p_2 = 0.005$. (iv) A threshold model, f^4, with $p = p_1 = p_2 = 0.15$. Two distinct score functions are used, S_{opt} and S_{pairs} (dotted and dashed lines respectively), $N = 25$, the thresholds $T = 2.0, 2.1, \ldots, 6.0$ and the number of simulations $J = 10000$.

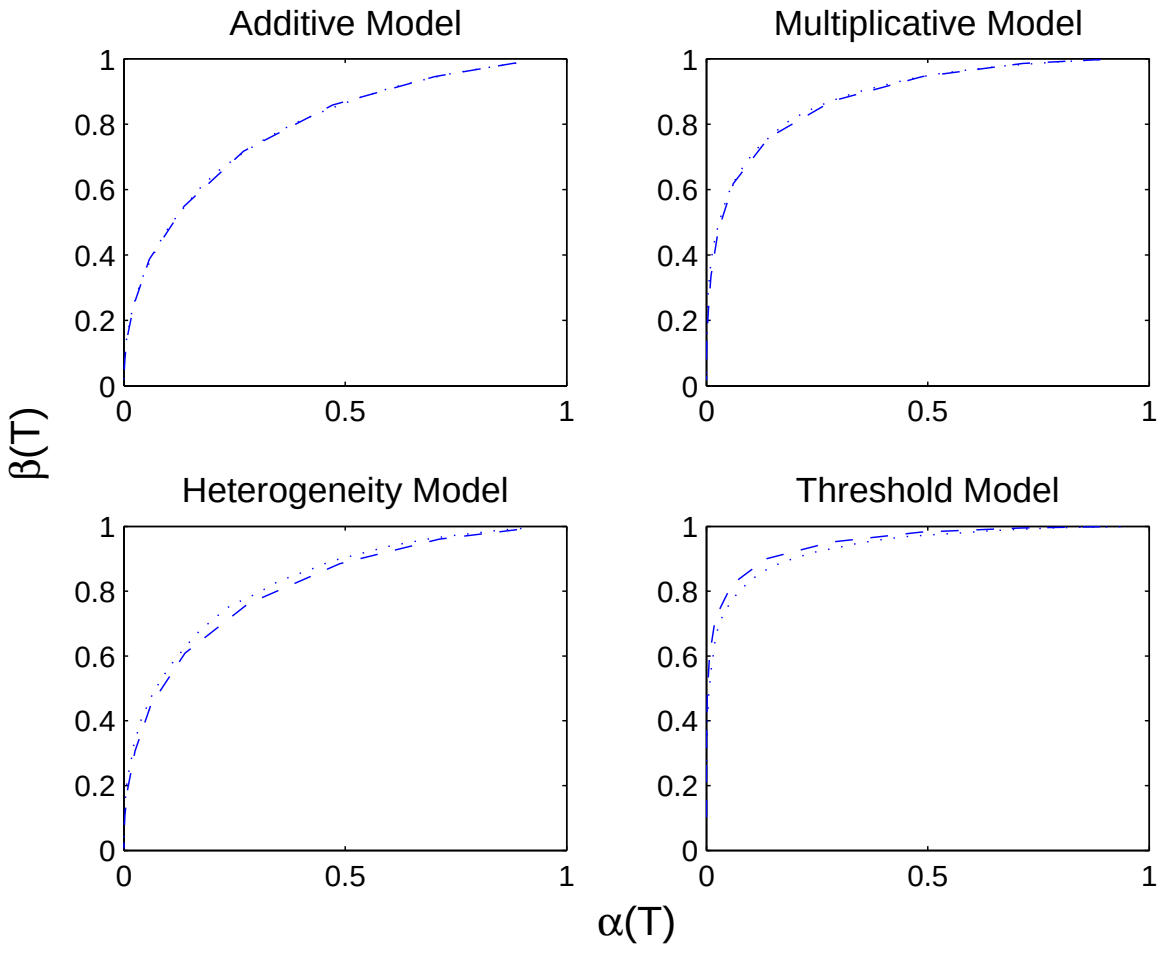

Figure 9: One-locus ROC-curves, using Pedigree 3. For more details, see Figure 8.

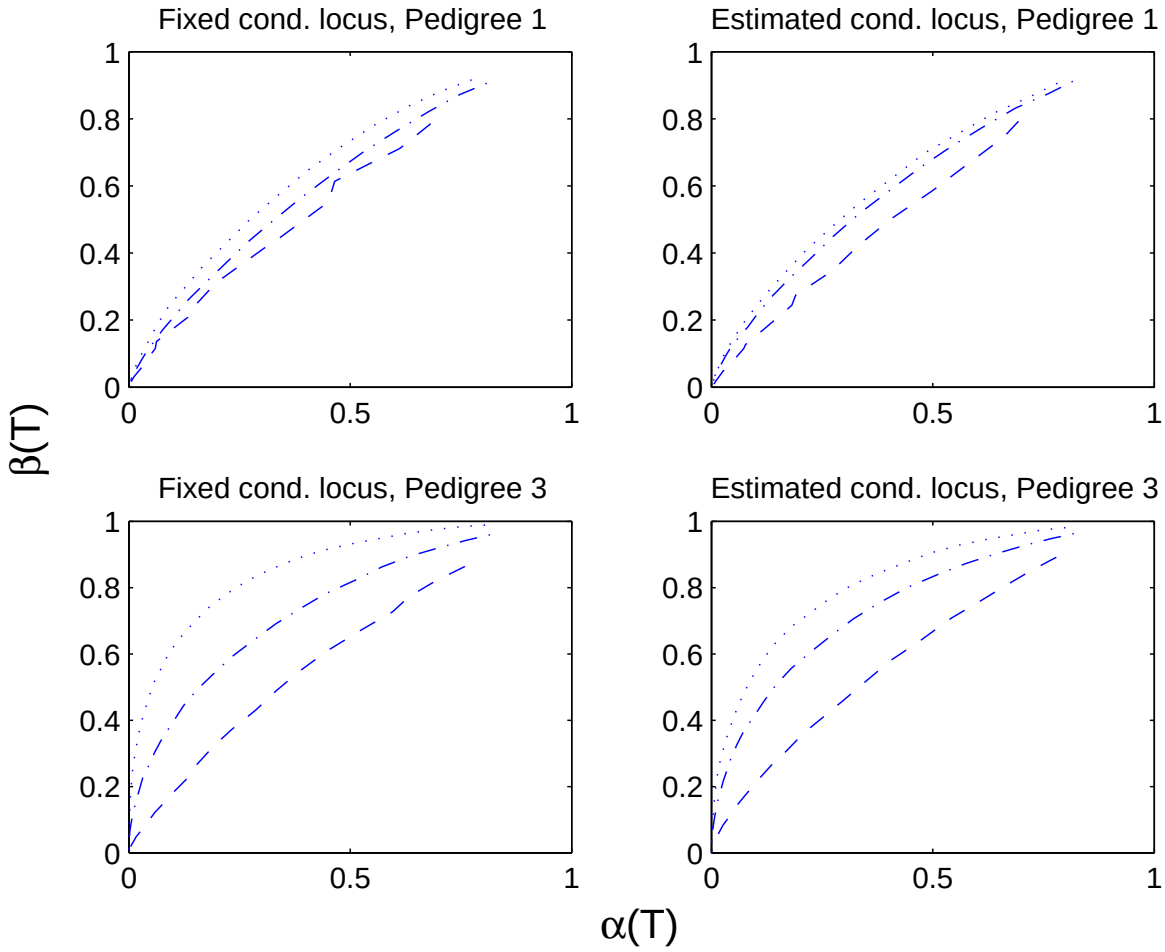

Figure 10: Conditional two-locus ROC-curves, using Pedigrees 1 and 3, under the additive disease model (f^1; d=0.99, x=0.2425) with $K = 0.01$, $p = p_1 = p_2 = 0.0075$ and three score functions (S^{Cox}_{pairs}, S^{2loc}_{pairs} and S_{opt}; dashed, dashed-dotted and dotted lines respectively). The number of pedigrees is $N = 25$, the thresholds $T = 2.0, 2.1, \ldots, 6.0$ and the number of simulations $J = 10000$.

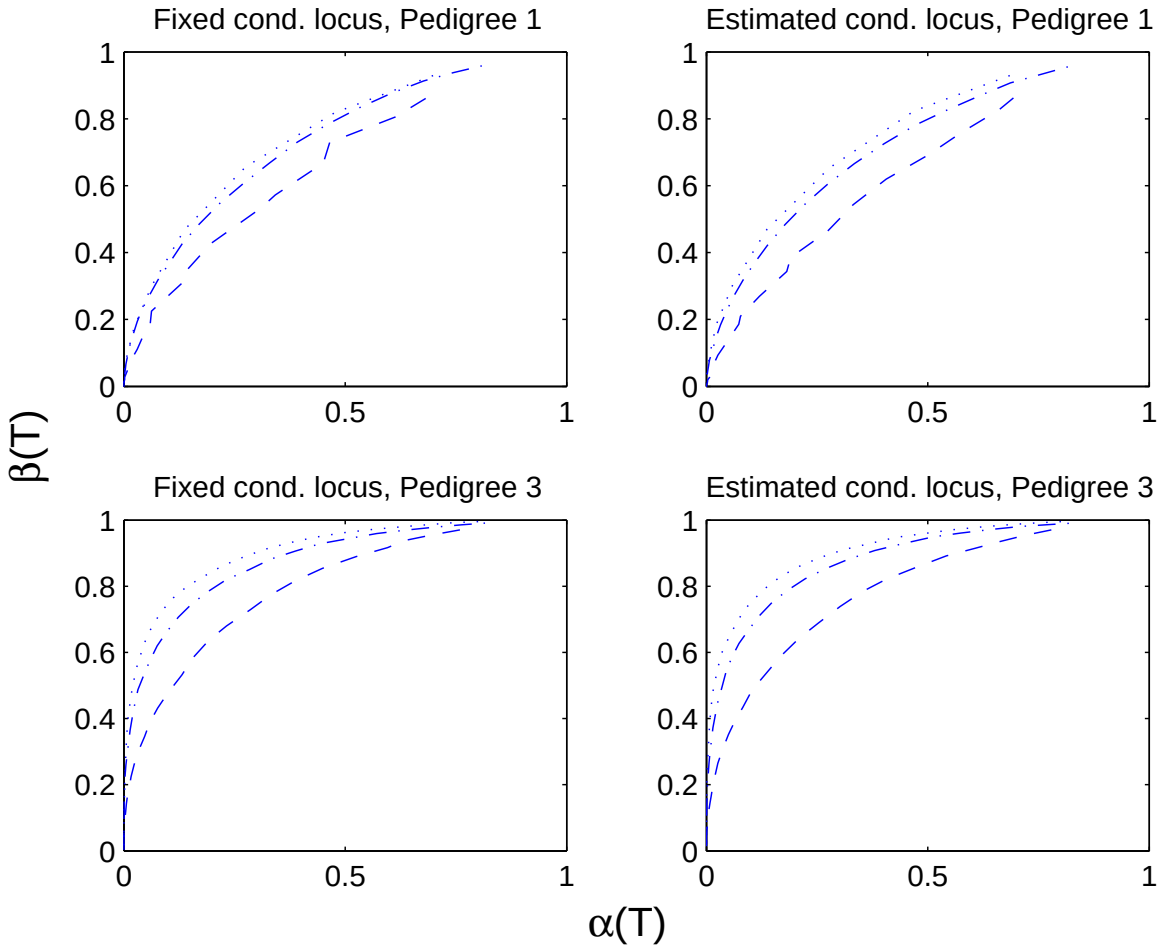

Figure 11: Conditional two-locus ROC-curves, using the multiplicative disease model (f^2; d=0.99, x=0.4925) with $p = p_1 = p_2 = 0.08$. For more details, see Figure 10.

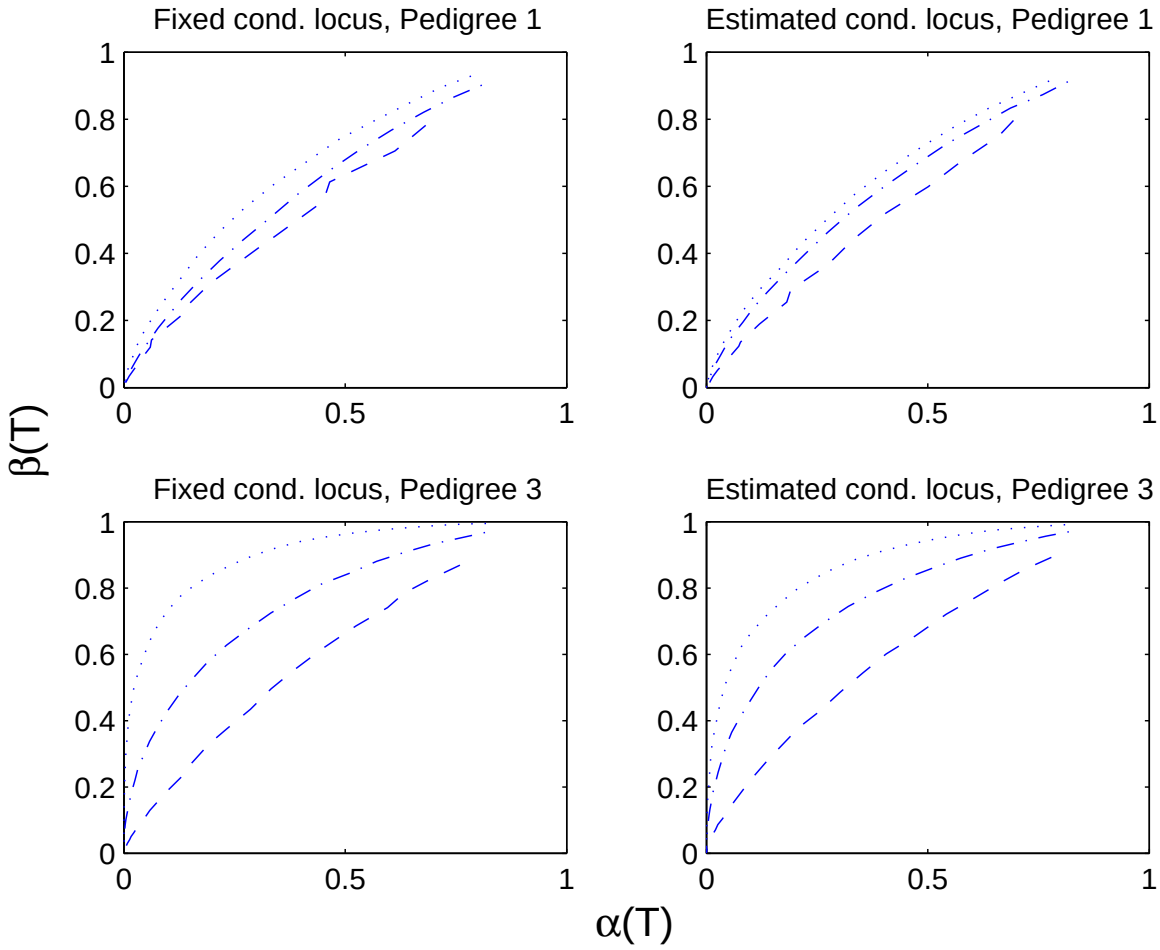

Figure 12: Conditional two-locus ROC-curves, using the heterogeneity disease model (f^3; d=0.99, x=0.4450) with $p = p_1 = p_2 = 0.005$. For more details, see Figure 10.

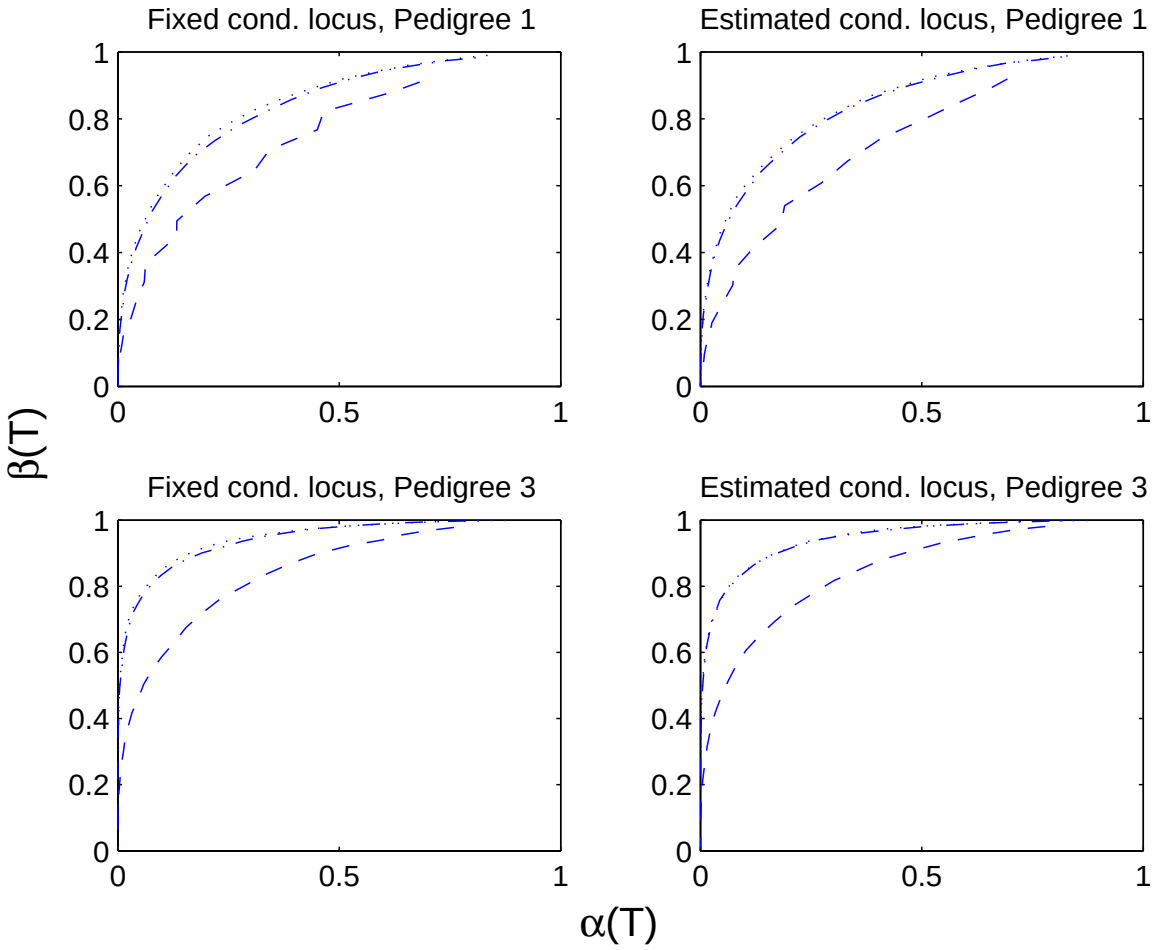

Figure 13: Conditional two-locus ROC-curves, using the threshold disease model (f^4; d=0.99, x=0.4900) with $p = p_1 = p_2 = 0.15$. For more details, see Figure 10.

Let $\hat{l}_1 = \arg\max_{x \in \Omega \setminus c(l_2)}$ be the estimated disease locus using the conditional NPL score. In Table 1 the proportions

$$r = \hat{P}\left[c(\hat{l}_1) = c(l_1)\right]$$

of chromosome-wise correctly estimated disease loci are displayed, including only those simulations for which $Z_{\max,y} \geq 3$.

Table 1: Proportions of estimated disease loci that are located on $c(l_1)$, given knowledge that the other disease loci is located on $c(l_2)$ at either a *fixed* or *estimated* position. For more details, see Figure 10.

Gen.Mod.	Sc.func.	Pedigree 1		Pedigree 3	
		Fixed	Estimated	Fixed	Estimated
f^1 (Addi.)	S^{Cox}_{pairs}	0.5226	0.5789	0.5287	0.5798
	S^{2loc}_{pairs}	0.5725	0.5879	0.7553	0.7579
	S_{opt}	0.6453	0.6134	0.8702	0.8422
f^2 (Mult.)	S^{Cox}_{pairs}	0.6568	0.6684	0.8073	0.8181
	S^{2loc}_{pairs}	0.7140	0.7282	0.8967	0.8938
	S_{opt}	0.7325	0.7480	0.9176	0.9185
f^3 (Hete.)	S^{Cox}_{pairs}	0.5258	0.5586	0.5552	0.5918
	S^{2loc}_{pairs}	0.5689	0.6038	0.7843	0.7886
	S_{opt}	0.6572	0.6243	0.9162	0.8886
f^4 (Thre.)	S^{Cox}_{pairs}	0.7456	0.7609	0.8501	0.8636
	S^{2loc}_{pairs}	0.8421	0.8611	0.9507	0.9545
	S_{opt}	0.8596	0.8609	0.9573	0.9542

Further, we consider a *random number* of conditioning loci. We select these loci by setting $z_c = 2.5$ in (31) for all $C = 4$ chromosomes in Ω and all three choices of score functions, see Figures 14-15. Since this case is simulation-wise more complex than the previous cases, we make some comments on the simulation procedure in Appendix 6.3.

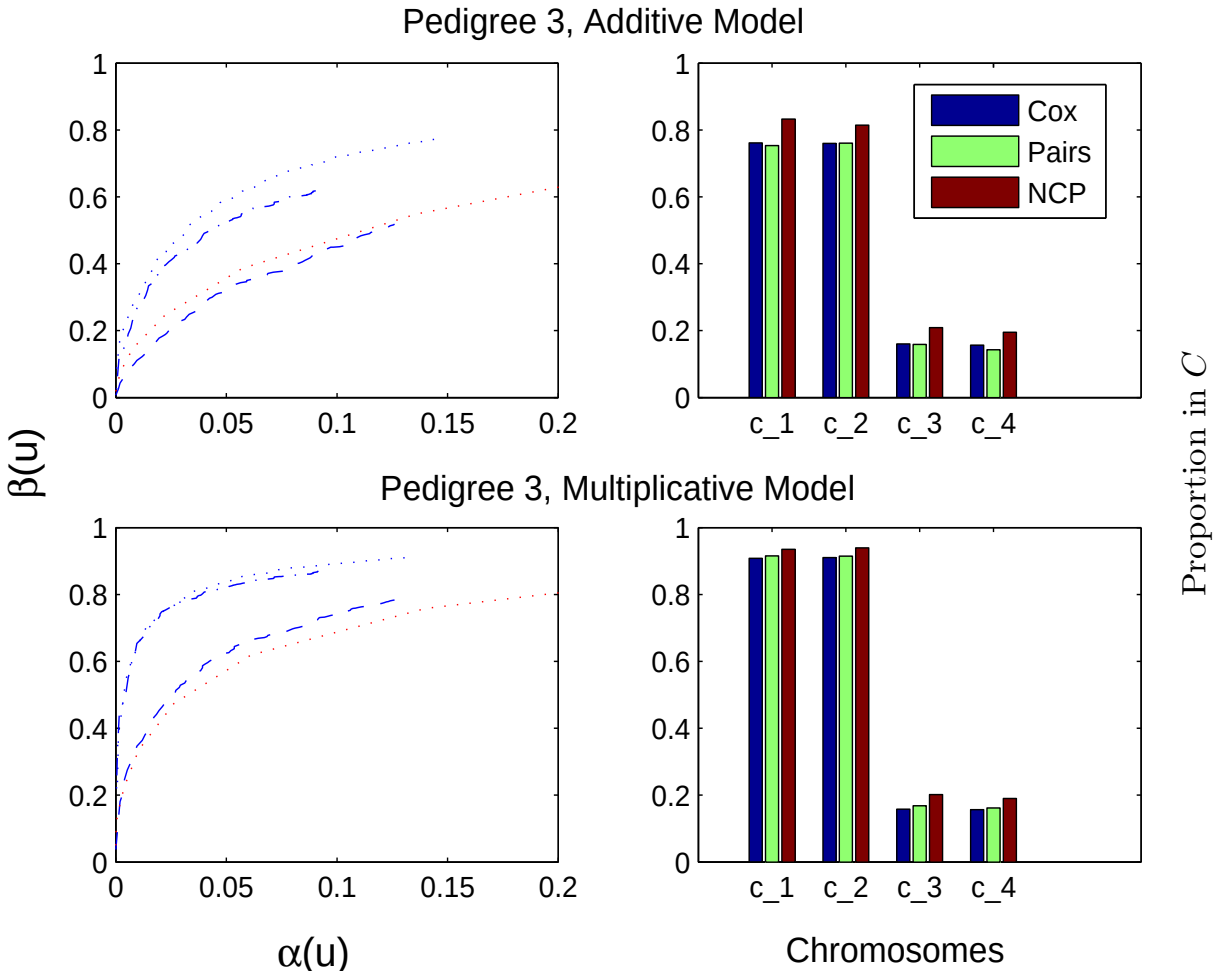

Figure 14: [Left] Conditional two-locus ROC-curves, with a random number of conditioning loci, for Pedigree 3 and the additive (f^1) and multiplicative (f^2) disease models, using three two-locus score functions, thresholds $u = 0, 0.0025, \ldots, 0.25$ and $J = 2500$ simulations. The one-locus ROC-curves based on S_{opt} is also displayed (lower dotted line). For further details, see Figure 10. [Right] Estimated probabilities for each chromosome of being selected as conditioning locus. Note that conditioning loci are selected through one-locus scores (S_{pairs} for analysis based on S^{Cox}_{pairs} or S^{2loc}_{pairs}; one-locus S_{opt} for analysis based on two-locus S_{opt}).

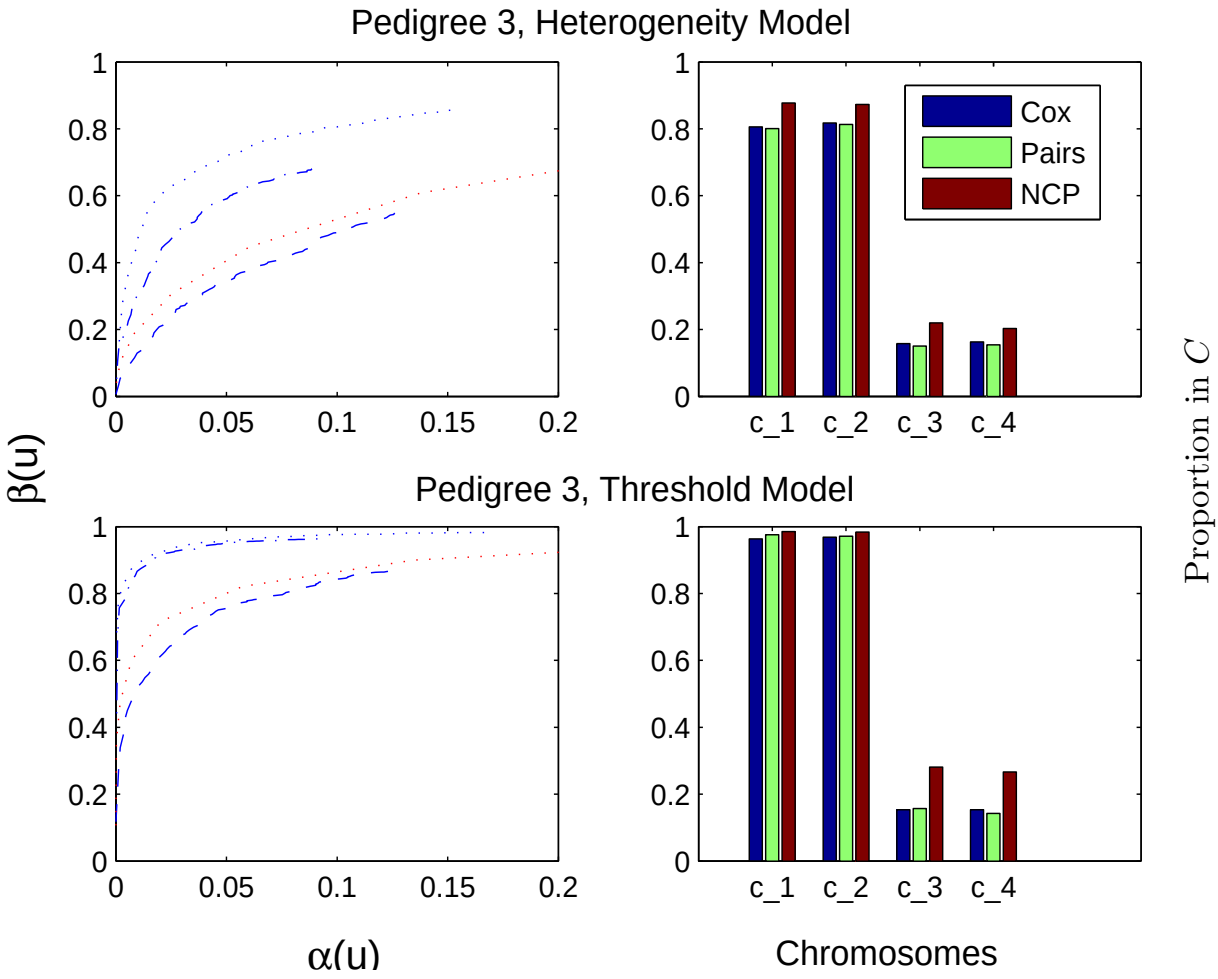

Figure 15: [Left] Conditional two-locus ROC-curves with a random number of conditioning loci, using the heterogeneity (f_3) and threshold (f_4) disease models. [Right] Estimated probabilities for each chromosome of being selected as conditioning locus. For further details, see Figures 10 and 14.

Also in Figures 14-15 the estimated probabilities

$$\hat{P}(c_i \in \mathbb{C}) = \frac{1}{J}\sum_{j=1}^{J} I(Z^j_{\max,c_i} \geq 2.5),\ i = 1,2,3,4,$$

of being selected as a conditioning locus is shown for each of the four chromosomes of Ω.

Remark 3 The interpretation of power with one mandatory conditioning locus versus a random number of loci is somewhat different. The latter case refers to the power to detect any disease locus (H_1), whereas in the former case power is restricted to detection of disease loci outside the conditioning chromosome c ($\bar{H}_1^c$). Hence, the corresponding ROC-curves are not directly comparable.

The score functions $S = S_{\text{opt}}$ in (44) which maximize NCPs do not have to maximize power, since NPL score distributions deviate from the standard normal and multiple testing is ignored in the NCP criterion. However, the score function that actually maximizes power should in most cases be close to S_{opt}. See for instance Feingold et al. (1993) how the NCP is related to an analytical approximation of the power in the one-locus case.

The methods of Section 3.2 and 3.4 include more multiple testing than those of Sections 3.1 and 3.3. For instance, this implies that a two-locus ROC-curve might be dominated by a one-locus ROC-curve, even though $\beta_{\text{tl}}(z) \geq \beta(z)$ for each threshold z.

6 Discussion

We have outlined and discussed procedures of conditional two-locus NPL analysis. Our primary focus has been approaches that facilitate the calculations of significance levels, power and noncentrality parameters.

6.1 Conditional Two-Locus Power Calculations

Given the situation, one may note that in all investigated cases S_{opt} turned out to be the most powerful score function, followed by S^{2loc}_{pairs} and S^{Cox}_{pairs}. A general observation is that, in many cases, the performances of the first two are quite similar, whereas the second one is far more powerful than the third procedure. This behaviour seems to be quite consistent with respect to disease models.

The main implication of this is that S^{2loc}_{pairs} clearly outperforms the well-known S^{Cox}_{pairs}. The performance of S^{2loc}_{pairs} may be improved on, and optimized, by adapting k (see Appendix 6.3) to the genetic model. This is a topic deserving further study.

6.2 Choice of Score Function and Conditional Method

The score function S_{opt} requires a known genetic disease model. However, for most real cases, the underlying disease model is, at least to some extent, unknown. Though this is a severe limitation, information from previous *segregation analyses* (Khoury et al., 1993; Haines and Pericak-Vance, 1998) may be helpful in the sense of suggesting, or at least narrowing down the set of plausible, disease models. If the model uncertainty is low S_{opt} might be used directly, whereas a higher degree of model uncertainty calls for adjusted approaches. Either a single robust score function, which performs well under a wide range of genetic models, may be used, or several distinct score functions. For more details on choosing score functions in NPL analysis, see McPeek (1999), Lange and Lange (2004), Hössjer (2005a, 2005b) and Ängquist (2006).

When there is a strong genetic component at l_2 and the size of the pedigree set is large, we may estimate l_2 with good precision through a one-locus analysis. In this case the procedure (34) of Section 3.4 will have virtually the same power as when l_2 is known in Section 3.3. On the other hand, if there is no single strong genetic component it might be more realistic to adopt the method based on unknown conditioning loci.

6.3 Comparisons Between One-Locus and Conditional Two-Locus NCPs and Power

Using the four families of disease-models derived as in Appendix 6.3, we calculated one-locus and conditional two-locus NCPs and powers. This was done on prevalence-levels $K = 0.01$ and $K = 0.1$ respectively.

Among our genetic disease models, the threshold model-class is the only one where conditional two-locus NCPs are significantly higher than one-locus NCPs for $K = 0.1$ (Figures 2-5).

For the most informative pedigree structure, Pedigree 3, one sees that conditional two-locus power is generally much higher than corresponding one-locus power for $K = 0.01$ (Figures 8-15). The results seem to be consistent between model-classes. With respect to power, the multiplicative and threshold models perform best. In these cases, S^{2loc}_{pairs} is closest to S_{opt}, i.e. when the

possibility of real findings is the largest given unknown disease model. Generally, conditional two-locus S_{opt} and S^{2loc}_{pairs} perform much better than S^{Cox}_{pairs} (Figures 14-15). For our models, S^{Cox}_{pairs} rarely outperforms the one-locus S_{opt}.

One might note that the performance of our threshold models is highly dependent on the prevalence K. If p is not very large, y in (C.2) is close to zero. In this case, for large K, most affecteds carries few copies of D, leading to low NCPs and power. This is reflected in Figure 5 (K=0.1) and Figures 13 and 15 (K=0.01).

In Figures 10-13 we compare conditional two-locus power to the contrasting behaviour of using a conditioning locus of known or unknown location, where in the latter case the conditioning chromosome is known. The proportions r_{fix} and r_{est} of correctly estimated second disease loci are given in Table 1. Generally $r_{fix} \approx r_{est}$, which seems surprising at first. The explanation is that, using an estimated conditioning locus, we condition on an inheritance vector corresponding to a high one-locus score, i.e. in many cases perfectly consistent with the disease model. In other words, when the first locus has high marginal penetrance, there is little performance loss in finding the second locus compared to when the first locus is known. Though the differences is small, r_{fix}/r_{est} seems largest and smallest for S_{opt} and S^{Cox}_{pairs} respectively.

Appendix A: The Inheritance Distribution Under Alternative Hypotheses

For a single pedigree, given H_1 and phenotypes Y one may calculate the inheritance distribution $P(v|Y, H_1)$ in (42), for each pair of inheritance vector $v = (w_1, w_2) \in \mathbb{V} \times \mathbb{V}$, at the disease loci pair l_1 and l_2. Using *Bayes Theorem*

$$P(v|Y, H_1) = \frac{P(v, Y|H_1)}{P(Y|H_1)} = \frac{P(Y|v, H_1)P(v|H_1)}{P(Y|H_1)} \propto P(Y|v, H_1),$$

where the final part follows since $P(v|H_1)$ and $P(Y|H_1)$ are constants over $\mathbb{V} \times \mathbb{V}$.

Brute force algorithms for calculating $P(Y|v, H_1)$ are based on,

$$\begin{aligned} P(Y|v, H_1) &= \sum_{G_f} P(Y|G_f, v, H_1)P(G_f|v, H_1) \\ &= \sum_{G_f} P\Big[Y|G(G_f, v), H_1\Big]P(G_f), \end{aligned} \tag{A.1}$$

where the summation and conditioning is done with respect to the collective (ordered) disease genotypes of the founders G_f at l_1 and l_2. We used that

genotypes for the pedigree members G at l_1 and l_2 is a deterministic function of G_f and v and that G_f is independent of the inheritance vector v (no segregation distortion).

Assuming no polygenic effects, one has,

$$P(Y|G, H_1) = \prod_{i=1}^{n} P(Y_i|G_i, H_1),$$

where $G_i = (G_{i1}, G_{i2})$ are disease genotypes at both disease loci for Individual i, and $P(Y_i|G_i, H_1)$ depends on the penetrance matrix f when Y_i is known and $P(Y_i|G_i, H_1) = 1$ otherwise. Finally,

$$P(G_f) = \prod_{i} P(G_i),$$

where the product is taken over all founders. Each term $P(G_i) = P(G_{i1})P(G_{i2})$ depends on the disease allele frequencies p_1 and p_2.

Remark 4 Using founder phase symmetry (Kruglyak et al., 1996; Gudbjartsson et al., 2000; Strauch et al., 2000), it suffices to calculate $P(v|Y, H_1)$ for $2^{2(m-f)}$ pairs of inheritance vectors. (Note that f here denotes the number of founders.) For large pedigrees, it is preferable to replace direct summation in (A.1) by recursive algorithms like peeling (Cannings et al., 1978; Lauritzen and Sheehan, 2003).

Appendix B: Proof of NCP-Optimal Score Functions

Proof: [Theorem 1] Let m be the common number of meioses of the homogeneous pedigree set. For a given score function, with weights $\gamma_k = 1/\sqrt{N}$ and standardization (22), the NCP parameters in (41) are defined through

$$A = \sum_{w_1} P_1(w_1)S(w_1),$$
$$B = \sum_{w_1,w_2} P(w_1, w_2)S(w_1, w_2),$$
$$D = \sum_{w_2} \mu(w_2)P_2(w_2),$$

where $\mu(w_2) = \sum_{w_1} S(w_1, w_2)P(w_1|w_2)$.

Using the methods in Hössjer (2005c), A, B and D are maximized with respect to S under the constraints (8), (15) and (21)-(22) respectively. Following an analogous version of Proposition 1 therein, the maximum NCPs

turn out to be as in (43) and the corresponding NCP-optimal score functions as in (44). Maximization of D is done in two steps. First, $S(w_1, w_2) = C(w_2)\,[P(w_1|w_2) - 2^{-m}]$ is derived by maximizing $\mu(w_2)$ for each w_2 subject to a constraint on $\bar{S}^2(w_2)$. Then $\sum_{w_2} \mu(w_2)P_2(w_2)$ is maximized with respect to $C(w_2)$ subject to (22), which for large samples can be written as $\sum_{w_2} \bar{S}^2(w_2)P_2(w_2) = 1$. The optimal choice $C(w) = c$ gives the desired solution. ■

Remark 5 Evidently $A \leq B$ and $A \leq D$, but there is no simple order relation between B and D. Often $B \geq D$. In fact, one can show this for weak genetic models, i.e. when $P(w_1, w_2)$ is close to 2^{-2m} for all w_1 and w_2.

Appendix C: Details On the NCP Calculations

For simplicity, we assume that both disease allele frequencies are equal, i.e. $p = p_1 = p_2$, and that the penetrance matrix f in (1) is symmetric. The last two properties imply that the marginal one-locus genetic models are equal. The disease *prevalence* is defined as,

$$\begin{aligned} K = P(\text{affected}) = f_{22}p^4 + 2(f_{21} + f_{12})p^3q + (f_{20} + f_{02})p^2q^2 \\ + 4f_{11}p^2q^2 + 2(f_{10} + f_{01})pq^3 + f_{00}q^4, \end{aligned} \tag{C.1}$$

where $q = 1 - p$ is the normal allele frequency at both loci and $f_{ij} = f_{ji}$.

Taking advantage of (C.1) we calculate a *one-parameter family* of penetrance matrices for the four distinct two-locus disease model-types defined in (2)-(5):

(i) A collection of *additive models*,

$$f^1 = \begin{pmatrix} 2y & K+y+x & K+y+2x \\ K+y+x & 2(K+x) & 2K+3x \\ K+y+2x & 2K+3x & 2(K+2x) \end{pmatrix},$$

which, given K, p and x, is defined through,

$$y = -\frac{4p(K+x) - K(1+2p^2)}{2(1-p)^2}.$$

(ii) A collection of *multiplicative models*,

$$f^2 = \begin{pmatrix} y^2 & y(K+x) & y(K+2x) \\ y(K+x) & (K+x)^2 & (K+x)(K+2x) \\ y(K+2x) & (K+x)(K+2x) & (K+2x)^2 \end{pmatrix},$$

which, given K, p and x, is defined through,

$$y = \frac{Kp^2 - 2p(K+x) + \sqrt{K}}{(1-p)^2}.$$

(iii) A collection of *heterogeneity models*,

$$f^3 = \begin{pmatrix} y(2-y) & y+(1-y)(K+x) & y+(1-y)(K+2x) \\ y+(1-y)(K+x) & 2(K+x)-(K+x)^2 & 2K+3x-(K+x)(K+2x) \\ y+(1-y)(K+2x) & 2K+3x-(K+x)(K+2x) & 2(K+2x)-(K+2x)^2 \end{pmatrix},$$

which, given K, p and x, is defined through,

$$y = \frac{Kp^2 - 2p(K+x) + 1 - \sqrt{1-K}}{(1-p)^2}.$$

(iv) A collection of *threshold models*,

$$f^4 = \begin{pmatrix} K-2y & K-y & K \\ K-y & K & K+x \\ K & K+x & K+2x \end{pmatrix}, \qquad \text{(C.2)}$$

which, given K, p and x, is defined through,

$$y = -\frac{p^3(p-2)x}{(p+1)(p-1)^3}.$$

For each one-parameter family (i)-(iv), in addition to our initial assumptions, the constraints $0 \leq f_{ij} \leq 1$ define a set of valid models $x_1 \leq x \leq x_2$. The larger x is within this interval, the stronger is the genetic component of the model.

Remark 6 All penetrance matrices $f^1 - f^3$ are derived using,

$$g = h = (y, K+x, K+2x)$$

in (2)-(4), whereas f^4 is directly defined through (5).

Appendix D: The Generalized Two-Locus Version of S_{pairs}

The unstandardized one-locus version of S_{pairs} is defined as

$$S_{\text{pairs}}(w) \propto \sum_{i<j} IBD_{i,j}(w),$$

where summation is over all pairs of affected individuals and $IBD_{i,j}(w)$ equals the number of alleles shared IBD by the i^{th} and j^{th} individual given w.

One may generalize this into a two-locus score function in several ways. We consider

$$S_{\text{pairs}}(w_1, w_2) = \sum_{i<j} \Big[IBD_{i,j}(w_1) + IBD_{i,j}(w_2) \Big]^k,$$

which for $k > 1$ may be thought of as capturing epistatic joint pairwise IBD-sharing within a pedigree. The case $k = 1$ corresponds to the additive score function $S_{\text{pairs}}(w_1, w_2) = S_{\text{pairs}}(w_1) + S_{\text{pairs}}(w_2)$ (Strauch et al., 2000). Throughout the analysis we use $k = 2$.

Appendix E: The Multiple Conditioning Loci Simulation Procedure

This case is computationally more demanding since p_c in (36) must be computed for all conditioning loci. Each p_c essentially corresponds to a one-locus p-value, hence any of the methods described in Section 5.2 can be used.

Defining the original number of conditional unconditional one-locus simulations as J_1 and the additional number of simulations for each p_c as J_2, one has the following possibilities: (i) Simulate all p_c through an *inner* loop. This occupies less memory, but is computationally demanding. (ii) Save all single-test specific information and estimate p-values for all $|\mathbb{C}^1| + |\mathbb{C}^2| + \cdots + |\mathbb{C}^{J_1}|$ selected conditioning loci of the J_1 original simulations, using the *same* set of simulated inheritance vectors at the conditiong loci in the second run of J_2 simulations. This occupies significantly more memory, but avoids the need for an inner loop.

In our case, we have adopted (ii) and chosen J_2=2500 when using S^{Cox}_{pairs} and J_2=1000 for S^{2loc}_{pairs} and S_{opt}. Further, J_1=2500 in all three cases.

Acknowledgement

This research has in part been sponsored by the Swedish Research Council and the Novo Nordisk Foundation.

References

Ängquist, L. (2001). *Conditional two-locus NPL-analyses: Theory and applications* (Master's thesis No. 2001:E22). 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., Anevski, D. and Luthman, H. (2005). *Unconditional two-locus nonparametric 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. and Hössjer, O. (2004). Using importance sampling to improve simulation in linkage analysis. *Statistical Applications in Genetics and Molecular Biology*, *3*(1:5). (Electronic journal, 24 pages)

Ängquist, L. and Hössjer, O. (2005). Improving the calculation of statistical significance in genome-wide scans. *Biostatistics*, *6*(4), 520–538.

Bacanu, S. A. (2005). Robust estimation of critical values for genome scans to detect linkage. *Genetic Epidemiology*, *28*, 24–32.

Barber, M. J., Todd, J. A. and Cordell, H. J. (2006). A multimarker regression-based test of linkage for affected sib-pairs at two linked loci. *Genetic Epidemiology*, *30*, 191–208.

Bengtsson, O. (2001). *Two-locus affected sib-pair identity by descent probabilities: Constraints, parameterisation and estimation* (Licentiate thesis). Göteborg: Department of Mathematical Statistics, Chalmers University of Technology, Göteborg University.

Boehnke, M. (1986). Estimating the power of a proposed linkage study: A practical computer simulation approach. *American Journal of Human Genetics*, *39*, 513–527.

Bradley, A. P. (1996). ROC curves and the χ^2 test. *Pattern Recognition Letters*, *17*, 287–294.

Cannings, C., Thompson, E. A. and Skolnick, M. H. (1978). Probability functions on complex pedigrees. *Advances in Applied Probability*, *10*, 26–61.

Chiu, Y. F. and Liang, K. Y. (2004). Conditional multipoint linkage analysis using affected sib pairs: An alternative approach. *Genetic Epidemiology*, *26*, 108–115.

Cordell, H. J., Todd, J. A., Bennett, S. T., Kawaguchi, Y. and Farrall, M. (1995). Two-locus maximum lod score analysis of a multifactorial trait: Joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. *American Journal of Human Genetics*, *57*, 920–934.

Cordell, H. J., Wedig, G. C., Jacobs, K. B. and Elston, R. C. (2000). Multilocus linkage tests based on affected relative pairs. *American Journal of Human Genetics*, *66*, 1273–1286.

Cox, N. J., Frigge, M., Nicolae, D. L., Concannon, P., Hanis, C. L., Bell, G. I. and Kong, A. (1999). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genetics*, *21*, 213–215.

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

Dupuis, J., Brown, P. O. and Siegmund, D. (1995). Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. *Genetics*, *140*, 843–856.

Farrall, M. (1997). Affected sibpair linkage tests for multiple linked susceptibility genes. *Genetic Epidemiology*, *14*, 103–115.

Feingold, E. (1993). Markov processes for modeling and analyzing a new genetic mapping method. *Journal of Applied Probability*, *30*, 766–779.

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*, 234–251.

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

Haines, J. L. and Pericak-Vance, M. A. (1998). *Approaches to gene mapping in complex human diseases.* New York: John Wiley & Sons.

Hernández, S., Siegmund, D. O. and Gunst, M. D. (2005). On the power for linkage detection using a test based on scan statistics. *Biostatistics*, *6*(2), 259–269.

Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics*, *4*, 701–709.

Holmans, P. (2002). Detecting gene-gene interactions using affected sib pair analysis with covariates. *Human Heredity*, *53*, 92–102.

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

Hössjer, O. (2005a). Combined association and linkage analysis for general pedigrees and genetic models. *Statistical Applications in Genetics and Molecular Biology*, *4*(1:11). (Electronic journal, 42 pages)

Hössjer, O. (2005b). Conditional likelihood score functions for mixed models in linkage analysis. *Biostatistics*, *6*(2), 313–332.

Hössjer, O. (2005c). Information and effective number of meioses in linkage analysis. *Journal of Mathematical Biology*, *50*(2), 208–232.

Kämpe, M. (2001). *Two-locus nonparametric linkage analysis for complex diseases* (Master's thesis No. 2001:E4). Lund: Lund Institute of Technology, Lund University.

Khoury, M. J., Beaty, T. H. and Cohen, B. C. (1993). *Fundamentals of genetic epidemiology.* New York and Oxford: Oxford University Press.

Knapp, M., Seuchter, S. A. and Baur, M. (1994). Two-locus disease models with two marker loci: The power of affected sib-pair tests. *American Journal of Human Genetics*, *55*, 1030–1041.

Kong, A. and Cox, N. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics*, *61*, 1179–1188.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics*, *58*, 1347–1363.

Lander, E. S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, *121*, 185–199.

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

Lange, E. M. and Lange, K. (2004). Powerful allele-sharing statistics for nonparametric analysis. *Human Heredity*, *57*, 49–58.

Lauritzen, S. L. and Sheehan, N. A. (2003). Graphical models for genetic analyses. *Statistical Science*, *18*, 489–514.

Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity*, *50*, 334–349.

Liang, K. Y., Chiu, Y. F., Beaty, T. H. and Wjst, M. (2001). Multipoint analysis using affected sib-pairs: Incorporating linkage evidence from unlinked regions. *Genetic Epidemiology*, *21*, 105–122.

MacLean, C. J., Sham, P. C. and Kendler, K. S. (1993). Joint linkage of multiple loci for a complex disorder. *American Journal of Human Genetics*, *53*, 353–366.

Malley, J. D., Naiman, D. and Bailey-Wilson, J. (2002). A comprehensive method for genome scans. *Human Heredity*, *54*, 174–185.

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

Ott, J. (1989). Computer-simulation methods in human linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(11), 4175–4178.

Ploughman, L. M. and Boehnke, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics*, *44*, 543–551.

Schulze, T. G., Buervenich, S., Badner, J. A., Steele, C. J. M., Detera-Wadleigh, S. D., Dick, D., Foroud, T., Cox, N. J., MacKinnon, D. F., Potash, J. B., Berrettini, W. H., Byerley, W., Coryell, W., Jr, J. R. D., Gershon, E. S., Kelsoe, J. R., McInnis, M. G., Murphy, D. L., Reich, T., Scheftner, W., Jr, J. I. N. and McMahon, F. J. (2004). Loci on chromosomes 6q and 6p interact to increase susceptibility to bipolar affective disorder in the National Institute of Mental Health Genetics Initiative pedigrees. *Biological Psychiatry*, *56*, 18–23.

Selin, I. (1965). *Detection theory*. Princeton, New Jersey: Princeton University Press.

Sengul, H., Weeks, D. E. and Feingold, E. (2001). A survey of affected-sibship statistics for nonparametric linkage analysis. *American Journal of Human Genetics*, *69*, 179–190.

Song, K. K., Weeks, D. E., Sobel, E. and Feingold, E. (2004). Efficient simulation of P values for linkage analysis. *Genetic Epidemiology*, *26*, 88–96.

Strauch, K., Fimmers, R., Kurz, T., , Baur, M. P. and Wienker, T. F. (2003). How to model a complex trait: 2. analysis with two disease loci. *Human Heredity*, *56*, 200–211.

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*, 1945–1957.

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

Tang, H. K. and Siegmund, D. (2002). Mapping multiple genes for quiantitative or complex traits. *Genetic Epidemiology*, *22*, 313–327.

Terwilliger, J. D., Speer, M. and Ott, J. (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genetic Epidemiology*, *10*, 217–224.

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

Wigginton, J. E. and Abecasis, G. R. (2006). An evaluation of the replicate-pool method: Quick estimation of genome-wide linkage peak p-values. *Genetic Epidemiology*, *30*, 320–332.